PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 1

Superconductivity

J. Volger

Although it was discovered half a century ago, the phenomenon of superconductivity had, until recently, still not yielded up many of its secrets. In the last few years the situation has changed. Although by no means all the experimental data have been adequately explained, we now have a reasonably good understanding of the nature of superconductivity and of the strange properties of superconductors of the second kind, discovered a few years ago. These properties are in many respects the opposite of those that for decades have been considered characteristic of a superconductor. Some members of this group, the "hard" superconductors, which remain superconducting in strong magnetic fields and when carrying high currents, have in a short time attracted considerable interest from the designers of magnet coils. An article about coils which have been made in Philips Research Laboratories for producing very high magnetic fields will shortly appear in this journal. In the article below a brief survey of present knowledge of the phenomenon of superconductivity is given.

Introduction

The discovery of superconductivity

The phenomenon of superconductivity, which in recent years has attracted a great deal of attention through the publication of new experimental or theoretical discoveries or promising technical applications, was first observed in 1911 by Kamerlingh Onnes and his research assistant Holst (*fig. 1*). At the Leyden Physics Laboratory, where Kamerlingh Onnes had in 1908 been the first to succeed in liquefying helium, and which was later named after him, a series of investigations was at that time being performed on the electrical conductivities of metals in the newly attained temperature range. The electrical resistance of metals exhibits a positive temperature coefficient, and it was considered probable that the resistance of a metal would drop to a very low level if the temperature were reduced to very close to absolute zero. However, in the course of some measurements with a mercury wire, it was then found that at a certain temperature (about 4 °K) the resistance, which was already very low, suddenly became too low to measure [1]. This discovery was followed by many similar ones over the years, and it is now known that about half the metals and metallic elements in the periodic system can become superconducting below a certain "transition" temperature T_c, while the same is also true for many hundreds of compounds and alloys.

Extensive research has made it clear that superconductivity should not be regarded as a *property* (the decrease of the resistivity to zero), but rather as a *state*. We now, therefore, speak of the *superconducting phase* [2]. The superconducting phase exhibits a large number of striking features whose interrelation has become much clearer over the past few years, and which are also of technical interest.

The availability of low temperatures

The ease with which low temperatures can be reached today (here we mean temperatures near that of liquid helium, about 4 °K), compared with the trials and tribulations of the pioneering years, is a factor of considerable significance in research into and application of superconductors.

Viewed more broadly, man's ability to achieve low temperatures is really of very recent date. The discovery

Prof. Dr. J. Volger is with Philips Research Laboratories, Eindhoven, as a Scientific Adviser.

[1] H. Kamerlingh Onnes, Comm. Phys. Lab. Univ. Leiden **12**, Nos. 122b and 124c, 1911. A survey of the initial period of research into superconductivity will be found in C. J. Gorter, Rev. mod. Phys. **36**, 3, 1964 and K. Mendelssohn, ibid. p. 7.

[2] For the "conventional" considerations on superconductivity see F. London, Superfluids, Dover Publ., New York 1961, and D. Shoenberg, Superconductivity, Cambridge Univ. Press, Cambridge 1952. A survey of the recent developments is also given in E. A. Lynton, Superconductivity, Methuen, London 1964.

and mastery of fire dates back thousands of years. The production of cold, on the other hand, was not successful until the last century, as can be seen clearly in *fig. 2*. The points in this figure represent the principal successes achieved through the years. The liquefaction of helium marked an important stage along this road. Further cooling would now only be possible by means other than the liquefaction of gases. It is indeed worthy of note that, for about thirty years, research at extremely low temperatures was possible at only three laboratories in the world, the Kamerlingh Onnes Laboratory at Leyden and laboratories in Toronto and Berlin. Now, however, there are many hundreds of research and development centres where helium can be liquefied and used as a coolant for all kinds of investigation. Among these centres are many large industrial laboratories; this comes about chiefly because of the keen interest at such laboratories in solid-state research, for which low temperatures are now essential. In addition there is more interest than ever before in the technical problem of liquefaction itself. We should mention here the spectacular success of Collins, who has considerably improved the expansion machine forming part of the Kapitza-type helium liquefier. The helium liquefier developed by Collins is in commercial production and has been on the market now for nearly twenty years. This has contributed considerably towards the post-war increase in cryogenic research and engineering. The development of a helium liquefier based on the Stirling cycle has also recently been successful [3] and this machine offers considerable promise.

An important factor in all this is the availability of helium itself. Its concentration in the atmosphere is very low (5×10^{-6}). Nevertheless, it is possible to obtain helium from air since it is one of the by-products of the large air-rectification installations now in use for the production of oxygen, argon, etc. The gas can also be obtained in small quantities by heating certain radioactive minerals. The most economical method of obtaining it is from sources of natural gas containing helium. Here, the helium is the end product of radioactive conversions, which has been retained under domed impermeable layers of rock and happens to have become mixed with natural gas trapped in the same geological structure. As far as is known, the natural gas fields in the

Fig. 1. H. Kamerlingh Onnes (right) and G. Holst, who was research assistant to Prof. Kamerlingh Onnes at the time of the discovery of superconductivity in 1911. Dr. Holst (later Prof. Holst) was to become the founder of the Philips Research Laboratories and their director till 1946.

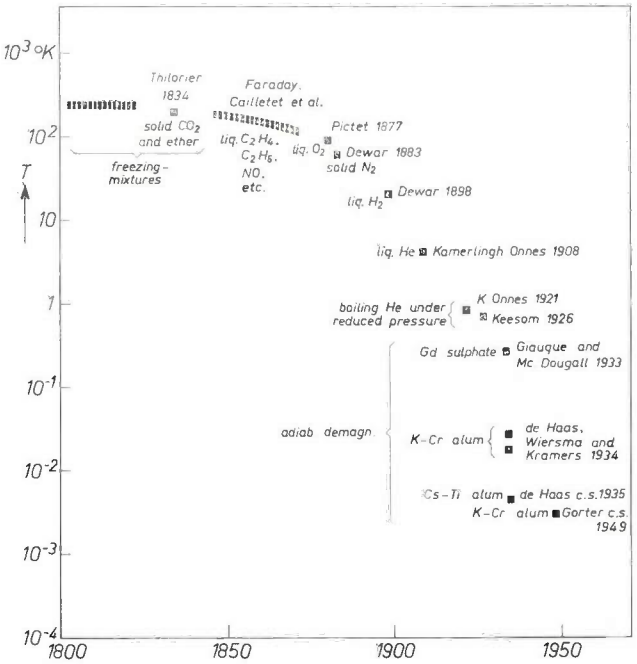


Fig. 2. A diagram which shows how recent is man's penetration of the field of low and very low temperatures. The dots are in effect the milestones in the history of cryogenic research.

south of the United States of America are the richest in helium. Many sources there contain a small percentage of helium, and some contain as much as 7 to 9 mol%. The helium is separated off by cryogenic methods. At present, annual production is estimated at 5×10^{10} litres at n.t.p. To obtain 1 litre of liquid helium about 800 litres of gas at n.t.p. are needed. The known world reserves should be some 5×10^{12} litres of gas at n.t.p. The consumption of helium in cryogenic research and technology is relatively low and amounts only to a small percentage of total production. More than half the world's production is used for welding and as a compressing gas for the liquid fuel in space craft.

Superconductivity and electrical technology

Industrial interest in a physical phenomenon lies in a "quality" or "outstanding feature" of that phenomenon. And in fact, the absence of electrical resistance unlocks a wealth of new possibilities. Although the highest transition temperature so far found is still as low as 20.05 °K, the possibility of using underground cooled superconducting cables for electricity supply, in certain densely populated areas such as London, is already being seriously examined to see what advantages it might offer. There is also interest in the use of superconducting elements for logic circuits ("cryotrons"). Some very interesting possible uses include superconducting magnet coils which can be used to give very high magnetic fields, sometimes with flux densities greater than 10 Wb/m² (100 000 gauss) over volumes large enough to be of interest in technical applications [4]. One wonders how many other even more impressive changes in electrical technology are to be expected if superconducting materials with a transition temperature of 50 or 100 °K are ever produced. Even though insulation losses are no problem today — modern insulating materials permit the escape of less than 1 W per m² at a temperature difference of 300 °K — the existence of such materials with higher transition temperature would appreciably reduce the technical difficulties and expense of cold production, both of which go up as the temperature goes down. Thus in practice the production of 1 kWh of cold at 77 °K requires 10 kWh of electric energy, but at 4 °K 1000 kWh are required.

Nevertheless, it is to be expected that, even if superconductors with transition temperatures well above 20 °K are never found, superconductivity will still find its applications in technology.

The superconducting state

Comparison with the normal state

The theoretical treatment of conduction in a normal metal is fairly simple. The electrons can be regarded as independent particles which travel through the crystal in all directions at random velocities. If there is no electric field, the average velocity is zero. If there is a field, however, the electrons move with a certain average velocity, the drift velocity, in a certain direction. The velocity distribution of the electrons is thus not symmetrical about zero.

In this situation the conduction electrons transfer at collisions more energy on average to the crystal lattice than they receive, and heat is developed. Between collisions, the energy which has been given up is recovered from the electric field. This is an irreversible process, so that the current quickly drops to zero once the field has been switched off. The velocity distribution then once more becomes symmetrical about zero.

If we are to apply quantum mechanics, we must consider the "cloud" of conducting electrons as a gas whose particles obey the rules of the Fermi-Dirac statistics: in accordance with Pauli's exclusion principle the various quantum states which we know from wave mechanics may only be occupied by one electron. This means that, at absolute zero ($T = 0$), and zero electric field, all levels with energy lower than a certain value E_F (the Fermi energy) are occupied and all higher levels are unoccupied. At a higher temperature the boundary between both regions is not so sharp. Roughly speaking this takes an energy interval of magnitude kT distributed about the value E_F (k is Boltzmann's constant). In this scheme, the flow of current means that some electrons have begun to occupy higher, previously unoccupied, energy levels by taking up energy from the field. The loss of energy on collision corresponds to falling back to a lower energy level. In the normal state, the flow of current is possible only through the excitation of separate electrons.

In a superconductor the situation is quite different. According to Bardeen, Cooper and Schrieffer [5], extensive ordering has been established in the material. Electrons with an equal but opposite electromagnetic momentum (see below), particularly those whose energy values lie in a narrow interval around E_F, have entered into a strong interaction in pairs, through the intermediary of the lattice vibrations. The details of this interaction, which can extend over a relatively large distance, will not be further dealt with here but we should point out that, in the ground state of the elec-

[3] G. J. Haarhuis, Proc. 12th Int. Refrigeration Congress, Madrid 1967, p. 894.

[4] The work done in this field at Philips Research Laboratories will be discussed in an article by A. L. Luiten which will shortly appear in this journal.
A general survey, with particular reference to large magnets, will be found in C. Laverick, Superconducting magnet technology, Adv. Electronics and Electron Phys. 23, 1967.

[5] A summary of the theory of Bardeen, Cooper and Schrieffer together with an extensive bibliography is to be found in J. Bardeen and J. R. Schrieffer, Progress in low temperature physics 3, 170-287, 1961.

tron pairs, the total electromagnetic momentum of the two electrons is exactly zero. (This should not however be taken to imply that their total mechanical momentum or the velocity of their common centre of gravity is also zero.) Moreover, the total energy of the electrons forming such a pair is slightly lower than it was before the interaction took place. Before going into the question of how such an ordering leads to superconductivity, let us have a quick look at the appearance of the band diagram in both cases, paying particular attention to the density of states.

The two graphs of *fig. 3* show the curve of the energy E against the momentum p of an electron on the right, and the curve of the density of states N against E on the left, all the curves referring to energy values near to E_F. The reference point chosen for the energy is E_F. For a normal metal (fig. 3*a*), N gradually rises with E; the relation between E and p can for example be parabolic. At $T = 0$, there are no E values greater than E_F and no p values greater than p_F. At $T > 0$, some electrons have a greater energy and, correspondingly, some levels with lower energies than E_F are unoccupied.

For a metal which is in the superconducting state (fig. 3*b*), there is a gap — a forbidden zone — of width 2Δ in the spectrum of the possible energy values of the electrons around E_F. Just above and below this gap, the density of states is much greater than in the normal state. The electrons which in the normal state populated the region between $E_F - \Delta$ and $E_F + \Delta$ can take up a position in these edge regions. The quantity 2Δ could be said to be the "binding energy" of the pairs which is liberated when they are formed. At $T = 0$, Δ has its maximum value which we call Δ_0, and the highest energy level which is occupied is $E_F - \Delta_0$. The value of Δ_0 is a characteristic of the material. At temperatures above absolute zero but below T_c, there is a certain thermal excitation and as a result of this the "bond" is broken for a number of pairs. The "loose" electrons formed have a higher energy than $E_F + \Delta$ and thus lie in the energy band above the gap; they may justifiably be regarded as "normal" electrons in a "sea" of "superconducting" electrons.

If the temperature is increased, the percentage of "normal" electrons increases while Δ decreases. Finally, $\Delta = 0$; this happens when T has reached the transition temperature T_c. An external magnetic field also has a similar effect, since Δ can be made to fall to zero by increasing the magnetic field above the critical value H_c. This value becomes smaller as the temperature increases.

We should also like to point out here that there are limits to the current in a superconductor, even at $T = 0$ and at zero external magnetic field. This may be

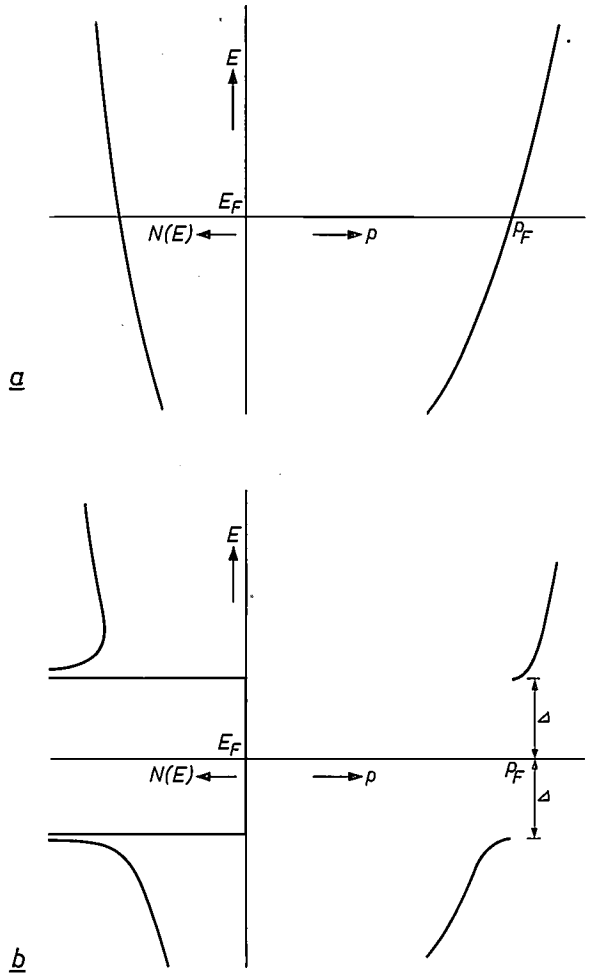


Fig. 3. *a*) The relationship between the energy E and the momentum p of the free electrons in a metal in the normal state is shown on the right. The abscissa axis is drawn at the level of the Fermi energy E_F. The curve of the density of states N against E, around $E = E_F$, is shown on the left.
b) The same for a metal in the superconducting state. $N(E) = 0$ in a zone of width 2Δ around $E = E_F$.

explained as follows: if the drift velocity of an electron — which increases with the current — exceeds a certain value ($2\Delta/p_F$), the energy per electron becomes greater than $E_F + \Delta$ and transitions to states of individual electrons with an energy $E_F + \Delta$ are possible. In other words, electrons forming a pair can then separate and once more become "normal". At a higher temperature or if there is a magnetic field, this situation occurs at a lower velocity, i.e. at a lower current.

Why, now, is the electrical resistance of a metal to which fig. 3*b* applies zero? Although it is very difficult to answer this question properly in words, i.e. without bringing in the quantum-mechanical treatment of the problem, we can at least say the following. The transition from the situation of fig. 3*a* to that of fig. 3*b* (the transition of the electrons from the free to the paired state) must be regarded as a "condensation" of the electron gas. Excitation of individual electrons is not now required for a current to flow: a current must

be regarded as a state of the condensate as a whole. This state exhibits perfect persistence because transition to an energetically lower state is impossible either because there is no such state or because the probabilities of transition are infinitely small; both cases will be discussed later. The transfer of energy to the lattice is therefore impossible and there is no development of heat and no resistance.

In the quantum-mechanical treatment of the superconducting state, the condensate of electrons corresponding to the above is described by one single wave function $\psi(r)$; here, r is the co-ordinate of position. Since $\psi(r)$ is a complex function, it can also be written in the form $\psi(r) = |\psi(r)| \exp j\varphi(r)$, i.e. as the product of amplitude and phase factors. The amplitude factor is in general independent of r and can be related to the concentration of the superconducting electrons.

The way of looking at the transition from the normal to the superconducting state as a condensation, and therefore as a phase transition, corresponds to the fact that it occurs at a sharply defined temperature, just as in the well-known phase transitions vapour → liquid and liquid → solid.

The temperature T_c and the width $2\Delta_0$ of the gap at absolute zero are proportional to each other. Bardeen, Cooper and Schrieffer, extending the theory of Fröhlich, have calculated that T_c and Δ_0 depend on the vibration spectrum of the lattice — particularly on the Debye temperature Θ_D —, on the density of states N of the electrons in the normal state close to E_F and on an interaction parameter V, according to the relation:

$$2\Delta_0 = 3.52\, kT_c \approx 3.52\, k\Theta_D \exp \frac{-1}{N(E_F)V}. \quad (1)$$

In many cases the experimental results could be described by a relation of this form.

An interesting implication of equation (1) is the "isotope effect". In this effect, the transition temperature of isotopes of the same element depends upon the mass M of the nucleus. Since Θ_D is inversely proportional to $\sqrt{M}$, the same must apply for the transition temperature. This effect had in fact already been found experimentally in 1950 [6].

In metals which have a slightly more complex band structure in the normal state (e.g. the s and the d band overlapping), the situation can be different from that shown in fig. 3*b*. Recently, cases have been found in which the drop in the density of states around $E = E_F$ is far less marked than the figure shows, and also the isotope effect is not found in all metals. Furthermore, it seems that the interaction between the electrons need not necessarily take place through the lattice vibrations. Thus the question as to what direction to take in looking for a superconducting material, i.e. the question of the key to the "electron engineering" of superconductors, has still not been answered satisfactorily.

Two kinds of current

To assist the understanding of what follows, we should point out here that we have to distinguish between two kinds of current in superconductors, the Meissner currents and the transport currents. In the first case, the total charge transport through every cross-section of the superconducting body is equal to zero, just as it is for an eddy current in an ordinary conductor; a Meissner current flows in the presence of a magnetic field. With the transport current, there is in fact a net charge transport through a cross-section, as in a normal conductor connected to a voltage source. We shall first of all discuss the connection between current and magnetic field.

London's first equation

We shall now show that London's first equation, which describes the connection between current and magnetic field for a superconductor, may be directly derived from the basic property of a pair of electrons, the property that its electromagnetic (or generalized) momentum P is zero [7]:

$$\mathbf{P} \equiv 2\, m\mathbf{v} + 2\, e\mathbf{A} = 0. \quad \ldots \quad (2)$$

Here m and e are the mass and the charge of an electron, $\mathbf{v}$ the velocity of the centre of gravity of the pair and $\mathbf{A}$ the vector potential. The occurrence of the term $e\mathbf{A}$ is a result of the presence of the magnetic field $\mathbf{H}$ whose relationship to the vector potential is given by:

$$\mathbf{H} = \operatorname{curl} \mathbf{A}, \quad \ldots \quad (3)$$

which, if $\mathbf{A}$ is to be uniquely determined, requires that:

$$\operatorname{div} \mathbf{A} = 0. \quad \ldots \quad (4)$$

If we now reduce (2) to

$$\mathbf{v} = -e\mathbf{A}/m, \quad \ldots \quad (2')$$

we see that the local value of $\mathbf{v}$ is proportional to that of the vector potential. From this we can derive the following. First of all, the normal component of $\mathbf{v}$ at the surface of a superconductor will be zero, and the same must therefore apply to the normal component of the vector potential, in agreement with the requirement $\operatorname{div} \mathbf{A} = 0$. Furthermore, using (2'), we can derive an expression for the current density $\mathbf{j}$ in a superconductor by substituting the expression found for $\mathbf{v}$

[6] E. Maxwell, Phys. Rev. **78**, 477, 1950; C. A. Reynolds, B. Serin, W. H. Wright and L. B. Nesbitt, Phys. Rev. **78**, 487, 1950.

[7] For London's own calculations, see his book [2].

in the relation $\mathbf{j} = ne\mathbf{v}$ (n is the concentration of the electrons):

$$\mathbf{j} = ne\mathbf{v} = -ne^2\mathbf{A}/m, \quad \ldots \quad (5)$$

which can be reduced by (3) to:

$$\text{curl } \mathbf{j} = -ne^2\mathbf{H}/m. \quad \ldots \quad (6)$$

This is London's first equation for a superconductor and takes the place of Ohm's law. (London's second equation relates to non-stationary states and does not need to be taken into account in this article.)

Just as (5) is a direct consequence of $\mathbf{P} = 0$, so can (6) be regarded as a direct consequence of the equation:

$$\text{curl } \mathbf{P} = 0. \quad \ldots \quad (7)$$

This is the most general expression for London's first equation.

The current flowing in a superconductor in the presence of a magnetic field should be regarded as an aspect of the ground state of the condensate, which is determined by the condition $\mathbf{P} = 0$. The application of the magnetic field only causes a slight shift in the energy value appropriate to the ground state, in exactly the same way as, say, the energy value of the electrons in the shells of an atom is shifted slightly when a field is applied.

We shall not deal with the quantum-mechanical treatment of this, but we should like to point out that the condition $\mathbf{P} = 0$ implies that the wave function $\psi(r)$ describing the condensate has no spatial variation: wave mechanics show that the gradient of ψ is, in fact, proportional to the momentum $\mathbf{P}$ and is therefore zero if $\mathbf{P} = 0$. The "wave" represented by the wave function is here one of constant amplitude (see above) and an infinitely long wavelength.

The Meissner effect; the penetration depth

The current just mentioned, which flows in a superconductor when a magnetic field is applied, has a very marked spatial variation, and the same is true of the magnetic field inside the superconductor. This will be seen directly if London's equation (6) is combined with the Maxwell equation:

$$\text{curl } \mathbf{H} = \mathbf{j}, \quad \ldots \quad (8)$$

which also applies in a superconductor. It is then found that:

$$\nabla^2\mathbf{H} = \mathbf{H}/\lambda^2, \quad \ldots \quad (9)$$

where

$$\lambda = (m/ne^2)^{1/2}. \quad \ldots \quad (10)$$

The parameter λ has the dimension of length. It is a "characteristic length", which will always appear in the solution to (9) whatever the boundary conditions, and is generally referred to as the *penetration depth*. If the values applying in a normal metal are substituted for m and n in (10), we find for λ a length of the order of only 100 nanometres (1000 Å).

The nature of the value λ and the meaning of the name "penetration depth" become clear if we look at the simple situation sketched in *fig. 4*. Here, a superconductor and a non-superconductor are separated by a plane but otherwise extend to infinity. In the non-superconductor there is a magnetic field of strength H_0, parallel to the plane of separation. If we now examine the way in which the strength H of the magnetic field in the superconductor varies as a function of the distance x from the plane of separation, we find:

$$H(x) = H_0 \exp(-x/\lambda). \quad \ldots \quad (11)$$

The magnetic field strength has thus already dropped to about a third of H_0 at the depth $x = \lambda$. (A solution of (9) like that of (11) but with a positive exponent can also be found, but this has no physical significance here.)

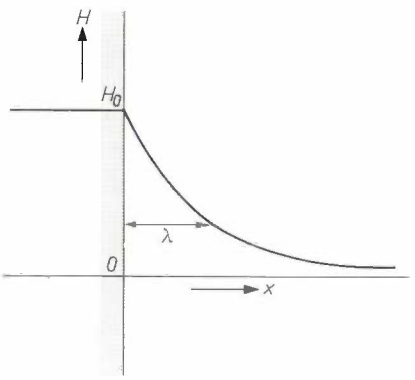


Fig. 4. The penetration depth λ characterizes the distance that a magnetic field can penetrate into a superconducting body. The strength H of this field has the value H_0 outside the body.

In a more complicated case, e.g. that of a cylinder of circular cross-section, the solution is also more complicated, but it is always found that H decreases as the depth inside the superconductor increases. The magnetic field inside a superconductor is reduced with respect to the external field by the action of a current. The spatial distribution of this "screening" or Meissner current satisfies an equation which, like (9), follows from (6) and (8) and has the same form as equation (9):

$$\nabla^2\mathbf{j} = \mathbf{j}/\lambda^2. \quad \ldots \quad (12)$$

Thus the Meissner current has the same depth of penetration as the magnetic field. The interior of a superconductor is field-free because the external field is exactly compensated by the magnetic field of the screening current.

The state of field exclusion also sets in spontaneously in a metal which is cooled in a magnetic field to below the transmission temperature. This very important feature was first established by Meissner and Ochsenfeld in 1933. It demonstrates that a superconductor differs fundamentally from a conductor of zero resistance and has pointed the way to the understanding of the superconducting state as a phase.

The transport current through a superconductor; the persistent current

The currents discussed above were reaction currents — or magnetization currents — brought about by the external field. It was tacitly assumed that we were dealing with a singly-connected body, i.e. one in which each path through its interior from a point P to a point Q can be transferred continuously by means of infinitely small changes into any other path between P and Q. If such a body is cut by a plane, the integral of these currents across the cut surface is equal to zero.

In order to prepare for situations which can occur in a multiply-connected body, let us now discuss the case of a superconducting rod with its ends connected to a battery by copper wires. A transport current can thus be passed through this rod, and the integral of the current across a cross-section is now not equal to zero. The forces which form the electron pairs remain effective, but the value of the generalized momentum is, in this case, no longer equal to zero:

$$\mathbf{P} = 2m\mathbf{v} + 2e\mathbf{A} \neq 0, \quad \ldots \quad (13)$$

as we shall explain in a moment.

A superconductor through which a transport current is flowing does not remain in the ground state. In this case the superconducting condensate enters a state of collective excitation without there being any question of excitation by decoupling of the pairs.

Here, too, the starting point for the treatment of the electrodynamic behaviour is London's equation of the form curl $\mathbf{P} = 0$. It follows from this equation that even though the field is now due to the flow of injected current, the current and field distribution are still determined by (9) and (12), and that, in particular, the concept of penetration depth retains its significance. Like the Meissner current, the transport current flows in a thin skin at the surface of a superconducting rod. There is neither current nor field within it but the vector potential $\mathbf{A}$ now has a value different from zero. The integral from which $\mathbf{A}$ can be calculated by means of Maxwell's theory:

$$\mathbf{A} = \int (\mathbf{j}/r)\, \mathrm{d}\tau. \quad \ldots \quad (14)$$

(where $\mathbf{j}$ is the current density in the volume element $\mathrm{d}\tau$ at a distance r from the point where $\mathbf{A}$ is required), now gives a value different from zero. This explains why $\mathbf{P}$ cannot be zero.

A special case of the collective excitation of the superconducting condensate noted above is the persistent current in a superconducting ring, i.e. a current which is induced in the ring and continues to flow unattenuated as long as the ring remains superconducting. Here again, the basis for an electrodynamic approach is one of the equations curl $\mathbf{P} = 0$ or curl $\mathbf{j} = -ne^2\mathbf{H}/m$, but, at a given external magnetic field, there are now several solutions for the currents in the superconductor, and not just one, as was the case for a singly-connected superconducting body.

As with the flow of a transport current in a superconducting rod connected to a current source, the flow of a persistent current in a ring must be regarded as a situation in which $\mathbf{P} \neq 0$. In the wave-mechanical approach given above, this means that the collective wave function will in this case have a spatial wave-like variation: if $\mathbf{P}$ is not zero, φ varies uniformly with the co-ordinates and ψ does have the character of a periodic function. We shall now see that this has rather remarkable consequences for the magnetic flux enclosed by the ring.

Flux quantization

A ring, like any other doubly-connected body, is characterized for superconductivity by the central region (the "hole") in which $\psi = 0$. The superconducting body has been pierced by either a hole, or a tube of normal material. In the latter case, the normal material need not be chemically different from the supercone ductor, a magnetic field in the central region could havbeen applied to remove the superconductivity. With this topological structure it is possible that $\mathbf{P} \neq 0$ without there being any external current source, but the related spatial dependence of φ now poses a problem connected with the uniqueness of ψ. If we assume that ψ extends over macroscopic distances, the ring must be a whole number of wavelengths along its circumference, just as it is for the path of an electron in an atom. As in atomic physics (Bohr, Sommerfeld), this is the case if the "phase integral" $\oint \mathbf{P}\mathrm{d}l$ is a whole multiple of h, Planck's constant:

$$\oint \mathbf{P}\mathrm{d}l = nh. \quad \ldots \quad (15)$$

Here, $\mathrm{d}l$ is an element of the integration contour and n is an integer. It can be shown directly from (15) and (3) (see below) that the magnetic flux Φ passing through the surface of a cross-section is therefore the n-times multiple of a flux quantum Φ_0, equal to $h/2e$:

$$\Phi = nh/2e = n\Phi_0. \quad \ldots \quad (16)$$

The flux quantum Φ_0 is about 2×10^{-7} gauss cm^2. This

is also true for the flux enclosed in a normal patch or cylinder within a superconducting object.

The persistent current in a doubly-connected body, which we have come to recognize as the form in which the excitation of the superconducting collective appears, is a quantum state which apparently makes the transition to the ground state only with great difficulty. No cases are known of transitions under interaction with an electromagnetic radiation field, i.e. with a single photon, like those which occur in atomic systems. Such a photon would in fact have to possess tremendous energy in a real superconducting system: with a persistent current of 100 A, a quantum jump $\Delta\Phi = \Phi_0$ corresponds to an energy jump of about 1 MeV. The extremely remote probability of such a transition is undoubtedly partly connected with the enormous disparity between the dimensions of the ring and the wavelength of the radiation (about 10^{-3} nm, or 10^{-2} Å).

Flux quantization is not a purely theoretical concept, but has also been demonstrated experimentally [8]. However, it is still, of course, an open question whether the relation (15) for the phase integral also applies if the integration contour is very long, as, for instance, in a short-circuited coil consisting of a superconducting wire a few miles long.

Later we shall discuss how flux can disappear from a superconducting ring; as far as we know, this is always a process of escape at the surface of the superconductor, in which in point of fact separate flux quanta are involved and where energy is converted into magnetic field energy elsewhere or into vibrational energy of the lattice.

Experimental determination of the width of the forbidden zone

The change in the energy spectrum that the material undergoes on cooling to below T_c, as discussed above, may be found experimentally in various ways. We should like to discuss two of them here.

A fairly direct method is the study of the absorption of a high-frequency electromagnetic field directed against a superconducting plate: for example in a microwave resonant cavity [9]. As the metal cools the normal resistance drops and so therefore does the depth of penetration of the electromagnetic field [10], and the absorption, but electromagnetic energy will always be dissipated as long as the electrical component of the electromagnetic field can interact with normal electrons. Even in the superconducting state this is still possible, because the alternating field still meets electrons that are normal — i.e. electrons unpaired because of thermal excitation — in the thin surface layer into which it penetrates. In the extreme case, however, of infinitely low temperature, this absorption would finally have to decrease to zero unless the energy $h\nu$ of the wave packets of the alternating field is itself high enough to decouple a pair of electrons and bring them across the forbidden zone to an excited state. This effect gives us the chance of accurate spectroscopic determination of the energy gap. The method is therefore quite comparable to the one used in semiconductor research, although there we find the fundamental absorption at wavelengths which are a few orders of magnitude smaller. *Fig. 5* gives the result of the classical experiments of Biondi and Garfunkel: the surface impedance (as a measure of the absorption) of a small plate of aluminium is plotted here in relative co-ordinates as a function of the value of $h\nu$ of the energy quanta of the microwave field (in units of kT_c) with the temperature as a parameter. It can clearly be seen, particularly at a relatively low temperature ($T \ll T_c$), that absorption sets in as soon as the quantum energy is higher than about 3.5 kT_c. A different value is found for some materials.

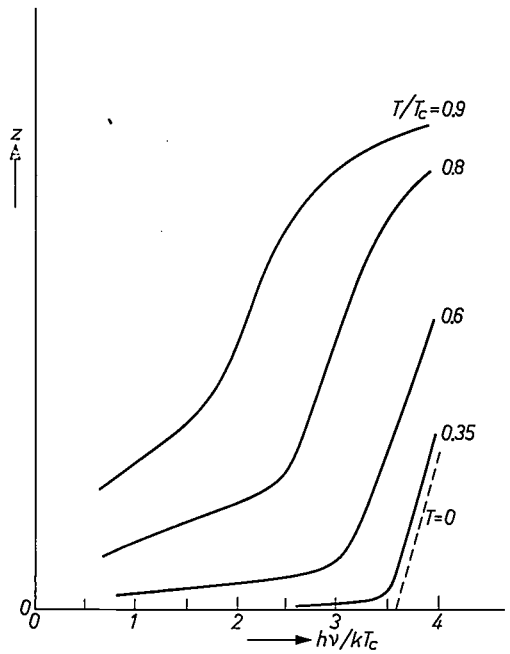


Fig. 5. The width 2Δ of the forbidden zone in the energy spectrum of the "superconducting" electrons can be determined by finding out how the absorption of microwaves in the surface layer of the superconductor depends on the quantum energy of the incident radiation. The surface impedance z (a measure of the absorption) is plotted against the quantum energy in units of kT_c. At very low temperatures absorption suddenly sets in at $h\nu/kT_c = 3.5$.

A particularly elegant method of spectroscopy — this is the second method that we shall discuss here — is that due to Giaever in which a tunnel diode is used [11]. Two metals are separated by a very thin non-metallic layer (a few nm thick). Electrons can pass through this skin from one metal to the other by a tunnelling process, that is to say, they do not need to make use of the high, thermally inaccessible conduction band of the intermediate material (usually an oxide); on the contrary, during the transition, they are located in the forbidden zone of the energy-level diagram of that material. Now, with this kind of contact the same thing happens as with a true contact: the Fermi levels of both plates reach exactly the same value after the exchange of a few electrons. If, however, one of the metals is a superconductor, the electrons on opposite sides are not "on speaking terms", that is to say, electrons that would be prepared to cross cannot find an available energy level on the other side (*fig. 6*), again particularly in the extreme case of infinitely low temperatures. A relative displacement of the level diagrams can be brought about by applying an electric voltage V, and this is possible because of the comparatively good insulation of the non-metallic

layer. A crossover of charge is then possible if $V \geqq \Delta/e$. The energy gap may therefore be read off immediately from the I-V characteristic of this Giaever diode as the double-threshold voltage. Experiments of this type also show that, in most cases, the energy gap is about $3.5\ kT_c$.

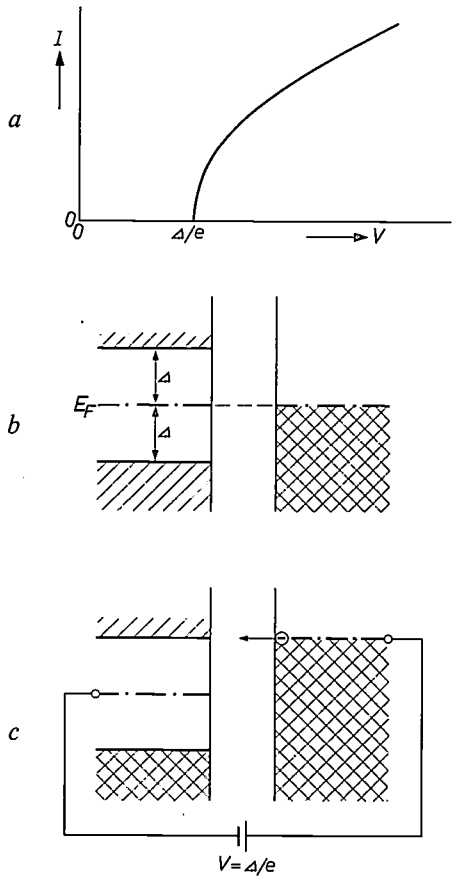


Fig. 6. Characteristic (*a*) and energy band diagrams (*b*, *c*) of a Giaever tunnel diode for determining Δ. A superconducting plate (on the left) and a normal plate of the same metal are separated by a layer of oxide so thin that the electrons can cross from one plate to the other by means of the tunnel effect. If the applied voltage V is lower than Δ/e, no current can flow because there are no energy levels (*b*) at the required height in the superconductor; if $V \geqq \Delta/e$ such levels are present (*c*).

Magnetization; superconductors of the second kind

Disruption of the superconducting ordering by a magnetic field

Up till now we have not dealt with the question of what exactly happens when the strength of an external magnetic field in which a superconductor is located is allowed to increase until it exceeds the value H_c at which superconductivity ceases. Neither have we enquired into the exact situation in the surface layer of the superconductor into which the field has penetrated.

If a magnetic field penetrates a superconductor, even if only to the penetration depth, this signifies an attack on the superconducting state. If the field is too high, the transition to the normal state takes place but, even where the superconducting state is retained, there is already an effect, in the form of a local reduction in the concentration n_s of the superconducting electrons, i.e. a reduction in the width 2Δ of the gap in the energy diagram. As we mentioned earlier, an increase in temperature also does this and it is known that both effects act on the penetration depth λ. There is therefore an interaction between n_s and H (or A), which means that London's equation does not hold in the regions penetrated by the magnetic field. A solution in this region can be found from the equations due to Landau and Ginzburg [12]. We shall not deal in detail with these equations but simply point out that they are two equations which can be used to find n_s and the vector potential **A** as functions of the position co-ordinates, provided that the boundary conditions are not too complicated. (Strictly speaking, the equations do not contain n_s, but a complex order parameter which can be identified with the wave function ψ mentioned earlier.) The coefficients which appear in the equations are characteristic of the material. The fact that there are now *two* equations leads directly to the possibility of more than one characteristic penetration distance. And in fact, besides the penetration depth λ for the magnetic field and the current (already given by London's theory), we now also find a characteristic penetration depth for the disturbance in the configurations, i.e. for the variation in $|\psi|$. This characteristic penetration depth is called the coherence length and is indicated by ξ. The term coherence length is perhaps a little confusing as "coherence" does not relate here to the maintenance of the phase coherence of the wave function, which as we have seen is maintained over a considerable distance, but to the "stiffness" of the wave function, in the sense that $|\psi|$ cannot vary to any great extent within the distance ξ. Like λ, ξ is generally small: we find values for ξ of the order of 10 to 1000 nm.

It will appear presently that it is of fundamental importance for the behaviour of a superconducting body in a magnetic field whether the penetration depth λ or the coherence length ξ is the greater. More precisely, the important question is whether λ/ξ is greater or smaller than $1/\sqrt{2}$. This can easily be shown to be reasonable from considerations of the surface energy. We shall therefore begin by discussing the main prin-

---

[8] R. Doll and M. Näbauer, Phys. Rev. Letters **7**, 51, 1961; B. S. Deaver, Jr. and W. M. Fairbank, Phys. Rev. Letters **7**, 43, 1961.

[9] M. A. Biondi and M. P. Garfunkel, Phys. Rev. **116**, 853 and 862, 1959.

[10] See H. B. G. Casimir and J. Ubbink, The skin effect I, Philips tech. Rev. **28**, 271-283, 1967 (No. 9), particularly equation (10).

[11] I. Giaever, Phys. Rev. Letters **5**, 147, 1960.

[12] V. L. Ginzburg and L. D. Landau, Zh. eksper. teor. Fiz. **20**, 1064, 1950. Important contributions have been made to the elaboration of this theory by L. P. Gorkov and A. A. Abrikosov. For complete bibliography see Lynton's book [2].

ciples of the thermodynamic treatment of the phase transition superconducting ⇄ normal.

Thermodynamics of the phase transition superconducting ⇄ normal

In the thermodynamic treatment of the phase transition superconducting ⇄ normal [13] use is made of the Gibbs' free energy G. Let us suppose that this has the value $G_s(H)$ in the superconducting phase with the value $G_n(H)$ in the normal phase. Now there is always some freedom of choice in the expression for the thermodynamic quantities when a magnetic field is present, as it is a question of taste whether all or only a part of the field energy is to be assigned to the material in which the field exists. The choice that we shall make is such that for an isothermal change:

$$dG = -M\,dH. \quad . \quad . \quad . \quad . \quad . \quad (17)$$

Here M is the magnetization of the body in the external field H, assumed to remain undisturbed. If the field is gradually increased in strength, the field intensity H_c at which the phase transition takes place must satisfy:

$$G_s(H_c) = G_s(0) - \int_0^{H_c} M\,dH = G_n(H_c). \quad (18)$$

We now assume that the superconducting body is perfectly diamagnetic, i.e. B is zero (according to Maxwell's theory, B actually has the significance of the average microscopic magnetic field intensity). The macroscopic magnetization M due to the screening current is then equal to $-H$, since $\mathbf{B} = \mathbf{H} + \mathbf{M}$. If, further, we assume that the magnetic susceptibility of the material in the normal state can be neglected, we can reduce (18) directly to:

$$G_n(0) - G_s(0) = \int_0^{H_c} H\,dH = \tfrac{1}{2}H_c^2. \quad . \quad (19)$$

To summarize, in the absence of a magnetic field, the free energy in the superconducting state is lower than that in the normal state; the value of H_c is determined by the difference.

The reduction of the free energy in the superconducting state with respect to that in the normal state can also be derived from the change in the energy spectrum of the electrons, in the following way. As we have seen, in the superconducting state, the electrons in the energy interval Δ just below the Fermi level E_F are accommodated at lower levels where there is a sufficiently high density of states available (fig. 3*b*). If $N(E_F)$ is the density of states at the Fermi level of the metal in the normal state, then we must have:

$$G_n(0) - G_s(0) = aN(E_F)\Delta_0^2. \quad . \quad . \quad (20)$$

Here, a is a coefficient of the order of unity, which depends on the precise form of the density of states in the superconducting state. If we compare (19) with (20), we see that H_c is proportional to Δ_0 and from (1) that H_c must therefore also be proportional to T_c.

When dealing with a phase transition, it is necessary to take into account the surface energy. In our case, this means the supplementary energy which results from the presence of a boundary between the superconducting and normal parts of the metal. Two contributions can be distinguished. First, there is a correction to the magnetic energy term $\int M\,dH$ as a result of the fact that the magnetic field penetrates the superconducting material to a distance λ. The field-free volume is thus slightly smaller than the volume of the superconductor itself by an amount λS in which S is the area of the boundary surface and the free energy is therefore smaller by an amount $\frac{1}{2}\lambda SH^2$ than the value taken into account above. This correction is therefore *negative.*

The second correction, on the other hand, which is a result of the disturbance of configuration at the surface, is *positive.* This correction term varies monotonically with ξ, the coherence length, and is dominant if ξ is sufficiently large. The surface energy is then positive and the boundary surface will become as small as possible or disappear entirely. If, on the other hand, ξ is relatively small, the magnetic term in the surface energy is dominant. When a magnetic field is applied, the boundary surface between superconducting and non-superconducting material then attempts to increase in area. In a body that is already totally superconducting, this can only happen through the onset of dispersion: a state arises in which the superconductor contains a large number of normal regions, whose total volume is very small, but whose total area is very large, and in which there is a magnetic field, as in the bordering penetration layers.

Two kinds of superconductors

The above thermodynamic considerations show that the picture of a superconductor given in the previous section is true only for materials where ξ is relatively large. In materials with a small ξ, the Meissner effect (i.e. almost perfect diamagnetism) does not occur. According to the theory of Landau, Ginzburg and Abrikosov, the first case is encountered when $\varkappa = \lambda/\xi$ is smaller than $1/\sqrt{2}$, and the other if $\varkappa$ is larger than $1/\sqrt{2}$. We speak of superconductors of the *first kind* (*with* Meissner effect) and superconductors of the *second kind.* The elements lead, tin and mercury belong to the first group and niobium, together with several alloys, to the second.

The superconductors of the second kind are the

ones which have received most attention in the last few years and for the moment offer most promise for technical applications. In alloys, ξ is usually small as a result of the effect of the mean free path of the electrons on the "stiffness" of the collective wave function (or order parameter) ψ. Gorkov has derived:

$$\varkappa = \varkappa_0 + b\varrho''. \qquad (21)$$

Here, $\varkappa_0$ is the $\varkappa$ value of the perfect lattice, i.e. the lattice of the appropriate material at $T = 0$ and with no crystal imperfections, and b is a coefficient which depends on the density of states of the electrons at the Fermi level in the normal state. The quantity ϱ'' is the residual resistance, i.e. the value that the resistivity of the material would have in theory at $T = 0$ if no superconductivity occurred. The value of ϱ'' increases with the number of irregularities in occupancy or other lattice defects in the metal crystal, which reduce the mean free path.

A superconductor of the second kind has a magnetization curve like the one shown in *fig. 7*. At small values of the applied field there is still an effective screening

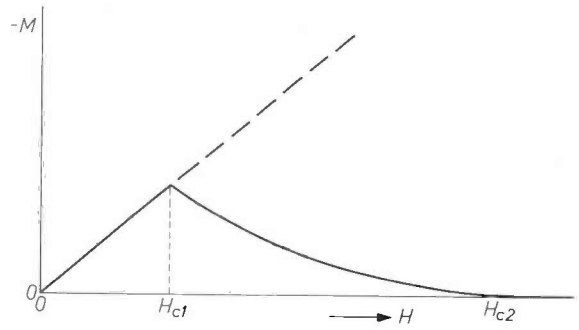


Fig. 7. Magnetization curve of a superconductor of the second kind. If the field H is increased, the magnetization M also increases at first, but begins to drop at the value H_{c1}. At $H = H_{c2}$, M has become equal to zero and the material reverts to the normal state.

current, but when the field exceeds a characteristic value H_{c1}, flux begins to penetrate. As we noted, there is then no longer any Meissner effect and we speak of the *mixed state*. The flux invasion takes place in the form of millions of very small flux "threads" or vortices. These vortices are very interesting entities, their main features being (see also *fig. 8*):

1) They can terminate only at a surface, or be closed on themselves.
2) They carry persistent and circular currents flowing over their entire length as current walls.
3) Inside them, and coupled to them, there is a magnetic field; current and magnetic field extend to a distance λ from the cylinder.
4) The flux threads have a non-superconducting core, a very thin cylinder of radius ξ.
5) The vortex is in fact an excited state of the persistent ring current type ($\mathbf{P} \neq 0$) that has already been discussed, and carries only a single flux quantum $h/2e$.

In the mixed state a part of the material — a very large part just above H_{c1} — is still in the superconducting state. Because of the flux penetration, this situation can be retained to a much higher field intensity than ever could be the case if the pure Meissner effect were maintained. The completely normal state is finally attained when the concentration of flux threads is so great that the entire volume of the material is filled by the normal cores. The value of H at which this happens is indicated by H_{c2}; this value increases with decreasing ξ. In various niobium-containing superconductors, values between 100 000 and 200 000 gauss are found for H_{c2}.

We may also note that for superconductors of the second kind, the area beneath the magnetization curve (fig. 7) is equal to the condensation energy (cf. eq. 19).

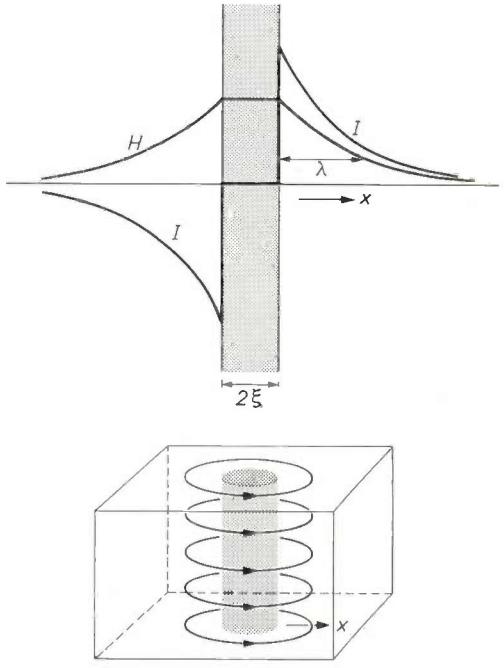


Fig. 8. Current and field curve in a longitudinal cross-section of a flux thread. The flux is trapped in a normal zone (of radius ξ) which has a circular current flowing at the wall. The magnetic field penetrates in the normal way into the surrounding material to a distance λ.

Induction effects

In the light of what has just been said, we should now like to take another look at the behaviour of a superconducting body in a magnetic field, this time allowing the body to be multiply-connected and the flux to be displaced. We shall deal with various cases in this section.

[13] C. J. Gorter and H. B. G. Casimir, Physica **1**, 306, 1934. This treatment will also be found in the book by Shoenberg [2].

Let us first take a ring or hollow cylinder of lead or other superconducting material of the first kind. When placed in a magnetic field, the ring will persist in the initial state: as long as it remains completely superconducting, the enclosed flux will not change (e.g. it will remain zero). In the interior of a cylinder which is long enough, the *field*, too, will not change (e.g. it will remain zero). Once the critical field strength is exceeded the field will, of course, be able to penetrate everywhere but now removal of the field will no longer result in the initial state. A quantity of magnetic flux has now been enclosed, and this can be supported by a persistent current.

A variation on this theme is found in Buckingham's "persistatron". A superconducting ring is asymmetrically connected to two current wires (*fig. 9*). If a current I, which is high enough to cause the ring to revert to the normal state at some point, is made to flow through the wires, the flux through the ring will be maintained when the current I is switched off, and supported by a persistent current in the ring.

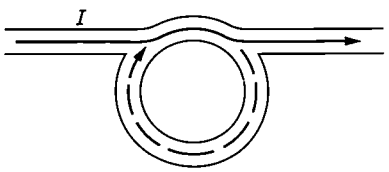


Fig. 9. Buckingham "persistatron". A current I strong enough to remove superconductivity locally is passed through part of a superconducting ring. The magnetic flux contained by the ring during the flow of I is enclosed when I is switched off.

In an apparently singly-connected body, too, flux can be trapped in such a way that an irreversible magnetization curve is produced. This will be the case if, for example, because of an unfavourable shape, part of the body becomes superconducting again when the magnetic field decreases and if that part is doubly connected. Now, flux can no longer escape from the superconducting ring thus formed and, when the field is switched off, there remains a normal central zone in which there is flux, thus effectively making the body no longer singly-connected.

We shall now direct our attention towards the process of *flux creep*. In the discussion of the magnetization of a superconductor of the second kind, we have already referred to the penetration of flux threads deep into the material. Without going in too great detail into the question of how these are released from the surface — they cannot arise spontaneously within the body — we shall now, in a very general manner, examine the significance of the transfer of flux across the superconducting current path in terms of electromagnetic induction. Let us therefore consider the situation in *fig. 10a*. We assume that there is a ring R with a zone P where the material is not superconducting, this zone being of such a size that it does not cut right through the superconducting ring. We also assume that this normal zone P encloses a flux Φ. This flux can be shifted, together with the zone, with the aid of a movable external auxiliary magnetic field (the "displacement field"). The flux Φ can thus be moved across the superconducting ring and, when the edge has been passed, the flux is enclosed within the hole in the ring. If the force which has taken the flux inwards is removed (i.e. the auxiliary field switched off), the flux is maintained by a persistent current in R. We may now consider the zone P with the moving flux Φ as being the source of a voltage V which has brought about an increase I in the persistent current in the ring of inductance L:

$$\int V\mathrm{d}t = LI = \Phi. \quad \ldots \quad (22)$$

If there is a regular quasi-continuous transfer of a large number of such flux tubes, then

$$V = \mathrm{d}\Phi/\mathrm{d}t \quad \ldots \quad (23)$$

is the average flux creep per second. If flux begins to move in the opposite direction, the current will decrease.

A very interesting case is the one in which the flux is set in motion under the influence of a current I_R already flowing in the superconductor, which "washes around" the patch of flux (fig. 10*b*). As a careful analysis of the total magnetic energy of the system as a function of a lateral shift of the zone has shown, such a current exerts a lateral force F_d on the tube of flux. This force is proportional to Φ and the current density. The negative induced voltage connected with the movement of the flux can be best interpreted in this case as *resistance* in the circuit. The situation is fully comparable to that in an ordinary type of d.c. dynamo; this can, of course, be operated as a current generator or as a motor according to the direction of energy flow, and when

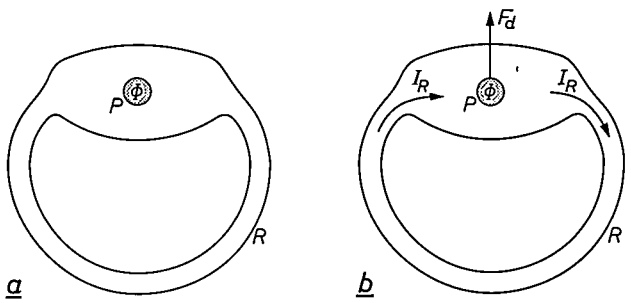


Fig. 10. *a*) Diagrammatic representation of a doubly-connected body (R) in which there is a non-superconducting zone (P) which does not disturb the doubly-connected nature of the superconducting region and encloses the flux Φ. If P is shifted from the outer to the inner edge of R, the flux contained by R can be increased by Φ.
b) If the location of P is not fixed and a current I_R flows in R, P will move across the direction of flow because the flow exerts a lateral force F_d on the tube of flux. F_d is proportional to Φ and to the current density at that point.

operated as a motor it appears in the circuit as a positive resistance.

The lateral force on a tube of flux is extremely important in all kinds of processes in superconductors. It is sometimes referred to as a Magnus force, as the effect shows a certain similarity to the Magnus effect in aerodynamics, but the force exerted on the flux can equally well be regarded as an electrodynamic force as described by Ampère or as an effect of the Lorentz force on the electrons. It is this force that helps to bring about the mixed state when a superconductor of the second kind is magnetized. When, as soon as H has become greater than H_{c1}, vortices begin to split off from the screening current (Meissner current) initially excited at the surface of the body, this force ensures that these vortices are evenly distributed through the superconductor. This follows because with an uneven distribution of the vortices the total current in the interior is not zero everywhere, as may be seen from *fig. 11*, and they will thus be subject to a restoring force. If the vortices are evenly distributed, this total current is zero and no change in the distribution can occur.

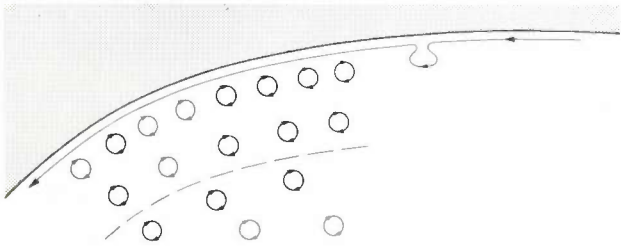


Fig. 11. What happens in a superconductor of the second kind as the magnetic field H increases, if $H_{c1} < H < H_{c2}$ (mixed state). Vortices containing a certain flux split off from the screening current at the surface. The free vortices attempt to distribute themselves as evenly as possible over the available space and thus initially move from the surface to the interior. The movement is produced through the agency of a force which is the result of the total current which flows locally where there is an uneven distribution; at the broken line, for instance, the current flows in five vortices to the left and only in three to the right.

Superconducting d.c. dynamos

The induction effect discussed above gives the basis for the design of superconducting dynamos [14]. These current generators consist essentially of a fixed system of conductors — there are therefore no brushes or commutators — in which both the Faraday induction and the periodic changes in the circuit are produced by an alternating field having the character of a periodic displacement or rotation field. The periodic changes in the circuit are effected by the field periodically bringing certain parts of the superconductor into the normal state — in effect this is a kind of commutator — thus producing a rectifying or unipolar effect. One of the first designs of the superconducting dynamo, which is very closely related to the conventional unipolar dynamo, has already been discussed in this journal [15]. *Fig. 12* gives a quick look at the family of superconducting dynamos, showing the way in which the flux is displaced. If a certain flux Φ is carried round in the dynamo at a frequency f, the e.m.f. is:

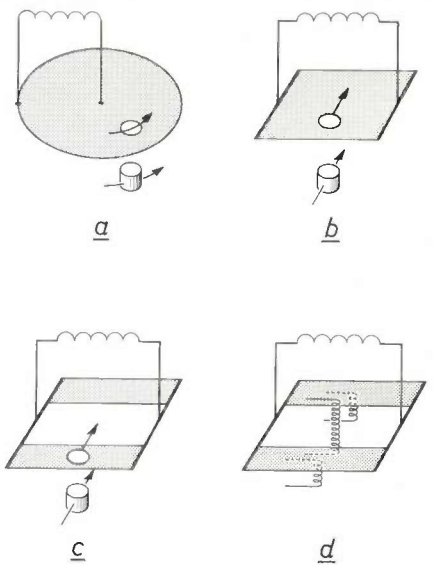


Fig. 12. Examples of the four most important types of design of the superconducting dynamo for the generation of a persistent current in a superconducting circuit.
a) With a rotating field provided by a rotating permanent magnet. *b*) With a displacement field. *c*) and *d*) Variations of (*b*) with a superconductor divided into two strips; in (*d*) the displacement field is obtained by means of coils.

$$V = \Phi f. \qquad (23')$$

There are certain attractive features in superconducting dynamos. They can, for instance, energize superconducting coils where they form closed circuits with these coils, the whole being kept in the bath of liquid helium. This means that no external supply wires are needed and there is therefore no inherent leakage of heat to the helium bath. As long as the dynamo is in operation, the current in the coil increases with time t:

$$I = Vt/L. \qquad (24)$$

Here, L is the inductance of the coil. Once the desired current intensity has been attained, the dynamo is simply stopped, at which it becomes a passive conducting element and the current remains circulating on its own. A persistent current is far more constant than that obtainable even with the best current stabilizers, and this is particularly attractive for precision work. A great advantage is that it is easy to generate very high currents up to thousands of amperes, with a relatively small dynamo. Small variations in the current are obtained simply by rotating the dynamo shaft a few turns [16]. The e.m.f. V with simple dynamo designs like that of *fig. 13* is of the order of 1-10 mV. When superconducting cable of very high current-carrying capacity (e.g. 1000 A) has to be used, this generally implies that the inductance of the circuit will be reasonably small, and so the time required to attain a high current will certainly be acceptable. A disadvantage of these generators is that they are not entirely loss-free. This is because the movement of the flux induces a

[14] J. Volger and P. S. Admiraal, Physics Letters **2**, 257, 1962.
[15] J. Volger, A dynamo for generating a persistent current in a superconducting circuit, Philips tech. Rev. **25**, 16-19, 1963/64. A survey of various possible designs will be found in J. van Suchtelen, J. Volger and D. van Houwelingen, Cryogenics **5**, 256, 1965.
[16] An application where use is made of this property has recently been described in this journal: F. W. Smith, P. L. Booth and E. L. Hentley, Masers for a radio astronomy interferometer, Philips tech. Rev. **27**, 313-321, 1966.

small (local) current in the normal zone, and this of course develops heat. In dynamos for use in large installations, special measures may have to be taken to get over this difficulty. However, the dissipation of perhaps a few watts which occurs in dynamos for energizing laboratory-type superconducting coils will never be prohibitive.

Fig. 13. A superconducting magnet coil (below) complete with superconducting dynamo (above).

Response of a superconductor of the second kind to the passage of a current

As discussed in one of the previous sections, a current flowing in a superconductor exerts a lateral force on a tube of flux passing through it, and the same applies to the elementary flux threads in the superconductor of the second kind when in the mixed state. If there are no opposing forces, i.e. if the flux threads are not fixed at some "anchor point", they will be set in motion. We then have the case of the macroscopically observable voltage determined by induction effects, i.e. the case where the superconductor exhibits a certain resistance. In agreement with the foregoing, we obtain for the voltage drop per unit length:

$$E = uB. \qquad (25)$$

Here, u is the velocity of the transverse movement of the flux threads and B the flux density they produce in the specimen.

There are one or two very convincing experimental indications of the reality of this flux movement in superconductors. Let us note first the measurements of the noise component of E in vanadium which were made by Van Ooijen and Van Gurp [17] at Philips Research Laboratories. These workers found a cut-off frequency in the noise spectrum which was equal to the reciprocal value of the crossing time of the flux threads calculated from (25) and the width of the measuring plate. A spectrum of this type is characteristic of a shot-noise effect, which would be expected here on account of the discrete values of the moving flux. The magnitude of the tubes of flux can be derived from the intensity of the noise signal. Calculation shows that this magnitude must depend upon the value of the primary current, but in the limit of an extremely high primary current, the flux crosses in separate elementary quanta without the formation of groups.

A second and very direct experimental proof has been given by Giaever [18]. He has examined the movement of the flux by bringing two very thin metal plates very close to each other. There was no metallic contact, but the magnetic patterns of the mixed state in both plates were coupled. If a current was then made to flow in one of the plates, it was found that a voltage was generated not only in this plate but also in the other one. The only way in which this voltage could have been excited was by the magnetic coupling between the two plates. This proves that the voltages must be due to the movement of the flux pattern.

The movement of flux has recently been made *visible* in experiments with very thin lead foil at Philips Research Laboratories [19]. A state of dispersion can in fact also occur in superconductors of the first kind, to which lead belongs, and in this case the dispersion is associated with the fact that the internal field strength in a body is in general not uniform, because of demagnetization [20]. This means that when the external field increases, not all the parts of the body change to the superconducting state at the same time. When the body is partly in the normal state and partly in the superconducting state — this situation is known as the intermediate state — the normal regions carry a flux which is much greater than one flux quantum. However, in the movement of flux under the influence of a current these regions are very similar to the flux threads of the mixed state. Their movement has been made visible with the aid of small grains of niobium: when the bundles of flux move through the lead foil, the niobium grains follow them. This has been observed and recorded with the aid of a specially designed microscope and a closed TV circuit (*fig. 14*).

Fig. 14. Experimental arrangement for visual demonstration of the movement of magnetic flux in lead in the intermediate state. Inside the helium cryostat, which can be seen in the middle of the picture, there is a special microscope, focused onto a lead foil which has been coated with niobium powder and which is placed in a variable magnetic field. (This field is provided by a superconducting coil.) A television camera is mounted above the cryostat, and part of the illuminating system of the microscope can be seen just to the left of the operator's hand. In the left background there is a television recorder, used for recording the movement of the niobium grains. The dark patches on the TV screen are niobium grains or small groups of niobium grains (magnified 10 000 diameters). The crazing in the lighter areas of the pattern is due to a lacquer coating applied to the lead to prevent oxidation.

The resistance in the mixed state

The voltage drop in a specimen of a superconductor of the second kind in the mixed state with a current flowing in it is thus determined by the rate of flux drift. The question of the factors determining this rate has been examined in detail in the last two years. The picture that we now have, which to some extent has been derived from the work at Philips Research Laboratories [21], is rather as follows. The flux threads are affected by a driving force with a value F_d per unit length, given by:

$$F_d = j\Phi. \qquad (26)$$

In the stationary state, the flux threads have a constant velocity, i.e. there is a frictional force F_f per unit length which exactly compensates F_d. As we mentioned above, the force F_d can be regarded as an electrodynamic force as described by Ampère. This also applies to the frictional force, at least where the braking is caused by the eddy currents generated by and at the moving flux threads. Looked at locally, $dB/dt \neq 0$ because of the movement of the flux thread, and there is thus a rotary electric field generating small eddy currents. It is

[17] D. J. van Ooijen and G. J. van Gurp, Physics Letters **17**, 230, 1965.
[18] I. Giaever, Phys. Rev. Letters **15**, 825, 1965 and **16**, 460, 1966.
[19] J. van Suchtelen and A. P. Severijns, not published.
[20] See for example the books by London and Lynton quoted under [2].
[21] For a survey of this with bibliography see the article by J. Volger in Quantum Fluids, Proc. Sussex Univ. Symp. 1965, North Holland Publ. Co., Amsterdam 1966, pp. 128-135.

these eddy currents, which also pass through the normal core of every flux thread, which provide the frictional force and in fact lead to energy dissipation mentioned earlier. If a flux thread is held fixed (by whatever mechanism), then the current will pass completely outside its normal core but, if the flux thread is moving, the current "corrected" by the eddy currents will pass through the vortex core. This picture thus immediately makes it clear that the normal or residual resistance of the material determines the resistance behaviour in the mixed state. These ideas have been examined theoretically by various authors, who have successfully demonstrated a connection with Kim's empirical relation:

$$\varrho_m = \varrho B/H_{c2}. \quad \ldots\ldots \quad (27)$$

Here ϱ_m is the resistivity in the mixed state and ϱ the resistivity in the normal state. The flux density B is the average internal magnetic field intensity in the specimen; B/H_{c2} is nothing more than the fraction of the volume taken up by the normal cores of the vortices.

Yet another interesting consequence of this theory is the existence of a Hall effect: the magnetic field in the vortex core initiates a Hall effect there. Because of this, the local current loops mentioned, which cause the frictional force, are rotated slightly with respect to their electric field through the influence of the magnetic field. This rotation is in fact through an angle which must be of the order of magnitude of the Hall angle that must apply in the core region. We do not propose here to discuss the difficulties which have to be tackled in an adequate treatment of this problem. The idea of a rotation in the frictional force F_f with respect to $-u$ (cf. 25) through such an angle is however undoubtedly correct. It leads to a deviation in the flux drift with respect to the ideal lateral movement and therefore to a transverse voltage with all the symmetry properties of the Hall voltage, which can simply be referred to as the Hall voltage of the superconductor in the mixed state [21]. This Hall voltage has been observed in these laboratories in samples of NbTa etc., by Niessen and Staas [22]. Their observations, however, and those published later by other authors, show that a good quantitative agreement between theory and experiment has by no means yet been achieved.

Since a normal flux vortex core represents a certain quantity of entropy, it is to be expected that thermal effects will also be connected with the flux movement. These have indeed been found and superconductors of the second kind, in the mixed state, also exhibit the whole family of thermo-galvanomagnetic effects of the second order, i.e., the Nernst, Ettingshausen and Righi-le Duc effects, besides the Hall effect.

[22] A. K. Niessen and F. A. Staas, Physics Letters **15**, 26, 1965.
[23] See for example: G. J. van Gurp and D. J. van Ooijen, J. Physique **27**, C3-51, 1966.

"Hard" superconductors

The superconducting materials which for a few years now have been attracting a great deal of attention because of possible applications belong basically to the group of superconductors of the second kind. They differ however from the materials discussed above in that the resistance remains zero right up to very high values of the external magnetic field strength or up to very high currents. This extremely favourable property is a result of the fact that, in such materials, the flux threads are held tightly in place and cannot, therefore, move at all in the first instance. They can be held fast by all kinds of structural faults, like dislocations, crystal boundaries, precipitations of a second phase, etc. [23]. As yet there is no proper quantitative understanding of the mechanism by which the flux threads are held in place.

Superconductors of the second kind belonging to this group are aptly enough known as "hard" superconductors — most of them are indeed mechanically hard due to the presence of many crystal imperfections — or as superconductors of the third kind. Examples of these are NbZr and other niobium alloys and the intermetallic compound Nb_3Sn, which is used for making cables for superconducting magnet coils. The properties of these materials will be discussed in the article by A. L. Luiten on superconducting magnets mentioned under [4], which will appear shortly in this journal.

Summary. Superconductivity, discovered in 1911 by Kamerlingh Onnes and Holst, remained for half a century a phenomenon which was not satisfactorily explained and which could not be used in any technical application. In the last ten years the situation has changed drastically in both respects. It is now possible to construct superconducting magnet coils for fields of 100 kilogauss or more; other technical applications are under consideration. A theoretical understanding of superconductivity has been obtained by treating it as a "condensation" of conduction electrons into pairs (Bardeen, Cooper and Schrieffer). The condensate assumes a macroscopic quantum state which cannot alter to an energetically lower state. There is therefore no development of heat and no resistance. If the temperature or an external magnetic field is increased, then there is an instant at which the superconducting state disappears. The theory of Landau and Ginzburg has made clear why superconductors have to be divided into two groups with regard to their behaviour in an external magnetic field: one group in which the interior of the superconductor remains field-free and another in which the field can penetrate the superconductor in the form of billions of flux threads if the field strength exceeds a certain value. Most members of this group exhibit a certain resistance to the passage of a current (superconductors of the second kind) while others do not do so and are capable of carrying very heavy currents (third kind). The superconducting materials of the third kind are used in the construction of superconducting magnet coils.

Low-pressure sodium lamps with indium oxide filter

The further development of sodium lamps in recent years has led to a considerable increase in luminous efficiency.

The higher efficiency has mainly been attained by improving the heat insulation of the lamps by means of heat reflecting filters. Most of the electrical energy which is supplied to a sodium lamp is wasted as undesired heat radiation from the discharge tube. These energy losses can be reduced considerably by means of a filter which has a high transmission for sodium light and a high reflectivity for the heat radiation from the discharge tube. The filter consists of a coating on the inside of the glass envelope which encloses the discharge tube.

Suitable filters have been developed in which heavily-doped semiconducting tin oxide is used. These films have been applied in commercial sodium lamps for some years, giving luminous efficiencies up to 150 lm/W [1][2][3]. The transmission of glass coated with such a tin oxide film is 87-89% for sodium light ($\lambda = 0.59$ μm) and this figure is only slightly less than that for uncoated glass (92%). As the heat radiation of the sodium lamp is emitted almost completely in the infra-red region at wavelengths > 3 μm the filter should also reflect strongly at $\lambda > 3$ μm. This condition is in fact fulfilled by the tin oxide films which attain a saturation reflection of 80% at $\lambda \approx 8$ μm.

The high infra-red reflection of the tin oxide films is related to their electrical properties. As can be shown from the dispersion theory of free charge carriers the concentration of free carriers must be higher than 3×10^{20}/cm³ to obtain high infra-red reflection for wavelengths > 3 μm. With the heavily-doped tin oxide films carrier concentrations of 6×10^{20}/cm³ can be achieved. An additional requirement for high infra-red reflection, which is also met by the tin oxide film, is a large value for the product of the effective electron mass and the mobility of the free carriers. The reflection increases as this product attains higher values (see reference [2], page 108).

Semi-conducting indium oxide films have recently been investigated [4] and these showed promise of even better results for sodium lamps than those obtained with tin oxide.

Indium oxide films can be prepared in a similar way to tin oxide films. A mixture of $InCl_3$ with an organic solvent such as butyl acetate is sprayed from an atomizer on to the hot glass. If the temperature is high enough, $InCl_3$ is converted to In_2O_3 which forms a thin layer on the glass. Indium oxide films prepared by the spray technique with no additional doping agent have carrier concentrations of only about 10^{19}/cm³ which are of course not sufficient to provide the desired high infra-red reflection. The carrier concentration can be increased considerably, however, by doping with tin, and for this purpose $SnCl_4$ is added to the spray mixture. *Fig. 1* shows the effect of different tin doping on the resistance per square of the indium oxide films. These films were prepared on polished borosilicate glass plates. The resistance per square $R_\square$ is reduced from 170 Ω to very low values of 7 to 8 Ω if the spray mixture is doped with 2-3 atomic percent of tin. The flat minimum in the resistance curve for tin doping of about 2.3 atomic percent corresponds to a concentration of free charge carriers $N = 5\times10^{20}$/cm³

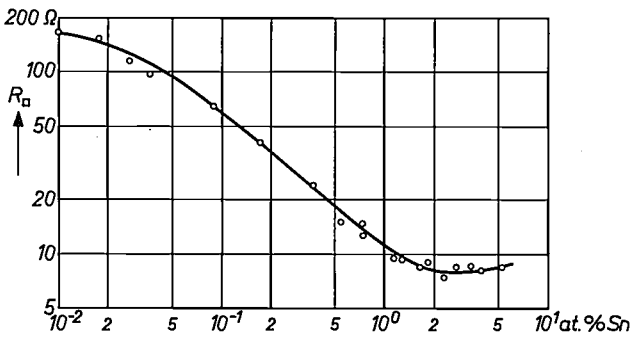


Fig. 1. Resistance per square of indium oxide films as a function of tin doping at room temperature. Spraying temperature 500 °C; film thickness 0.32 μm.

and a mobility for the free carriers $\mu = 50$ cm²/Vs. The concentration achieved in indium oxide can be seen to be nearly the same as for tin oxide ($N = 6\times10^{20}$/cm³) but the mobility is much higher than for tin oxide ($\mu = 20$ cm²/Vs).

These results led us to expect that the infra-red reflection of these indium oxide films would be higher than that of tin oxide provided that the effective mass of the free electrons, which has not yet been determined, was not much lower than it is in tin oxide.

This expectation was confirmed by optical measurements. *Fig. 2* gives a graph of the spectral transmission and reflection of such an indium oxide film. (The corre-

[1] M. H. A. van de Weijer, Recent improvements in sodium lamps, Philips tech. Rev. **23**, 246-257, 1961/62.
[2] R. Groth and E. Kauer, Thermal insulation of sodium lamps, Philips tech. Rev. **26**, 105-111, 1965.
[3] H. J. J. van Boort, Electrotechniek **43**, 509, 1965.
[4] R. Groth, Phys. Stat. sol. **14**, 69, 1966.

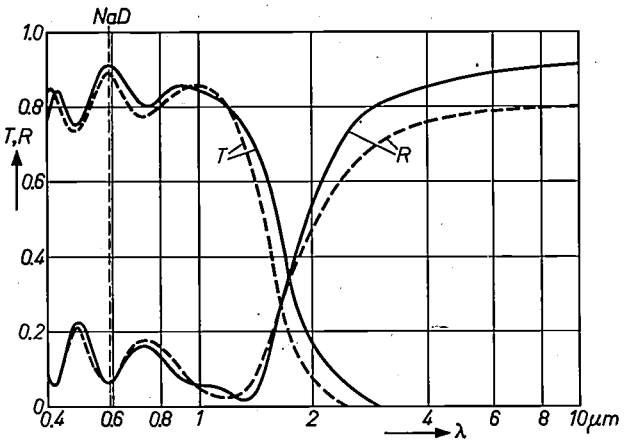


Fig. 2. Solid curves: spectral transmission T and reflection R of an indium oxide film. Spraying temperature 500 °C; film thickness 0.31 μm; concentration of free charge carriers 5×10^{20} cm^{-3}; mobility of the free carriers 50 cm^2/Vs.
Dashed curves: the same quantities for a tin oxide film. Spraying temperature 460 °C; film thickness 0.32 μm; concentration of free charge carriers 6×10^{20} cm^{-3}; mobility of the free carriers 20 cm^2/Vs.

sponding curves for the tin oxide film mentioned previously are given for comparison.) As the filters hardly absorb at all in the visible region, the transmission in this region is only modified by the interference. The film thickness of 0.31 μm for the indium oxide film has been chosen to obtain a maximum of transmission for the wavelength of the sodium D lines. For this wavelength the transmission of the glass coated with the film is 91%, which very nearly corresponds to the transmission of uncoated glass (92%). The infra-red reflection of the filter is higher than for tin oxide and reaches 90%. Films prepared on the inside of soft glass tubes showed the same filter characteristics as films prepared on polished borosilicate glass plates.

A consequence of the spraying technique used is that only 10-15% of the indium metal in the spraying solution is actually used for the formation of the indium oxide film on the glass envelope. Since indium is more expensive than tin, it is necessary to reduce this loss of indium. It has been found that much of the vapour which contains the excess indium and leaves the spraying oven on the opposite side to the spraying gun can be dissolved in a spray deduster. After concentration of the solution in the spray deduster the indium metal can be recovered by normal chemical processes.

In commercial low-pressure sodium lamps of the SOX-type the sodium discharge is generated in a U-tube. This discharge tube is sealed into a cylindrical outer vacuum bulb. Coating this bulb with a tin oxide layer on the inner wall gave a luminous efficiency of 150 lm/W at an input power of 200 W. When an In_2O_3 coating of adequate thickness is applied to this type of lamp instead of a SnO_2 coating a luminous efficiency of 175 lm/W is obtained at an input power of 180 W. In *Table I* the efficiency of several SOX lamps which have In_2O_3 coatings is shown alongside results for comparable lamps with SnO_2 coatings. In general, the In_2O_3 coating gives an improvement in luminous efficiency of about 15% when compared with a SnO_2 coating, the dimensions, current and lumen output of the lamp remaining unchanged. This improvement considerably enhances the superiority in luminous efficiency of low-pressure sodium lamps over that of other gas discharge lamps.

Table I. Lamps from the SOX series.

SOX with SnO_2 coating		SOX with In_2O_3 coating	
100 W	125 lm/W	90 W	144 lm/W
150	137	135	157
200	150	180	175

It is a well-known fact that sodium lamp installations are usually installed because of the need to keep down the cost of electricity consumption. The further reduction of power consumption by 15% due to the improvement in the thermal insulation will therefore be an important factor in favour of the application of this light source.

H. J. J. van Boort
R. Groth

Drs. H. J. J. van Boort is with the Development Laboratory at Turnhout (Belgium) of the Philips Lighting Division; Dr. R. Groth is with the Aachen laboratory of Philips Zentrallaboratorium GmbH.

The use of digital circuit blocks in industrial equipment

C. Slofstra

Digital equipment contains large numbers of certain basic types of circuit, such as counters, decoders and shift registers. The article below describes how some of these widely used circuits can be made up from the digital circuit blocks described earlier in this journal. This requires a more detailed description of the actual basic circuits than the usual type of treatment, which presents these circuits as "black boxes" of specified function without telling the reader what is inside them. As an example of the application of digital circuit blocks in industrial equipment, the article describes an equipment which measures the length of sheets which are cut off in the manufacture of corrugated cardboard.

Digital circuit blocks — units containing complete basic digital circuits — can readily be combined with one another to build up complicated pieces of equipment. The circuit blocks are commercially available in series of units, which include bistable circuits and logic circuits. The individual units have the form of small blocks provided with connecting pins (*fig. 1*). The advantages of such a series were dealt with in a previous article [1], which described the various types of circuit blocks marketed by Philips. This article will henceforth be referred to here as 1. In the present article we shall consider their applications.

Circuit blocks have to meet different requirements in different fields of application. Circuit blocks for computers, for example, should have high speed, small volume and low dissipation. For applications in industrial equipment the chief requirement is one of insensitivity to electrical interference. The lower speed and higher dissipation which this entails do not give rise to difficulty in these applications.

In the present article we shall confine ourselves to the industrial applications of circuit blocks and we shall base our treatment on the range of blocks which Philips have designed specially for these applications and which is known as the 10-series. After a brief review of this series, we shall describe a number of widely used circuits, such as logic circuits, counters and shift registers, which can be made by combining various circuit blocks. We shall then conclude with a description of a complete piece of equipment containing such circuits.

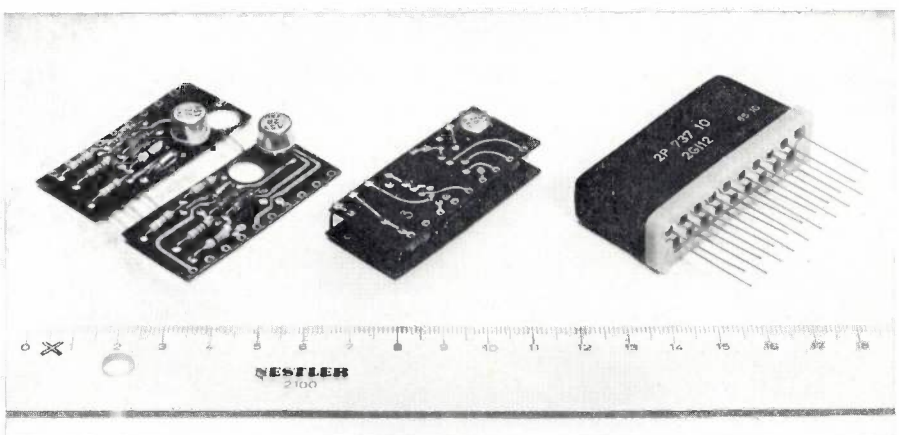


Fig. 1. One of the 10-series circuit blocks, containing two logic circuits. The circuits are mounted on printed wiring boards which fit into a metal case.

Brief review of the 10-series

1) Logic units. The basic feature of the logic circuits which can be built up from these units is a choice be-

Ir. C. Slofstra is with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven.

[1] E. J. van Barneveld, Digital circuit blocks, Philips tech. Rev. **28**, 44-56, 1967 (No. 2).

tween two alternatives, represented by low and high voltage levels at the output. This choice of the high or the low level is determined by the presence or absence of a particular combination of input voltages. The GI circuit blocks (type numbers 2.GI 10, 2.GI 11 and 2.GI 12) contain two logic circuits, each consisting of an AND circuit for positive logic (see below) followed by an inverting amplifier (GI stands for gate inverter). The numbers 10, 11 and 12 indicate different numbers of inputs. The GA circuit blocks (GA 10 and GA 11) contain an AND circuit for positive logic, followed by a non-inverting amplifier (GA for gate amplifier). This amplifier can take a far heavier load than the inverting amplifier in the GI blocks.

7) Indicator tube driver. This circuit block, the ID 10, contains, in addition to a decoder (from the binary 8-4-2-1 code to the decimal code) ten transistors for illuminating the correct digit of the indicator tube.

The series also includes various input and output circuits and a number of power supply units. To simplify the assembly of equipment, standardized assembly components are available, such as printed wiring boards, connectors and racks.

Any circuit block should be represented in a drawing by a rectangle containing the type number (see *fig. 2c*). The inputs and outputs are denoted by their letter code, together with numbers corresponding to the pins. These symbols are available in the form of stickers, so that

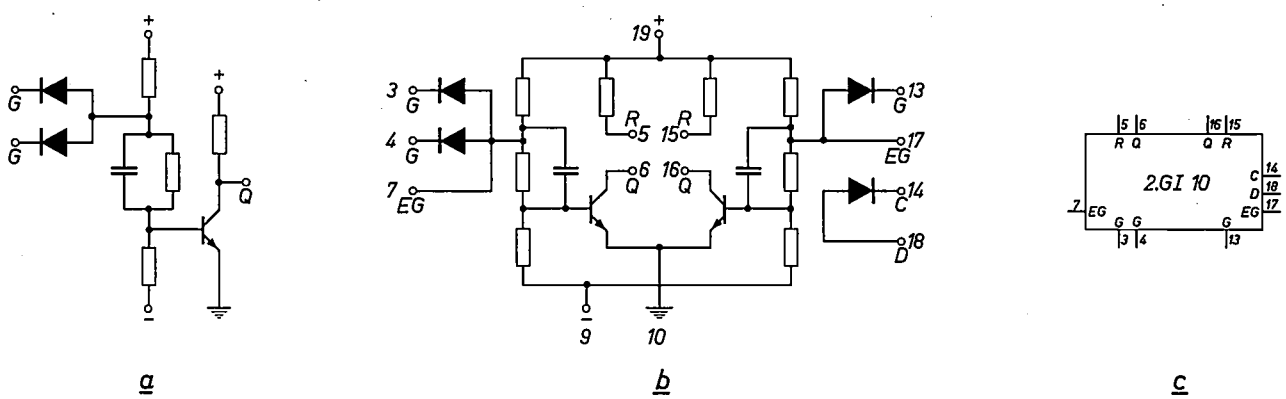


Fig. 2. *a*) The logic circuit of the GI blocks. *G* inputs. *Q* output.
b) The circuit of block 2.GI 10. *EG* points where the number of inputs can be expanded by the addition of diodes (e.g. the diode between *C* and *D*). For normal operation points *5* and *6* and points *15* and *16* are short-circuited externally.
c) Symbol used for these circuit blocks. The numbers of the outputs correspond to the pin numbers.

2) Bistable units. Circuit blocks FF 10, FF 11 and FF 12 (FF for flip-flop) each contain a bistable circuit; types FF 11 and FF 12 are also provided with trigger gates at the inputs of the bistable circuit.

3) Trigger gates. The three types in this series are 2.TG 13, 2.TG 14 and 4.TG 15. The first two each have two trigger gates and the last one four.

4) Delay devices. These include the monostable circuit (or one-shot multivibrator) OS 10, the timer unit TU 10, and the pulse driver PD 10. They differ in output power and in the time delay obtainable. The input of all these delay circuits is a trigger gate.

5) Pulse shapers. The PS 10 block was developed for restandardizing the edges of pulses in time and amplitude. This circuit block contains a Schmitt trigger followed by an inverting amplifier.

6) Power amplifiers. These include the RD 10 (relay driver) and the PA 10 (power amplifier). The first can operate small relays and pilot lamps, the second handles higher powers.

circuit diagrams can be formed quickly and simply which are clear both to the designer and the wireman.

Application of logic units

The logic circuit used in the GI blocks is shown in fig. 2*a*. When both inputs *G* are at the higher potential (level H) the output *Q* is at the low potential (level L). In the GI blocks two of these circuits are combined (fig. 2*b*), the left-hand one with two inputs and the right-hand circuit with one. One of the two circuits can be given an additional input by connecting a separate diode, included in the circuit block, to one of the points *EG*.

Table I is the truth table relating to the circuit of fig. 2*a*, and gives the output level for all possible combinations of input levels. In the mathematical treatment of logical operations with the aid of Boolean algebra, the two states of a system are denoted by "0" and "1". These states can be related in two ways to the voltage levels of the circuit. In what is termed *positive logic*,

Table I. Truth table for the circuit in fig. 2*a*.

General		Positive logic		Negative logic	
inputs	output	inputs	output	inputs	output
H H	L	"1" "1"	"0"	"0" "0"	"1"
H L	H	"1" "0"	"1"	"0" "1"	"0"
L H	H	"0" "1"	"1"	"1" "0"	"0"
L L	H	"0" "0"	"1"	"1" "1"	"0"

state "1" is taken as the high level H; reference to table I shows that the circuit in fig. 2*a* then represents what is called a NAND function. In *negative logic* state "1" is taken as the low level L, and the circuit represents a NOR function. These terms NAND and NOR are contractions of NOT and AND and of NOT and OR. Here we should just briefly mention the significance of these functions. The AND function $Z = A.B$ is "1" if *A and B* are "1" and "0" in all other cases; the NOT function $Z = \bar{A}$ gives an inversion (interchange of "0" and "1"), the NAND function $Z = \overline{A.B}$ is therefore "0" if *A and B* are "1", and "1" in the other cases. If we take Z as the state at the circuit output and A and B as the states at the inputs, then we see that this function is represented by the central part of table I. The OR function $Z = A + B$ is "1" if *A or B*, or both, are "1"; the NOR function in these cases is therefore "0", and is "1" if *A and B* are "0". This is to be seen in the right hand part of table I.

As a general rule, when changing from one kind of logic to the other an AND circuit will become an OR circuit, and vice-versa. In order therefore to be able to describe a circuit as an AND, OR, NOR or NAND circuit, it is necessary to decide beforehand on the logic to be applied. Following normal practice, we shall adopt positive logic in the 10-series; the GI circuit is then the realization of a NAND.

Fig. 3 shows how the three logic functions AND, OR and NOT can be realized with NAND circuits. The symbol used in this figure for the NAND circuit was previously introduced in article I; the small circle at the output indicates that the circuit is inverting. Fig. 3 is easily verified from the truth table for positive logic in table I, if we remember that an unconnected input behaves as if it were at the high potential (level "1"). In this way any logic function that can be defined by Boolean algebra can be produced with the GI circuit blocks.

Apart from being able to take a greater load, GA circuit blocks have the advantage that the permitted voltage limits for the low level at the input are much wider than in the GI blocks (0 to $1\frac{1}{2}$ V, as against 0 to 0.3 V in the GI blocks). For this reason the GA circuit blocks are particularly useful as the output stage of a complex logic circuit in which only the output is loaded (e.g. the one in *fig. 4*). The logic circuit is then built up with diodes and resistors and connected to the GA input. Due to the voltage drop across the diodes of this circuit, the input voltage at the low level can be

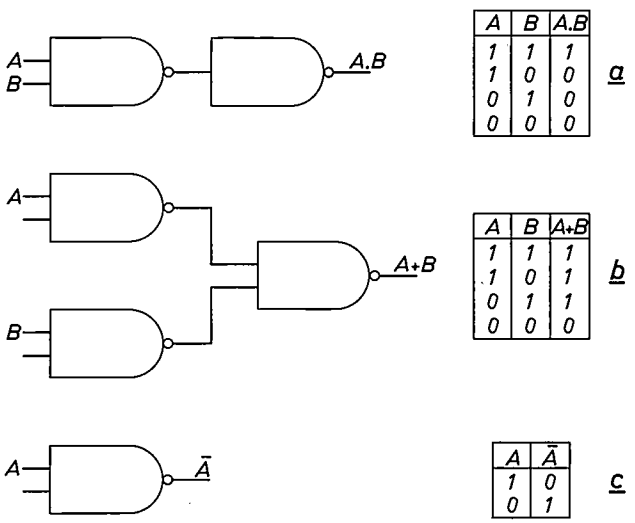


Fig. 3. The basic logic functions realized with NAND circuits. *a*) The AND function *A.B*. *b*) The OR function $A + B$. *c*) The NOT function $\bar{A}$. The truth table for each function is given.

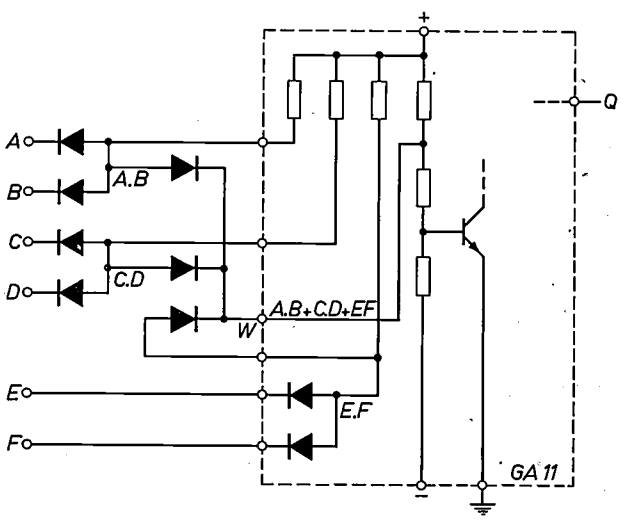


Fig. 4. Realization of the AND-OR function $A.B + C.D + E.F$ using diodes and a GA 11 block. The logic circuit is built up partly from external diodes and partly with diodes and resistors included in the circuit block. The amplifier of the GA 11 is used as output stage. The logic circuit is connected to the input *W* of the amplifier. A high load can be applied to the output *Q*.

higher than 0.3 V; as long as this voltage does not exceed 1.5 V, the GA block delivers the right output signal.

Application of the bistable circuit blocks

The bistable units of the 10-series were described in some detail in article I (p. 52). It will therefore be sufficient here to give the circuit diagram and a brief description of the most commonly used type, FF 12 (*fig. 5a*). The bistable circuit consists of two cross-coupled NAND circuits with transistors Tr_1 and Tr_2 and outputs Q_1 and Q_2. It can be brought into one of its two possible states by bringing one of the S inputs (S for set) to the low level. If, for example, the voltage at S_1 is made low, no current flows through Tr_1, Q_1 becomes high and therefore Q_2 low. The state is then of S, G and T inputs by means of external diodes. Circuit block FF 11 is like FF 12, but without the S inputs; FF 10 contains only the bistable circuit, without trigger gates.

Fig. 5*b* gives the symbol used for the FF 12. Here again all the inputs and outputs are indicated, and the pin numbers are shown. To avoid making the figures in this article too complicated, we shall henceforth indicate only the inputs used in the particular circuit under discussion. The two parts of a bistable circuit will then be distinguished by the subscripts 1 and 2.

Counting circuits

With circuit block FF 12 it is an easy matter to make a circuit that divides by a factor of two, called a scale-of-two counter. This is done by cross-connect-

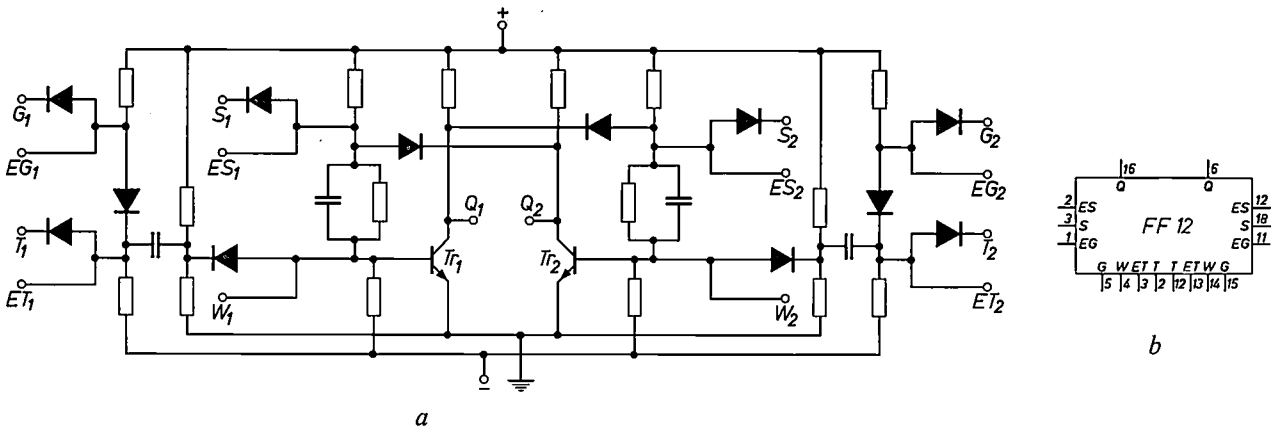


Fig. 5. *a*) The bistable circuit with trigger gates in the circuit block FF 12. *b*) Symbol for this circuit block.

Q_1 = "1". On each side of the circuit a trigger gate is connected to the base of the transistor (W_1 and W_2), and this trigger gate can be used to bring the bistable circuit into a particular state when a particular condition is satisfied. If, for example, the condition input G_1 is at the high level, an abrupt negative-going voltage step at the trigger input T_1 will stop the flow of current through Tr_1, resulting in the state Q_1 = "1". If G_1 is at the low level, a voltage step at T_1 has no effect and the state of the bistable circuit remains unchanged. Thus, a trigger gate is opened if G = "1" and closed if G = "0". The two trigger inputs of a bistable circuit with trigger gates may be connected, for instance, to a clock pulse generator; the voltages at the condition inputs, derived perhaps from logic circuits, will then at any time determine into which state the circuit is driven by the clock pulse.

The inputs *ES*, *EG* and *ET* of circuit block FF 12 were provided in order to be able to expand the number ing the outputs Q_1 and Q_2 with the condition inputs G_2 and G_1 and interconnecting the trigger inputs T_1 and T_2 (*fig. 6a*). If Q_1 is at the "0" level (and hence Q_2 = "1") the left-hand transistor will be conducting and the right-hand transistor will not. As a result of the connections

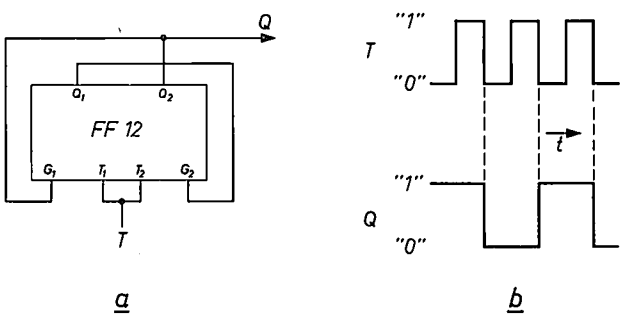


Fig. 6. *a*) The FF 12 used as a scale-of-two counter. When a series of pulses is applied to point T, pulses with a period twice as long appear at the output Q. *b*) Voltage waveform at points T and Q.

made, the condition input G_1 is now at the "1" level and G_2 at the "0" level, so that the left-hand trigger gate is open and the right-hand one closed. A voltage step from high to low at the common trigger input T will now have an effect on the left-hand trigger gate but not on the right-hand one. Consequently the current through the left-hand transistor will go to zero, Q_1 will change to the "1" level and Q_2 to the "0" level, which means that the bistable circuit has changed to the other stable state. The left-hand trigger gate is now closed and the right-hand one open, and therefore the arrival of the next trigger pulse [2] returns the bistable circuit to its original state. Two pulses at the input T thus result in one complete pulse at the output Q (fig. 6*b*), i.e. the circuit divides the number of incoming pulses by two.

The connection of n such units in cascade gives a circuit that divides by 2^n; an example is shown in *fig. 7* for $n = 4$. By making a few small modifications this circuit can be turned into a decade counter (*fig. 8*). A diode is introduced between the output Q_1 of D and the EG_2 input of B, and the T inputs of the last unit (D) are separately connected. If we consider the output voltages of the scale-of-two counters D, C, B and A, we see that the application of pulses to input I gives rise to the successive situations shown in *Table II* (see also fig. 8*b*; the reason for reversing the sequence will appear later). The process up to and including the eighth input pulse will be identical with that of a scale-of-16

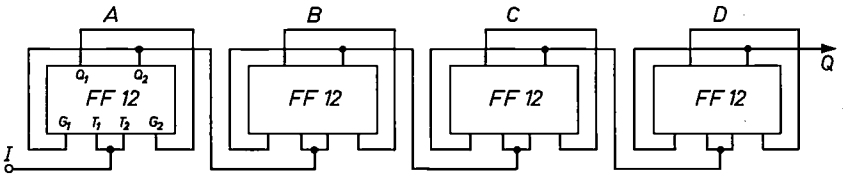


Fig. 7. Circuit of a scale-of-16 counter made up from four FF 12 circuit blocks.

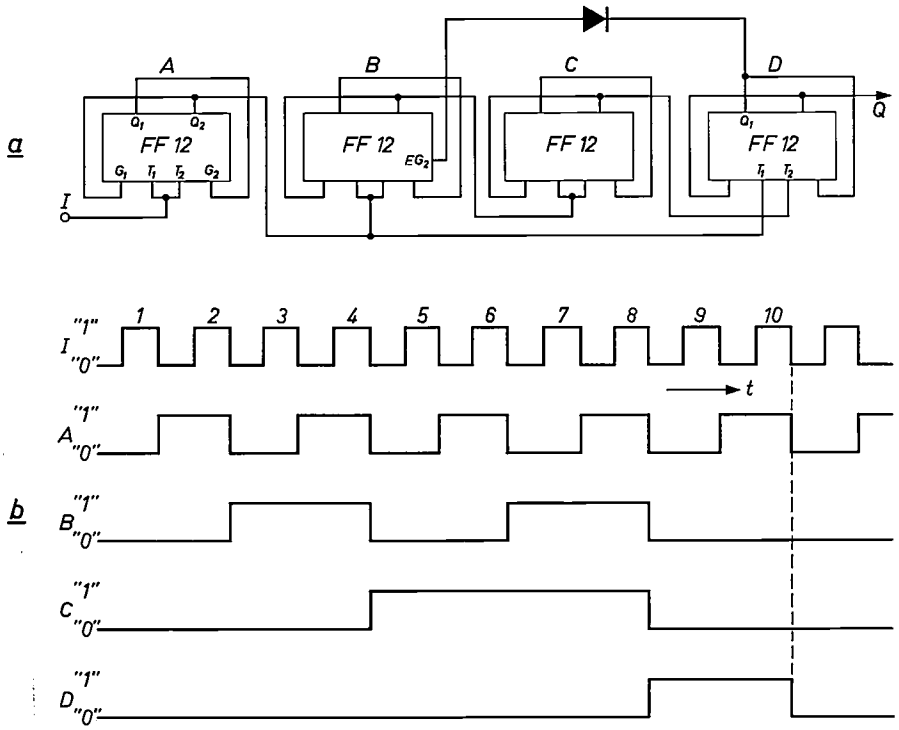


Fig. 8. *a*) A decade counter derived from the scale-of-16 counter of fig. 7 by the addition of a feedback diode. *b*) Voltage waveform at the circuit input I and at the outputs of the four stages A, B, C and D. For every ten input pulses a single pulse appears at the output of D, and this can be passed to the next counter unit which therefore indicates the tens.

[2] The bistable circuit responds only to an abrupt voltage drop from the high to the low level, i.e. to the trailing edge of a pulse. The leading edge has no effect. For simplicity, we shall not always mention this but simply state that a pulse causes the bistable circuit to change its state.

Table II. The states of the four scale-of-two counters D, C, B and A, of the decade counter in fig. 8 when pulses are applied to the input. Bold figures show the states of significance for decoding.

	D	C	B	A
initial state	**0**	**0**	**0**	**0**
after 1 pulse	**0**	**0**	**0**	**1**
after 2 pulses	0	**0**	**1**	**0**
„ 3 „	0	**0**	**1**	**1**
„ 4 „	0	**1**	**0**	**0**
„ 5 „	0	**1**	**0**	**1**
„ 6 „	0	**1**	**1**	**0**
„ 7 „	0	**1**	**1**	**1**
„ 8 „	**1**	0	0	**0**
„ 9 „	**1**	0	0	**1**

counter, because the output Q_1 of D is always "1" in this process, so that the right-hand trigger gate of B connected to it remains open in the normal way. The pulses from the output Q_2 of A will also have no effect on the left-hand transistor in D since this is not yet conducting (up to now Q_1 of D is in state "1"). After the eighth input pulse the situation changes. Circuit D switches over, Q_1 in this block changing to state "0" and Q_2 to state "1", thus opening the left-hand trigger gate. In addition the right-hand trigger gate of B is closed, since the EG_2 input is at the "0" level; a pulse at the T input, can therefore no longer change the state of this circuit. The ninth input pulse triggers the scale-of-two counter A in the normal way, causing its output Q_2 to switch from "0" to "1", but nothing else changes. At the tenth pulse, Q_2 of A again goes from "1" to "0". Because of the changed situation of B and D, circuit B now does not switch over, but remains in state "0", while D changes its state, becoming "0" as well. The result is that after the tenth pulse the initial situation is reached again. The circuit thus amounts to a decade counter. If we assign a value of 8 to the state D = "1", a value of 4 to C = "1" and, in the same way, values of 2 and 1 to B and A, this counter will operate in the 8-4-2-1 code; this means that the states of the units, in the sequence indicated in table II, indicate the number of counted pulses in the usual binary manner. The voltage drop appearing at the output Q_2 of D after the tenth input pulse can be used for controlling a second counter which therefore indicates the tens. Thus, by connecting a number of decade counters in cascade, we have a binary coded decimal (BCD) counter.

Decoding of counters

To give a visual indication of the contents (the reading) of a decade counter like the one shown in fig. 8, e.g. using indicator lamps or an indicator tube, a decoder consisting of ten AND circuits is generally used. Simple circuits of resistors and diodes (see fig. 7*a* of I) are quite adequate for this. When the counter reads 0 the outputs Q_2 of all the bistable circuits (fig. 8) are at the "0" level and the outputs Q_1 are thus at the "1" level. If we now connect the Q_1 outputs to an AND circuit, the output from this gate will therefore be "1" only when the counter reads 0. In order to decode reading 1 of the counter we connect to an AND circuit the output Q_2 of A and the outputs Q_1 of B, C and D, which are all in state "1" when the counter reads 1.

An AND circuit with four inputs is not needed for every reading of the counter. If we consider the states of the bistable circuits at the ten readings of the counter as given in Table II, we see that the output Q_2 of D is at the "1" level only at readings 8 and 9; at all other readings this output is at the "0" level. Readings 8 and 9 differ from each other only in the level of A, and therefore for decoding these readings only AND gates with two inputs are required. Readings 0 and 1 require four inputs, while three inputs are sufficient for the remaining readings. Altogether 30 inputs are needed. The states of the scale-of-two counters which have to be used for decoding are indicated in Table II by bold figures.

Fig. 9 shows one way in which a decoder of this kind could be made. The vertical wires are connected to the outputs of the scale-of-two counters; the horizontal wires can be connected via transistors to an indicator tube, which will then show the reading of the counter at any given moment. A more complicated circuit is used in block ID 10, which will not be dealt with here; this circuit block also contains ten driving transistors. This compact assembly makes it possible to accomodate two decade counters with indicator tube drivers

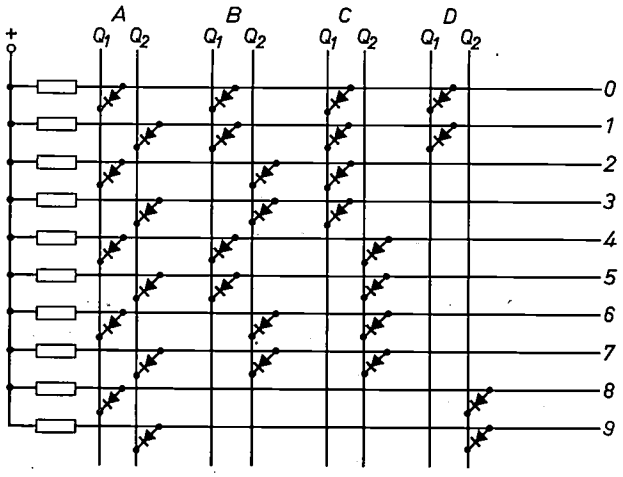


Fig. 9. Circuit for decoding the contents of a decade counter using four scale-of-two counters A, B, C and D. The outputs Q_1 and Q_2 of A, B, C and D are connected to the vertical wires; a horizontal wire goes to the "high" voltage when the counter reads the corresponding number.

on a single, standard printed wiring board, an example of which is to be seen in I, fig. 2.

Presetting of counters

In many cases a counter is required to deliver a signal when a preset reading is reached. This can be done with "thumbwheel" switches (*fig. 10*); the number is preset by turning a small wheel. *Fig. 11* shows the circuit diagram of a thumbwheel switch; the inputs of an AND circuit are connected through sliding contacts to the

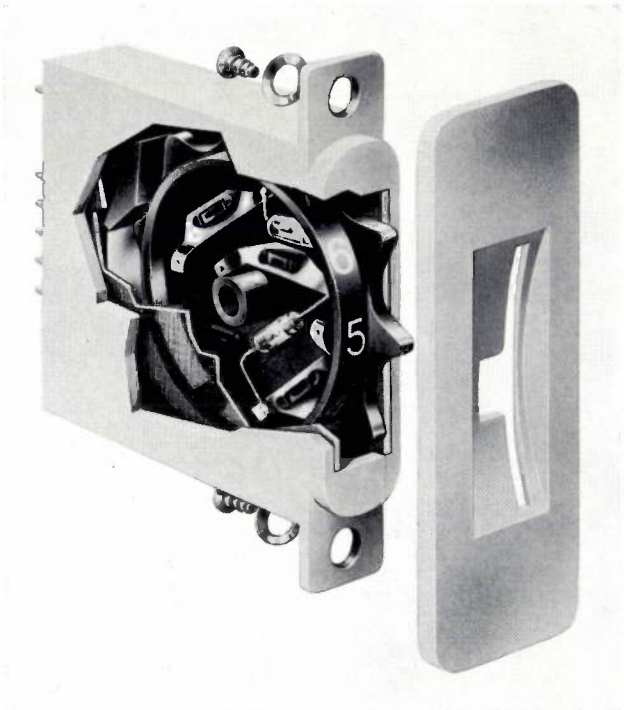


Fig. 10. Cut-away drawing of a thumbwheel switch. When connected to a decade counter, this switch gives a signal when the counter reaches the value to which the wheel has been preset. The diodes of the AND circuit can be seen inside the wheel and also the sliding contacts which make the connections with the counter through code discs on the inside of the case (dimensions 12.7 × 65 × 57 mm).

four bistable circuits of the counter, in such a way that at any given position of the wheel (i.e. of the sliding contacts) the four input levels are "1" when the counter reads the number corresponding to the position of the wheel. The output Q is then at the "1" level. The connections to A, B, C and D naturally follow the same pattern as in fig. 9. Thus, a decimal counter with a capacity of 10^n, i.e. consisting of n decade counters, has n thumbwheel switches. The AND circuits of these switches can then be combined to form a single AND circuit with $4n$ inputs, which responds only when all decade counters have reached the preset reading.

In some cases, particularly in numerical control, the

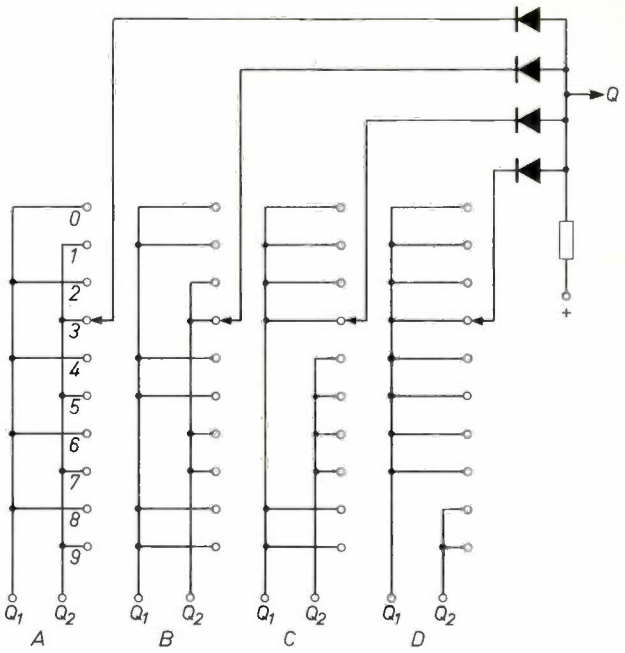


Fig. 11. Basic circuit of a thumbwheel switch. The four inputs of an AND circuit are connected by four contacts to the outputs Q_1 and Q_2 of the bistable circuits A, B, C and D of a counter. The output Q goes to the "1" level when the counter reaches the preset reading.

preset numbers have to be changed frequently and quickly, which rules out the use of thumbwheel switches. The required number is then fed into a register store consisting of the same number of bistable circuits as the counter. A logic circuit is then used to determine whether the counter has reached the preset reading.

Shift registers

A circuit which is widely used is the shift register. It consists of a series of bistable circuits connected up in such a way that if a pulse appears at a particular input, the state of each bistable circuit is passed on to the next one in the series, so that the information stored in the register moves up one place. *Fig. 12* shows a shift register made up from FF 12 circuit blocks. In this device the T inputs of all the bistable circuits are connected in parallel and the outputs of each bistable cir-

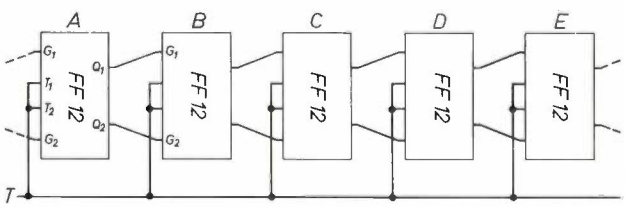


Fig. 12. Shift register built up from FF 12 circuit blocks. The outputs Q_1 and Q_2 of each block are connected to the condition inputs (G_1 and G_2) of the next bistable circuit. When a pulse appears on the common trigger input T, bistable circuit B will take over the state of A, C the state of B, and so on, thus shifting the contents of the register one place to the right.

cuit are connected to the G inputs of the next block in the series. The operation may be simply explained as follows.

Suppose that bistable circuit A is in the state where Q_1 is at the "1" level and Q_2 is therefore at the "0" level. The input G_1 of B is now at the "1" level as well, and G_2 at the "0" level, so that the upper trigger gate of B is open and the lower one closed. If a pulse now appears at the trigger input, the upper transistor of B will be cut off, irrespective of the previous state of B, so that Q_1 goes to the "1" level and hence Q_2 goes to the "0" level. This means that B has taken over the state of A, and in this way the state of each bistable circuit is passed on to its right-hand neighbour.

As soon as a bistable circuit switches over, the voltages across the condition inputs of the next bistable circuit also change. This circuit, however, must not yet react to these new values but has to take over the old state of the first bistable circuit. In each trigger gate a storage action is therefore required in order to remember the old voltages until the next bistable circuit has switched over. This is the purpose of the capacitor in the trigger gate (to be seen in fig. 5), which gives the trigger gate a specific response time (see also I, p. 53).

A shift register can be modified by connecting the outputs of the last bistable circuit to the condition inputs of the first one. This gives a register in which the information circulates. It is also possible, of course, to connect a number of shift registers in parallel and drive them with the same trigger pulses; with four rows of bistable circuits a shift register can be made for a decimal figure in the 8-4-2-1 code.

Application of circuit block PS 10

The PS 10 circuit block contains a Schmitt trigger and an inverting amplifier (*fig. 13a*). The circuit is designed in such a way that the output voltage V_Q is always at the "0" or at the "1" level within the tolerances required for the 10-series. The transition from "1" to "0" is fast enough to be able to drive a bistable circuit with this signal. The circuit block was designed primarily for shaping a pulse to the right height and edge steepness for driving the other circuit blocks. Fig. 13*b* shows the relation between the input voltage V_B and output voltage V_Q. It can be seen that the transition from "1" to "0" takes place at an input voltage different from that for the transition from "0" to "1". This hysteresis effect can be put to good use in several applications; we shall consider here the generation of a square-wave signal.

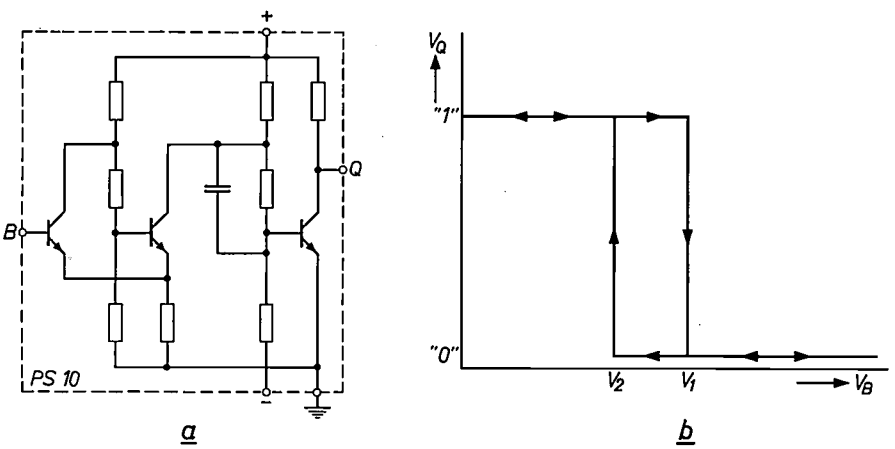


Fig. 13. *a*) Basic circuit of the PS 10 circuit block, consisting of a Schmitt trigger and an inverting amplifier. *b*) The output voltage V_Q of the PS 10 as a function of the input voltage V_B. If V_B is low, V_Q is at the "1" level. As V_B increases, state "1" suddenly changes to state "0" at a voltage V_1 which is higher than the voltage V_2 at which "0" returns to "1" when V_B decreases. This hysteresis region lies between the values $V_2 = 0.13\ V_p$ and $V_1 = 0.36\ V_p$, where V_p is the applied positive voltage corresponding to state "1".

Fig. 14 shows a square-wave oscillator whose frequency can be preset within a wide range. We start from the situation in which C is not charged, so that V_B is low and V_Q is at the "1" level. Capacitor C now starts to charge up through R_1, R_2 and R_3. As soon as this causes V_B to exceed the value of V_1 in fig. 14, V_Q goes to the "0" level. Diode D now conducts and C starts to discharge through R_3, R_2 and D. This continues until V_B is lower than the other change-over point of the PS 10, whereupon V_Q returns to the "1" level and the process is repeated. The oscillation can be stopped by putting one or both inputs G_1 and G_2 at the "0" level, thus holding V_B at a low voltage. The frequency can be set anywhere in a range from below 1 Hz to several tens of kHz by choosing the value of C.

A square-wave voltage with a frequency of 50 or 100 Hz can be obtained from the PS 10 circuit block by feeding a half-wave or full-wave rectified mains signal into the input. This arrangement is frequently used for time measurements in industrial equipment; the mains frequency is usually sufficiently accurate for this purpose.

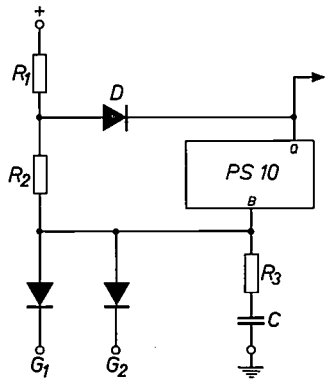


Fig. 14. Square-wave oscillator. As long as V_B is low and V_Q is at the "1" level, charge flows into C through R_1, R_2 and R_3. When V_B has risen to a value at which V_Q goes to the "0" level, C discharges through R_3, R_2 and D until V_B has fallen far enough for V_Q to return to the "0" level. The oscillation can be stopped by bringing G_1 or G_2 to the "0" level.

the need for very fast or non-wearing input elements, various electronic devices have been designed, some of which will now be discussed.

The first to be mentioned is the Vane Switched Oscillator (VSO). This is an oscillator circuit containing two coils, one in the base circuit of a transistor and the other in its collector circuit; the oscillator output is taken to a rectifier via a coupling winding in one of the coils (*fig. 15*). The oscillator starts when the supply voltage V_p is applied; the output voltage V_Q from the rectifier is then positive. If a piece of metal is now inserted between the two coils of the oscillator, the coupling between the coils is reduced and the oscillator stops. The output signal then drops to the "0" level.

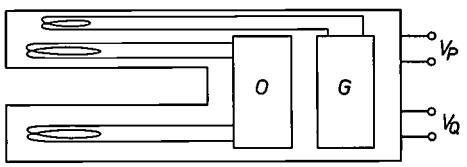


Fig. 15. Diagram of a Vane Switched Oscillator (VSO). The oscillator O, which operates when a supply voltage V_p is applied, stops when a metal vane enters the gap between the two coils. The output voltage V_Q of the rectifier G then drops from a positive value to zero.

The other circuit blocks

The 10-series also contains blocks with delay circuits and amplifiers. We shall not mention any special applications of these types, but we shall consider type PD 10 in somewhat more detail because this circuit block is used in the device described in the last part of this article.

Circuit block PD 10 contains a trigger gate followed by a monostable circuit. In the steady state the output voltage of PD 10 is at the "1" level. If the voltage across the trigger input drops from the "1" to the "0" level, the output voltage goes to the "0" level, and returns after a certain time interval to the "1" level. This output pulse can be used for driving many inputs of other circuit blocks. Apart from its pulse driver function, the PD 10 can be used as a delay circuit; the time interval during which the output voltage is at the "0" level can be adjusted between four microseconds and several tens of milliseconds.

Input and output devices

Since the only language the circuit blocks understand is that of the "0" and "1" levels, whereas the variables of the processes being controlled are velocities, angular displacements, temperatures, times, etc., these variables have to be "translated" by transducers acting as input and output devices for the electronic system. In view of

The complete circuit is enclosed in a case of the form shown in fig. 15. One application is to mount the VSO near a rotating shaft in such a way that a small metal vane fixed to the shaft traverses the gap of the VSO once per revolution, giving a voltage pulse at the output of the VSO. In practice a PS 10 is usually connected in series with the VSO to ensure that the pulses have the right height and edge steepness. An input circuit operating on the same principle is the Electronic Proximity Detector (EPD); in this device the oscillation is stopped when a piece of metal comes close to the coil.

Photoelectric cells are widely used as input devices; they are commercially available in various types. The simplest type is based upon the interruption of a beam of light incident on a photo cell. A more complicated photoelectric device is the rotation transducer, which delivers several hundred pulses for every revolution of a shaft, thus allowing the position or speed of rotation of the shaft to be determined.

The last input device we shall mention is the ultrasonic detector. When this device is used together with flexible tubes as acoustic waveguides, signals can be detected at places otherwise very difficult to reach.

The output devices include, apart from the indicator tubes already discussed, two types of power amplifier, the RD 10 which can drive small relays and indicator

lamps, and the PA 10 for higher powers. For heavy-duty power handling one can use the thyristor, which also permits continuous control of the power output.

Example: Design of an equipment for measuring the length of cardboard sheets

As an example of the industrial application of the 10-series circuit blocks, we shall now describe an equipment which measures the length of sheets of cardboard and gives an alarm if there is too large a deviation from the required length. The equipment was designed to be fitted to one of the production machines at the Philips Corrugated-cardboard Works in Eindhoven. The manufacture of corrugated cardboard has previously been discussed in this journal [3]. We shall confine ourselves here to a brief description of the part of the machine where the cardboard is cut to length (*fig. 16*).

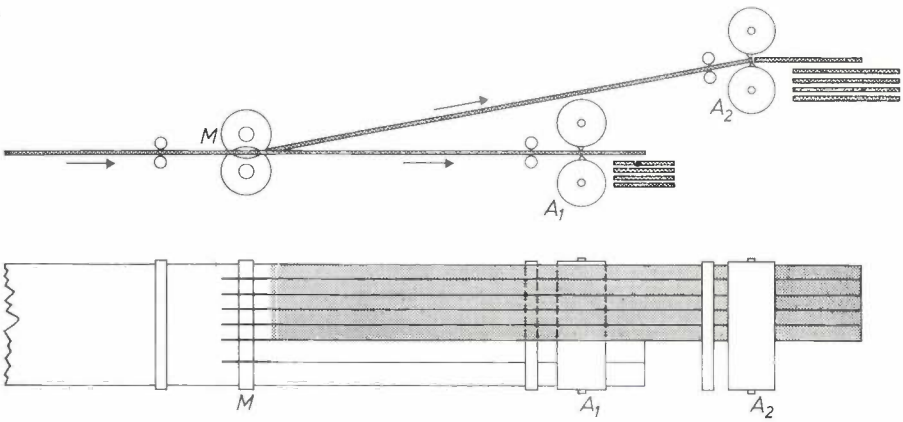


Fig. 16. Diagram of part of a machine for the production of corrugated cardboard. This is the part in which the cardboard is cut to size. The cardboard strip coming from the left is cut lengthwise by a number of cutting wheels M whose spacing can be adjusted. The resultant strips are divided between two pairs of cutting blades A_1 and A_2, which cut the strips to the required lengths. These blades are attached to rotating cylinders that are actuated by an eccentric mechanism which ensures that they travel along with the cardboard strip as it is being cut, so that this strip is not bruised or torn.

The continuous cardboard strip (width 2170 mm), travelling at a speed of about one metre per second, is cut into rectangular sheets by two kinds of cutting mechanism. The strip is first cut lengthwise into narrower strips by rotating cutting wheels M. These strips are then cut transversely to the required lengths by two pairs of cutting blades A_1 and A_2 placed at different heights. These blades are mounted on two rotating cylinders and at each revolution a cut is made. To obtain sheets of the required lengths the speed of rotation can be separately adjusted for each set of blades. However, the blades have to travel along with the moving strip of cardboard as they cut; this is achieved by driving the cylinders via an eccentric mechanism. The cardboard strips cut lengthwise can be divided as required between the two pairs of cutting blades; the arrangement thus offers considerable flexibility.

The length of the sheets of cardboard cut off is determined by the speed of the moving cardboard strip and by the speed at which the cylinders of the cutting blades rotate. If the cardboard travels at v metres per second and the cylinders complete one revolution in T seconds, then vT metres of cardboard are cut off. The speed of rotation can be adjusted by hand to ensure that the pieces are cut to the right length. As a result of slight variations, however, mainly in the speed of the cardboard strip, deviations from the proper length do arise after some time. The equipment described here gives a direct visual indication of these deviations, enabling the speed of rotation to be adjusted accordingly.

Fig. 17 gives a schematic diagram of the various parts of the equipment. A wheel W is driven by the moving strip of cardboard B. Attached to the rim of the wheel are a number of small magnets which generate pulses in a magnetic detecting head PU situated beside the wheel. The wheel and detector are so arranged that a pulse is delivered for every millimetre displacement of the cardboard strip. The time of revolution of the cylinders with the cutting blades is measured by a vane switched oscillator (*VSO* in the figure). Attached to one of the cylinders is a metal vane which passes through the gap in the VSO once every revolution, so that the VSO delivers a pulse after every T seconds. The output

[3] H. W. van den Meerendonk and J. H. Schouten, Trim-losses in the manufacture of corrugated cardboard, Philips tech. Rev. **24**, 121-129, 1962/63.

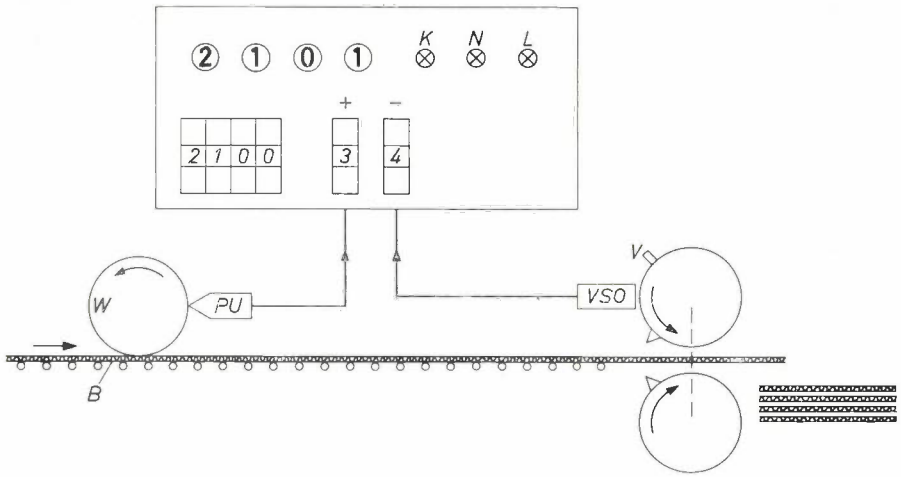


Fig. 17. Diagram of the equipment for measuring the length of the sheets of cardboard cut off. The wheel *W* is driven by the corrugated-cardboard strip. The magnetic detecting head *PU*, which is energized by small magnets fixed to the rim of the wheel, gives a pulse (a wheel pulse) for every millimetre displacement of the strip. A VSO, whose gap is traversed by a metal vane *V* mounted on one of the cylinders of the cutting blades, delivers a pulse (a blade pulse) for every revolution of the cylinder. Both signals are fed to a cabinet containing the digital circuit. The front panel of this cabinet is arranged as follows: at the bottom there are four thumbwheel switches for presetting the required length and two thumbwheel switches for positive and negative tolerance limits; at the top of the panel there are four indicator tubes which read out the measured length and, on the right, three lamps *K*, *N* and *L* for indicating the situations "too short", "normal", "too long". In the situation illustrated here, the length is within the tolerance limits and the lamp *N* is therefore switched on.

signals from the magnetic detecting head (the wheel pulses) and of the VSO (the blade pulses) are fed into a digital circuit, which is also supplied with information about the required length (in mm) and the positive and negative tolerance limits (in mm) by means of thumbwheel switches.

With the aid of the signals from the magnetic detector and from the VSO the equipment determines the length cut off and compares this with the required length. If the tolerance limits have been exceeded, this is indicated by lamps. Indicator tubes are used for reading out the length of

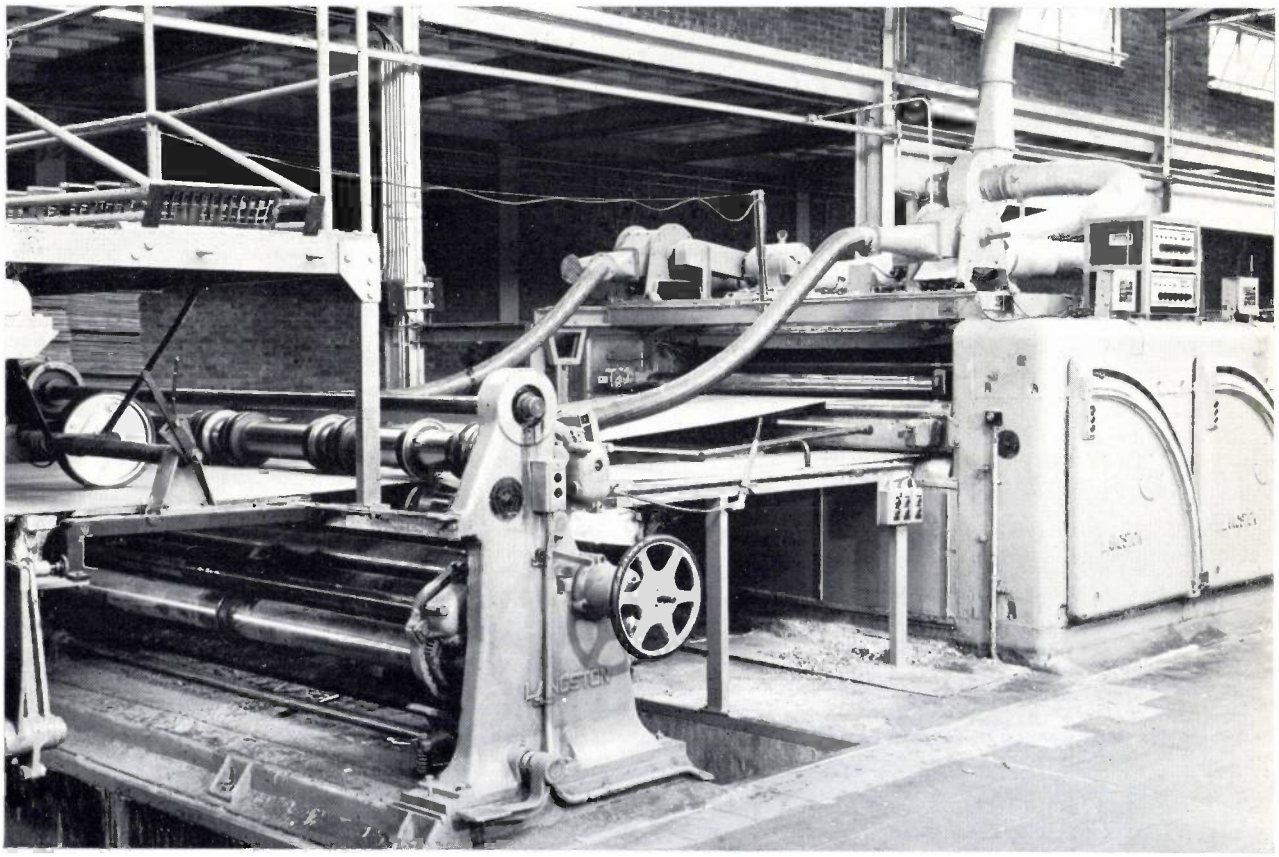


Fig. 18. The part of the production machine where the corrugated-cardboard strip is cut (cf. fig. 16). On the left can be seen the wheel which measures the speed at which the strip is travelling, and on the right of it the cutting wheels which cut the strip lengthwise. The two cabinets containing the digital circuits are on the right, one for each pair of cutting blades. Below them can be seen the mechanism for adjusting the speed of rotation of the cylinders.

every second sheet cut off. *Fig. 18* shows the part of the corrugated-cardboard machine to which the equipment is fitted. On the left can be seen the wheel *W* and on the right the two cabinets, one for each pair of cutting blades. Located under these cabinets is the mechanism for adjusting the speed of rotation of the cylinders.

adjusting the required length. If this length is reached before the next blade pulse appears, the remaining wheel pulses are also fed to the "auxiliary counter", which therefore counts up the excess length of the sheet cut off. The auxiliary counter, which consists of a single decade circuit, is linked with a thumbwheel switch for

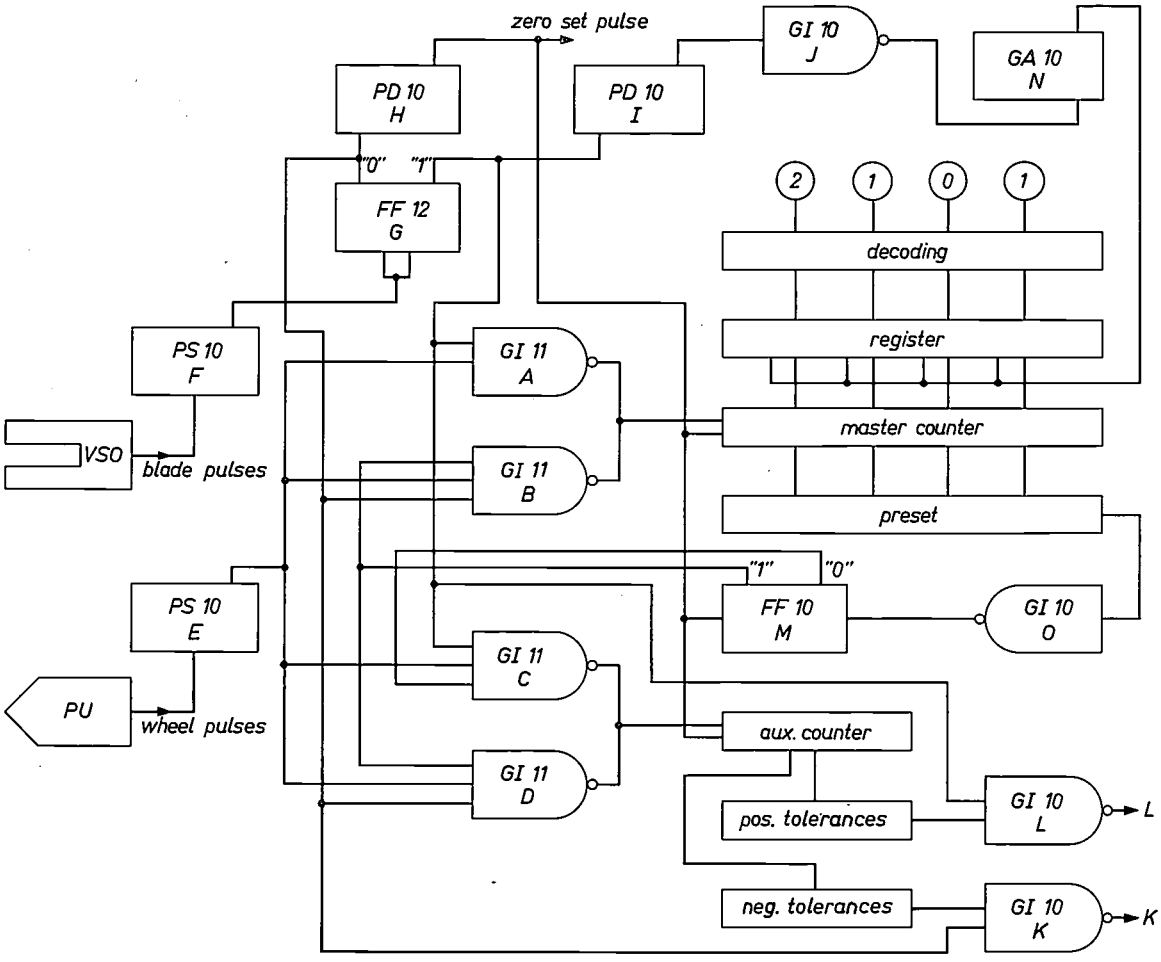


Fig. 19. The digital circuit. The wheel pulses from the magnetic detector *PU*, which appear between two blade pulses from the VSO, are counted in the master counter and their number compared with the preset value. The number of wheel pulses representing the difference between the measured and preset value is counted in the auxiliary counter. The NAND circuits *A*, *B*, *C* and *D* form trigger gates, controlled by the blade pulses and by the output signal from the thumbwheel switches used for presetting. The wheel pulses go from the trigger gates to the counters. After each measurement the final reading of the master counter is shown on indicator tubes, with the aid of a register and a decoder. The tolerances are preset by thumbwheel switches connected to the master counter; if a tolerance limit is exceeded, a pulse appears at the output *L* or *K*.

The digital circuit

The heart of the digital circuit (*fig. 19*) is the "master counter", a decimal counter with a capacity of four decades. This counts the number of wheel pulses delivered by the detector *PU* between two blade pulses (this number being the length cut off in mm). The master counter is connected to four thumbwheel switches for

adjusting the positive tolerance limit. If the limit is exceeded, this switch gives a signal which lights up an indicator lamp. Another thumbwheel switch is connected for adjusting the negative tolerance limit. When the sheet cut off is too short, i.e. when the blade pulse appears before the master counter has reached the preset value, wheel pulses are also supplied to the

auxiliary counter and these pulses indicate by how much the sheet is too short.

The wheel pulses are passed to the master counter and the auxiliary counter through gate circuits consisting of GI 11 circuit blocks *A*, *B*, *C* and *D* (fig. 19). The outputs of these NAND circuits, are interconnected *A* to *B* and *C* to *D*. A combination of this kind represents a NOT-AND-OR function (see I, page 51), which means that the common output goes to the "0" level if all inputs of one or both circuit blocks are at the "1" level. The output is thus at the "1" level if one or more inputs of both blocks are at the "0" level. We shall now consider in detail how these circuits are used to pass on the train of pulses.

As the starting situation we take the moment at which the VSO has just delivered a pulse; in this situation the outputs of circuit blocks *G* and *M* are at the levels indicated. The pulses from the VSO and from *PU* are first shaped to the appropriate height and edge steepness by the PS 10 blocks *F* and *E* (we recall here that only the transition from "1" to "0" is of importance). The wheel pulses are now fed to the circuit blocks *A*, *B*, *C* and *D*. At this moment the upper input of *A* is at the "1" level. If now *E* gives the "1" level as well, the common output of *A* and *B* will be at the "0" level. If *E* gives the "0" level, the output will be at the "1" level. This means that for every wheel pulse ("1" to "0" and back) the voltage at the input of the master counter goes from "0" to "1" and back, thereby activating the master counter. Block *A* thus passes on the wheel pulses to the master counter. The lower input of block *B* is at the "0" level, so that this circuit lets no pulses through. Blocks *C* and *D* also both have an input at the "0" level; *C* obtains this voltage from the FF 10 block *M*, and *D* obtains it from the FF 12 block *G*; this gate is thus closed and no pulses get through to the auxiliary counter.

Let us now first assume that the length cut off is equal to the preset value of, say, 2100 mm. The master counter will then read 2100 at the moment the blade pulse appears. This pulse activates the bistable circuit FF 12 *G* via the PS 10 *F*, causing the FF 12 *G* to change polarity at the output. At the same time the positive voltage now appears at the output of the thumbwheel switches used for presetting, so that the right-hand *S* input of the FF 10 *M* is brought to the "0" level via GI 10 *O* (used as an inverting amplifier), thus causing the FF 10 *M* to switch over. The switch-over of *G* has the effect of blocking *A*; *B* would be opened by the "1" level of the left-hand output of *G*, but is at the same time blocked again by the "0" level of the left-hand output of *M*. In the same way as *B*, both *C* and *D* remain blocked. Neither of the counters receive any further pulses.

The reading of the master counter must now be shown by indicator tubes. For this purpose the counter contents are first stored in a register in order to release the counter for the next measurement. The register contains as many bistable circuits as the counter (16, four for each decade). These bistable circuits are connected to four decoding circuits (fig. 9), which control the indicator tubes. For transferring the contents of the counter to the register, the outputs of the bistable circuits in the counter are each connected to the corresponding condition input of the bistable circuits in the register; the trigger inputs of the register are interconnected. The transfer is effected by supplying a pulse to these trigger inputs (it is thus comparable with one step of a shift register).

For transferring the contents of the counter into the register, the voltage change from "1" to "0" at the right-hand output of FF 12 *G* is used. This actuates the PD 10 circuit block *I*, whose output voltage goes to the "0" level and then, after a slight delay, returns to the "1" level. The latter voltage pulse, after being inverted by GI 10 *J* and amplified by GA 10 *N*, is used for the transfer. The delay is necessary because otherwise the blade pulse might appear immediately after, say, the thousandth wheel pulse; in that case the counter must still change from 999 to 1000 before the contents are stored in the register, which would take a relatively long time because of the four transfers. The delay does not upset the next measurement because, as we shall explain presently, only every other sheet is measured. This can also be seen with the PD 10 *H*. At the first blade pulse the input of this circuit block goes from the "0" to the "1" level. This has no effect, however; only when the second pulse arrives, and the voltage drops again, does *H* deliver a pulse which is used for resetting the counters to zero and returning FF 10 *M* to the initial state.

Let us now consider the case where the sheet of cardboard cut off is longer than the preset value. The pulses are again passed in the same way to the master counter, but now the preset value is reached before the blade pulse appears, and FF 10 *M* thus switches over before FF 12 *G*. The lower input of GI 11 *C* has now arrived at the "1" level, while its upper input is still at the "1" level. This means that *C* is open and the next wheel pulses are supplied to the auxiliary counter as well as to the master counter. If the blade pulse now appears, the FF 12 *G* also switches over, whereupon all NAND circuits are again blocked (*A* and *C* by the "0" level of *G*, and *B* and *D* by the "0" level of *M*), so that all counters stop. The reading of the master counter is again shown on the indicator tubes as a result of the switch-over of *G*.

The auxiliary counter has meanwhile received a

number of pulses. If the positive tolerance limit has been exceeded, the relevant thumbwheel switch delivers a corresponding positive output voltage at that reading. Both inputs of GI 10 *L* are at the "1" level (since FF 12 *G* is still in the state indicated in fig. 19) so that the output *L* of the circuit block goes for a short time to the "0" level.

Finally, there is the case where the sheet is cut off shorter than the preset value. FF 12 *G* will then switch over before FF 10 *M*. As soon as *G* switches over, the contents of the master counter will again be taken over into the register in the manner described above, so that the measured length will be indicated. Now, however, the master counter does not stop yet. Although *A* is blocked by the switch-over of *G*, block *B* at the same moment starts to let pulses through (the lower input being at the "1" level because of the state of *G* and the upper input at the "1" level because of the state of *M*). Simultaneously the GI 11 *D* starts to pass pulses to the auxiliary counter.

As soon as the master counter has reached the preset value, circuit *M* changes state, so that just as above, all the NAND circuits are blocked and the counters stopped.

The auxiliary counter has now counted the number of millimetres by which the sheet cut off is too short. If this has reached the negative tolerance limit, the voltage at both inputs of the GI 10 *K* will just have been at the "1" level, resulting in the appearance of a pulse at the output *K* of this block.

In the last case we allowed the master counter to go on counting after the blade pulse had appeared. In other words, we have used for the measurement pulses which should really have been used for measuring the length of the next sheet. With this arrangement, therefore, we can only measure every second sheet, but in view of the slow variations this is not a drawback. Some of the consequences of this have already been noted above.

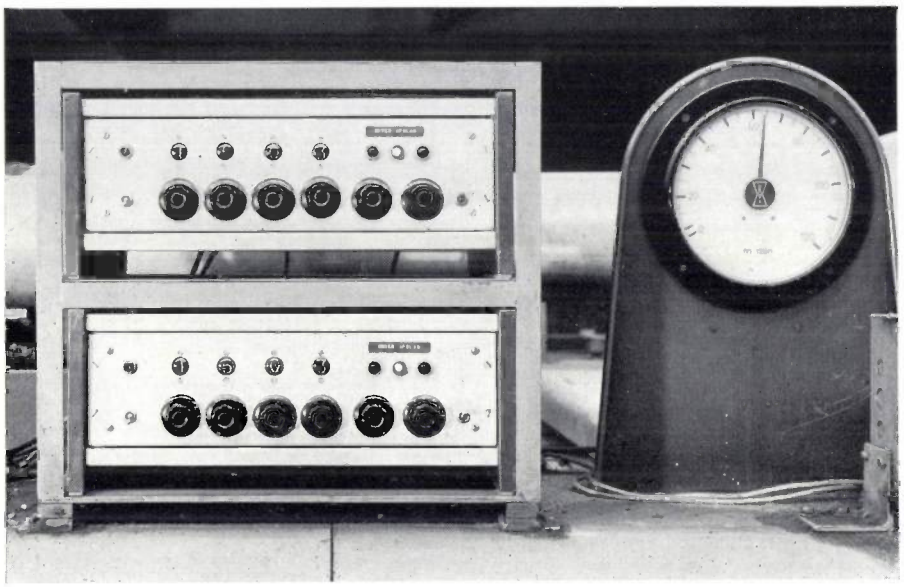


Fig. 20. The cabinets containing the digital circuits (one complete circuit for each pair of cutting blades; cf. fig. 19). The indicator tubes and the indicator lamps can be seen on each cabinet, and below them the switches for presetting the length and the tolerances (in this model the switches were similar to those of fig. 10 but with knobs instead of thumbwheels). The meter at the right of the cabinets indicates the speed of the cardboard strip.

A red lamp lights up if a measured sheet is too long or too short (*L* or *K* respectively in fig. 17); a green lamp is illuminated when the length is within the tolerances (*N* in fig. 17). The lamps are controlled by a circuit which is fed by the reset (zero) pulses and the pulses appearing at outputs *L* and *K*. We shall not go into the operation of this circuit since it offers no new information about the application of the circuit blocks.

The digital circuits add up in total to 65 circuit blocks, mounted on nine printed wiring boards. Together with the power pack they are contained in a cabinet (*fig. 20*) which measures only $20 \times 30 \times 50$ cm and can easily be fitted to the corrugated-cardboard machine. *Fig. 21* shows the interior of the cabinet.

The equipment described here has proved to be exceptionally reliable. High safety factors are built into

the design of the circuit blocks, and if they are used in a correct circuit, equipment results which requires no technical attention after installation. The equipment has been in operation for the last three years in the corrugated-cardboard works without ever giving any trouble. Maintenance is limited to periodically cleaning the rotating parts of the wheel-pulse and blade-pulse generators.

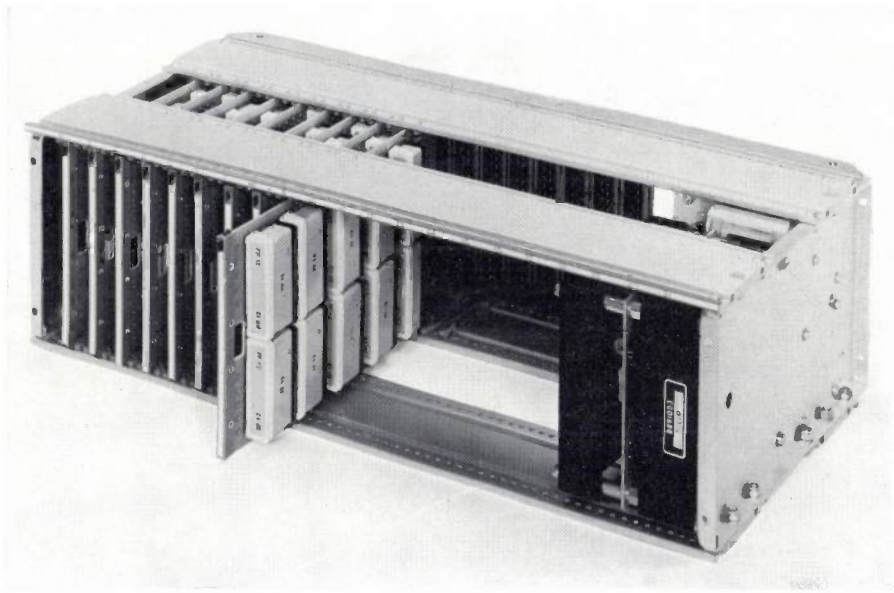


Fig. 21. Interior of one of the cabinets of fig. 20, seen from the rear. The printed wiring boards with circuit blocks are fitted in a rack of standard size (20 × 30 × 50 cm), which also accommodates the power pack on the right. One of the boards has been partly pulled out from the rack. The boards are connected together by means of the connectors at the front of the rack. Since nine boards only are required for the circuit blocks used in the equipment for one pair of cutting blades, not all the space in the cabinet is occupied.

Summary. Digital circuit blocks contain complete basic digital circuits, such as logic circuits, bistable circuits, etc. The series of circuit blocks marketed by Philips have been described in a previous article in this journal. The present article describes how a number of frequently used circuits can be built up with these circuit blocks. After a brief discussion of logic circuits, some applications of bistable circuits are described. These include decimal counters, the decoding of these counters for indication, the presetting of counters by means of thumbwheel switches, and shift registers. Some applications of a circuit block containing a Schmitt trigger are considered, and this is followed by a discussion of various input and output devices which can be used to form the link between a controlled process and a digital circuit. As an example of a piece of industrial equipment made up from circuit blocks, the last part of the article gives a description of an equipment for measuring the length of sheets cut off in the manufacture of corrugated cardboard. In the manufacturing process a continuous strip of cardboard is cut into sheets by a system of cutter blades on rotating cylinders. The speed of the cardboard strip and the speed of rotation of the cylinders are measured, giving the input signals from which the equipment determines the length cut off. This is compared with the required length, which can be preset in the measuring unit. If the difference is greater than a certain tolerance (also preset), the equipment puts a warning light on. The use of circuit blocks gives a compact equipment which is extremely reliable and requires virtually no maintenance.

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, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, 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.

G. A. Acket & H. 't Lam: Behaviour of *n*-type gallium arsenide in strong microwave fields.
Electronics Letters **3**, 258-259, 1967 (No. 6). *E*

A. C. Aten: Conductivity range and absorption edge of zinc seleno-tellurides.
J. Phys. Chem. Solids **28**, 1340-1342, 1967 (No. 7). *E*

P. Baeyens & G. Krijl: Philips formaldehyde bright tin.
Trans. Inst. Metal Finishing **45**, 115-121, 1967 (No. 3).

P. Beekenkamp: De structuur van glasachtige boraten en borofosfaten.
Klei en Keramiek **17**, 230-249, 1967 (No. 8). *E*

H. C. J. van Beukering & H. H. M. van der Aa: A rolling diaphragm seal for high pressures and high speeds.
3rd Int. Conf. on Fluid Sealing, Cambridge, England, 1967, p. G4 21-35. *E*

G. Blasse: The absorption of the Cr^{3+} ion in host lattices containing pentavalent cations.
J. inorg. nucl. Chem. **29**, 1817-1820, 1967 (No. 7). *E*

G. Blasse: Magnetische wisselwerking via verschillende diamagnetische ionen.
Chem. Weekblad **63**, 361-362, 1967 (No. 32). *E*

G. Blasse & A. Bril: Energy transfer from trivalent rare earth ions to Cr^{3+}.
Physics Letters **25A**, 29-30, 1967 (No. 1). *E*

G. Blasse & J. de Vries: A new family of lanthanide compounds: lithium lanthanide silicates and germanates.
J. inorg. nucl. Chem. **29**, 1541-1542, 1967 (No. 6). *E*

G. Blasse & J. de Vries: On the Eu^{3+} fluorescence of mixed metal oxides, VI. Temperature dependence of the fluorescence.
J. Electrochem. Soc. **114**, 875-877, 1967 (No. 8). *E*

R. Bleekrode: Absorption and emission spectroscopy of C_2, CH and OH in low-pressure oxyacetylene flames.
Thesis, Amsterdam 1966. *E*

R. Bleekrode & C. van Trigt: Spectroscopic determination of $Cs(6^2P_{1/2})$ excited state concentrations in Cs-Ar low-pressure gas discharges.
Appl. Phys. Letters **10**, 352-354, 1967 (No. 12). *E*

P. F. Bongers: Toepassing van de ligandenveldtheorie in de kristalchemie.
Chem. Weekblad **63**, 353-361, 1967 (No. 32). *E*

J. Borne: Dispositif électronique d'aide à l'enseignement.
Onde électr. **47**, 682-685, 1967 (No. 482). *L*

G.-A. Boutry: La télévision dans l'invisible.
Rev. gén. Electr. **76**, 1018-1026, 1967 (No. 7/8). *L*

P. C. Brandon: Temperature features of enzymes affecting Crassulacean acid metabolism.
Plant Physiology **42**, 977-984, 1967 (No. 7).

J. van den Broek: The electrical behaviour of vapour-deposited lead-monoxide layers.
Philips Res. Repts. **22**, 367-374, 1967 (No. 4). *E*

A. Broese van Groenou, J. L. Page & R. F. Pearson: Magnetic after-effect rotational hysteresis — theory and experiment on Si-doped YIG.
J. Phys. Chem. Solids **28**, 1017-1025, 1967 (No. 6). *M*

A. Broese van Groenou & R. F. Pearson: Anisotropy relaxation in manganese ferrous ferrites.
J. Phys. Chem. Solids **28**, 1027-1036, 1967 (No. 6). *M*

G. Brouwer: Control of the surface potential of germanium by means of a variable *p*H electrolyte containing hydrogen peroxide and potassium chloride.
J. Electrochem. Soc. **114**, 743-748, 1967 (No. 7). *E*

H. A. C. M. Bruning: Homogeneity regions and superconducting transition temperatures in the system V-V_3Si.
Philips Res. Repts. **22**, 349-354, 1967 (No. 4). *E*

K. H. J. Buschow & J. F. Fast: Crystal structure and magnetic properties of some rare earth germanides.
Phys. Stat. sol. **21**, 593-600, 1967 (No. 2). *E*

P. Chevalier: Photomultiplicateur à haute résolution utilisant un multiplicateur semi-conducteur.
Nucl. Instr. Meth. **50**, 346-348, 1967 (No. 2). *L*

B. H. Clarke: The anomalous ferrimagnetic resonance properties of ytterbium-doped yttrium iron garnet.
Brit. J. appl. Phys. **18**, 727-738, 1967 (No. 6). *M*

A. Cohen: On distinctiveness as a criterion in language analysis.
Acta linguistica hafn. **9**, 179-191, 1966 (No. 2). *E*

A. Cohen: On methods in the analysis of speech perception.
Synthese **17**, 331-343, 1967. *E*

H. J. van Daal & A. J. Bosman: Hall effect in CoO, NiO, and α-Fe_2O_3.
Phys. Rev. **158**, 736-747, 1967 (No. 3). *E*

A. M. van Diepen, H. W. de Wijn & K. H. J. Buschow: Nuclear magnetic resonance in rare-earth-aluminum intermetallic compounds: RAl_3.
J. chem. Phys. **46**, 3489-3492, 1967 (No. 9). *E*

J. A. W. van der Does de Bye: The role of acceptors in the luminescence of *n*-type GaAs.
J. Phys. Chem. Solids **28**, 1485-1491, 1967 (No. 8). *E*

F. C. M. Driessens: Place and valence of the cations in Mn_3O_4 and some related manganates.
Inorg. chim. Acta **1**, 193-201, 1967 (No. 1). *E*

G. Engelsma: Photoinduction of phenylalanine deaminase in gherkin seedlings, I. Effect of blue light.
Planta **75**, 207-219, 1967 (No. 3). *E*

W. Ermrich & A. van Oostrom: Experimental evidence for tunnel-resonance effects in field electron emission.
Solid State Comm. **5**, 471-474, 1967 (No. 6). *A, E*

J.-M. Goethals: Cyclic error-locating codes.
Information and Control **10**, 378-385, 1967 (No. 4). *B*

C. A. A. J. Greebe, W. F. Druyvesteyn & W. J. A. Goossens: Use of electromagnetic resonances in flat metal boxes for the measurement of d.c. size effects.
Physics Letters **24A**, 727-728, 1967 (No. 13). *E*

C. Haas, A. M. J. G. van Run, P. F. Bongers & W. Albers: The magnetoresistance of *n*-type $CdCr_2Se_4$.
Solid State Comm. **5**, 657-661, 1967 (No. 8). *E*

W. van Haeringen: Polarization properties of a single-mode operating gas laser in a small axial magnetic field.
Phys. Rev. **158**, 256-272, 1967 (No. 2). *E*

J. Haisma: Enige karakteristieke eigenschappen van het actieve medium van gaslasers.
Ned. T. Natuurk. **33**, 105-124, 1967 (No. 4/5). *E*

P. Hansen, J. K. Vogel & G. Winkler: Das Verhalten der Spinwellenlinienbreite beim Einbau von Seltenen Erden in YIG und von Übergangsmetallen in Spinellferrite.
Z. angew. Physik **23**, 190-193, 1967 (No. 3). *H*

P. A. H. Hart & L. Kranenbarg: Crystal-mixer noise as a function of temperature at 1400 MHz.
Philips Res. Repts. **22**, 402-418, 1967 (No. 4). *E*

J. Hasker: Short-length oscilloscope tubes with a high deflection sensitivity by bending the electron beam.
Philips Res. Repts. **22**, 419-442, 1967 (No. 4). *E*

H. Hörster, H. Lydtin, O. Reifenschweiler & K. G. Fröhner: Bildung und Transport kleiner Teilchen bei Verdampfungsvorgängen in Gegenwart von Inertgasen.
Z. angew. Physik **22**, 203-208, 1967 (No. 3). *A, E*

R. J. Klein Wassink: Enkele fundamentele aspecten van de bevochtiging bij het solderen.
Lastechniek **33**, 173-175, 1967 (No. 9). *E*

H. Koelmans & H. C. de Graaff: Drift phenomena in CdSe thin film FET's.
Solid-State Electronics **10**, 997-1005, 1967 (No. 10). *E*

J. A. Kok, J. W. Poll & C. E. G. van Vroonhoven: Electrostatic purification of insulating liquids.
Appl. sci. Res. **17**, 461-478, 1967 (No. 6). *E*

W. Lems: Magnetic properties of electrodeposited FeNi films, I.
Philips Res. Repts. **22**, 388-401, 1967 (No. 4). *E*

P. Leuthold: Filternetzwerke mit digitalen Schieberegistern.
Promotionsarbeit, Zürich 1967. *E*

J. Liebertz & C. J. M. Rooymans: Phase behaviour of Li_2MoO_4 at high pressures and temperatures.
Solid State Comm. **5**, 405-409, 1967 (No. 5). *A, E*

M. H. van Maaren, G. M. Schaeffer & F. K. Lotgering: Superconductivity in sulpho- and selenospinels.
Physics Letters **25A**, 238-239, 1967 (No. 3). *E*

H. A. Meijer & J. J. de Jong: Einfluß der Austenitisierungsbedingungen und einer Tiefkühlbehandlung auf die mechanischen Eigenschaften von gehärtetem Werkzeugstahl.
Archiv Eisenhüttenw. **38**, 735-742, 1967 (No. 9). *E*

E. A. Muijderman: Analysis and design of spiral-groove bearings.
Trans. ASME: J. Lubrication Technology **1**, 291-306, 1967 (No. 3). *E*

G. T. M. Neelen: De Philips-Stirlingmotor. Van klassiek idee tot moderne krachtbron.
Schip en Werf **34**, 407-421, 1967 (No. 17). *E*

A. K. Niessen, F. A. Staas & C. H. Weijsenfeld: The Hall angle of various superconducting niobium-tantalum alloys in the mixed state.
Physics Letters **25A**, 33-35, 1967 (No. 1). *E*

C. van Opdorp & J. Vrakking: Photo-effects in isotype heterojunctions.
Solid-State Electronics **10**, 955-971, 1967 (No. 9). *E*

P. Penning: Dynamical theory for simultaneous X-ray diffraction.
Adv. in X-ray Anal. **10**, 67-79, 1967. *E*

G. Pietri: Les photomultiplicateurs en physique nucléaire: limites qu'ils introduisent dans la précision des mesures.
Bull. Instr. nucl. No. 29, suppl. au Bull. Inform. sci. techn. C.E.A. No. 115, p. 36-50, 1967. *L*

D. Polder & W. van Haeringen: Circular polarization in a $j = 1 \rightarrow j = 0$ transition laser.
Physics Letters **25A**, 337-338, 1967 (No. 4). *E*

H. J. Prins: The distribution of the brightness of stars. A Poisson process with "time"-dependent parameter.
Statistica neerl. **21**, 131-149, 1967 (No. 2). *E*

A. Rabenau, H. Rau & P. Eckerlin: A halogen-deficient phase in the system tellurium-iodine.
Angew. Chemie: Int. Edit. in English **6**, 706, 1967 (No. 8); German Edit. **79**, 688, 1967 (No. 15). *A*

H. Rau & A. Rabenau: Hydrothermalsynthese von CuS und CuSe in 48%iger HBr.
Mat. Res. Bull. **2**, 609-614, 1967 (No. 6). *A*

J. C. Richard & P. Saget: Diagrammes de diffraction d'électrons lents obtenus sur monocristaux d'halogénures alcalins au moyen d'un dispositif expérimental simple.
Le Vide **22**, 99-104, 1967 (No. 128). *L*

C. Romers, C. J. M. Rooymans & R. A. G. de Graaf: The preparation, crystal structure and magnetic properties of $Na_3Fe_5O_9$.
Acta cryst. **22**, 766-771, 1967 (No. 6). *E*

H.-D. Rüpke: Magnetodynamische Eigenschwingungen in anisotropen Ferriteinkristallen.
Z. angew. Physik **23**, 194-199, 1967 (No. 3). *H*

P. Schnabel: Vierpunktmethode zur Messung der elektrischen Widerstandsanisotropie.
Z. angew. Physik **22**, 136-140, 1967 (No. 2). *A*

J. F. Schouten: Behavior, physiology and model.
Communication: Concepts and Perspectives, 2nd Int. Symp. Comm. Theory Res., 1966, p. 181-190; 1967. *E*

G. Schulten & J. P. Stoll: A high-precision wideband wavemeter for millimeter waves.
IEEE Trans. **MTT-15**, 430-431, 1967 (No. 7). *H*

H. Severin & J. P. Stoll: Permeabilität einiger Ferrite mit magnetischen Verlusten im Bereich der Zentimeter- und Millimeterwellen.
Z. angew. Physik **23**, 209-212, 1967 (No. 3). *H*

M. Sintzoff: Existence of a van Wijngaarden syntax for every recursively enumerable set.
Ann. Soc. Scient. Bruxelles **81**, 115-118, 1967 (No. 2). *B*

L. A. Æ. Sluyterman: Reversible inactivation of papain by cyanate.
Biochim. biophys. Acta **135**, 439-449, 1967 (No. 3). *E*

L. A. Æ. Sluyterman: The effect of methanol, urea and other solutes on the action of papain.
Biochim. biophys. Acta **139**, 418-429, 1967. *E*

L. A. Æ. Sluyterman: The activation reaction of papain.
Biochim. biophys. Acta **139**, 430-438, 1967. *E*

D. R. Tilley: Proximity effect of a superconductor on an antiferromagnetic metal.
Physik kondens. Materie **6**, 236-242, 1967 (No. 3). *M*

A. Venema: Thermionic emission.
Handbook of Vacuum Physics, ed. A. H. Beck, Vol. 2, Parts 2/3, p. 179-298; Pergamon Press, Oxford 1966. *E*

J. Verweel: Über die Blochwände in Zn_2Y-Einkristallen.
Z. angew. Physik **23**, 200-204, 1967 (No. 3). *H*

W. L. Wanmaker & J. W. ter Vrugt: Luminescence of alkaline-earth pyrophosphates, activated with divalent europium.
Philips Res. Repts. **22**, 355-366, 1967 (No. 4).

J. G. W. van de Waterbeemd & G. W. van Oosterhout: Effect of the mobility of metal atoms on the structure of thin films deposited at oblique incidence.
Philips Res. Repts. **22**, 375-387, 1967 (No. 4). *E*

C. Weber: Analogue and digital methods for investigating electron-optical systems.
Thesis, Eindhoven 1967. *E*

H. W. Werner & J. M. Nieuwenhuizen: On the distribution of AgBr grains in photographic plates for ion detection in mass-spectrometers.
Z. Naturf. **22a**, 1035-1039, 1967 (No. 7). *E*

W. J. Witteman: A sealed-off Michelson type CO_2 laser for diagnostic studies of gaseous plasmas.
Appl. Phys. Letters **10**, 347-349, 1967 (No. 12). *E*

H. C. Wright, R. E. Hunt & G. A. Allen: Crystal efficiency for injection luminescence in GaAs.
Solid-State Electronics **10**, 633-639, 1967 (No. 7). *M*

PHILIPS TECHNICAL REVIEW

Volume 29, 1968, No. 2

Non-polar electrolytic capacitors

J. C. W. Kruishoop

Electrolytic capacitors are widely used on account of their large capacitance for a relatively small volume. One well-known application is their use as smoothing capacitors in the power supply circuits of radio receivers. For many other applications, however, particularly where an alternating voltage is applied to the capacitor, they have until now not been suitable, since this type of capacitor always had to be connected with the proper polarity. A further development is the non-polar electrolytic capacitor described here, which does not have this disadvantage and has the same relatively large capacitance as a conventional electrolytic capacitor.

In this article we describe a new type of electrolytic capacitor which, unlike conventional electrolytic capacitors, can be connected up with either polarity, and which we have therefore called non-polar. The electrolyte in such a capacitor has to meet two essential requirements: it should contain no hydrogen ions and it should have self-healing properties in *both* voltage directions. An example of such an electrolyte is a solution of anhydrous calcium nitrate in pyridine. We begin with a brief description of the formation and leading characteristics of conventional electrolytic capacitors, and of the mechanism responsible for their polar nature.

The conventional electrolytic capacitor

Electrolytic capacitors have a large capacitance in a relatively small volume (capacitances lie between about 1 μF and 1 F), and for this reason they are widely used in electrical and electronic circuits. Electrolytic capacitors have therefore been a subject of research at Philips since 1932 [1]. The difference in size between a paper capacitor and an electrolytic capacitor with the same capacitance and working voltage is illustrated in *fig. 1*.

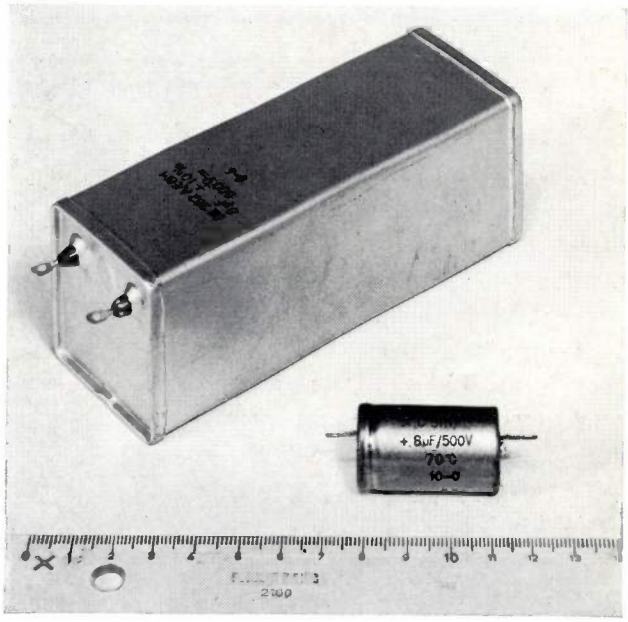


Fig. 1. A paper capacitor (above) and an electrolytic capacitor of the same capacitance and working voltage.

A schematic diagram of an electrolytic capacitor is shown in *fig. 2*: two electrodes are contained in an electrolytic solution, one of them (electrode *A*) being coated with a thin insulating oxide film. This oxide film functions as the dielectric and is bounded by the electrode *A* and the electrolyte *El* which act as the plates of

Drs. J. C. W. Kruishoop is with Philips Research Laboratories, Eindhoven.

[1] See W. Ch. van Geel and A. Claassen, Philips tech. Rev. **2**, 65, 1937.

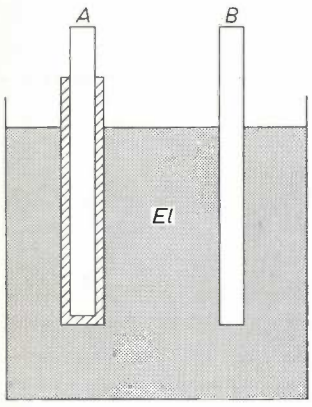


Fig. 2. Schematic diagram of a conventional electrolytic capacitor. *A* electrode covered with an insulating oxide film. *B* counter electrode. *El* electrolyte. The electrodes are usually of aluminium. The electrolyte used is frequently water or glycol containing a mixture of boric acid and sodium borate or ammonium borate.

a capacitor. Electrode *B* serves only as contact for the external connection. We shall henceforth refer to electrode *A* as the oxidized electrode and to electrode *B* as the counter electrode. The capacitance of a capacitor is given by the well-known relation:

$$C = \frac{\varepsilon A}{3.6\pi d},$$

where C is the capacitance (in picofarads), ε the dielectric constant, A the area of the plates (in cm^2) and d the thickness (in cm) of the dielectric. The capacitance of electrolytic capacitors is so very large because the thickness d is very small (0.01-1 μm), while the film is still capable of withstanding breakdown at the voltages used.

The very thin dielectric film is obtained by electrolytic oxidation of a suitable metal, usually pure aluminium or tantalum. The oxidation process may be described as follows (see *fig. 3*). The metal is given a positive potential with respect to the electrolyte, the voltage applied being higher than that at which the capacitor will be used. Under the influence of this voltage, the metal ions migrate through the existing naturally present oxide film to the electrolyte side. Here the ions combine with the oxygen of the anions from the electrolyte, causing the oxide film to grow to the required thickness [2].

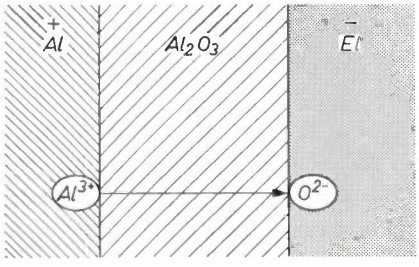


Fig. 3. Illustrating the anodic oxidation of aluminium.

Fig. 4 shows the relation between the field strength F and the current density J associated with the migration process. It can be seen in the figure that, at a field strength greater than 5×10^6 V/cm, a considerable current flows, so that the oxide film increases in thickness; this therefore determines the minimum field strength needed for making the capacitor. At a value lower than 5×10^6 V/cm the current is negligible, and so therefore is the growth of the oxide film. In other words, the current is blocked. At values of F between 5×10^6 V/cm and zero the oxide film is stable and can be used as a dielectric. This value of 5×10^6 V/cm is a

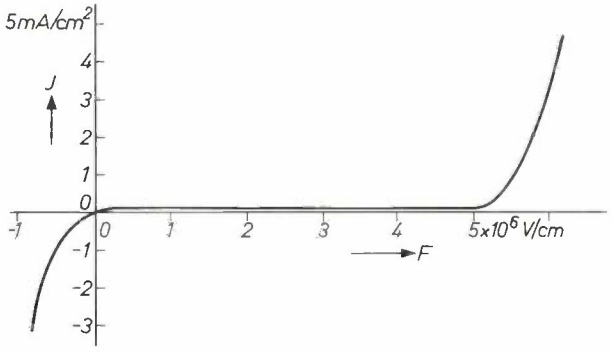


Fig. 4. Current density J measured as a function of field strength F in an oxide film formed by anodic oxidation of aluminium.

factor of about ten higher than the breakdown field strength of most other dielectric materials. This is due to the self-healing action of the electrolytic capacitor: in the event of damage due to breakdown, the oxide film reforms by oxidation of the underlying metal where the film was damaged. The maximum permissible working voltage V_{max} in an electrolytic capacitor depends on the thickness of the oxide film and varies in practice from 6 to 500 V. If the surface area is held constant, the product CV_{max} of the permissible voltage and the capacitance is approximately constant.

However, if only a relatively small field of the

[2] From recent work it seems likely that oxygen ions can also migrate in the reverse direction through the oxide film. This makes no essential difference however to what we have said about the growth of the film.

[3] Sometimes referred to as "non-polarized" or "symmetrical" electrolytic capacitors.

[4] L. Young, Anodic oxide films, Academic Press, London 1961.
D. A. Vermilyea, Adv. Electrochem. electrochem. Engng. **3**, 211, 1963.
Extended Abstracts of the Spring Meeting of the Electrochemical Society, Dallas 1967, Dielectrics and Insulation Division.
W. S. Goruk, L. Young and F. G. R. Zobel, Ionic and electronic currents at high fields in anodic oxide films, in: Modern Aspects of Electrochemistry No. 4 (ed. J. O'M. Bockris), Butterworths, London 1966, pp. 176-250.

[5] The complete results will be published shortly in J. Electrochem. Soc.

[6] Arguments in support of this view are summarized in: A. Middelhoek, J. Electrochem. Soc. **111**, 379, 1964.

opposite polarity is applied, the result is a high leakage current which permanently damages the oxide film and short-circuits it. The reason for this will be discussed under the next heading.

For applications where this single polarity is a disadvantage it has been the practice to use what are called bipolar electrolytic capacitors [3]. These consist essentially of two conventional electrolytic capacitors connected in series in such a way that the current is blocked in both directions. This improvement is obtained, however, at the expense of the CV_{max} product which is reduced by a factor of about two (*fig. 5*). The non-polar electrolytic capacitor does not have this disadvantage.

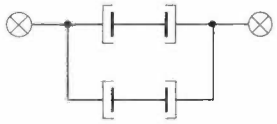


Fig. 5. Two conventional electrolytic capacitors connected in series-opposition can be used for both polarities (bipolar). The capacitance, however, is then only half that of a separate electrolytic capacitor. To obtain the same capacitance it is therefore necessary to use two of these bipolar capacitors connected in parallel.

Mechanism underlying the polar nature of the conventional electrolytic capacitor

The polar nature, or rectifying action, of the conventional electrolytic capacitor has been the subject of numerous investigations and has given rise to many conflicting theories [4]. We shall confine ourselves here, however, to our own views [5].

The electrolytes used in conventional electrolytic capacitors all contain hydrogen ions. When a sufficiently high voltage is applied in the *forward direction* (i.e. in the reverse direction for use as a capacitor, the oxidized electrode then being negative), the hydrogen ions can migrate through the oxide film and are discharged at the metal-oxide interface [6] *(fig. 6)*. This mechanism has become clear to us chiefly through our investigations on oxidized *tantalum* electrodes. The

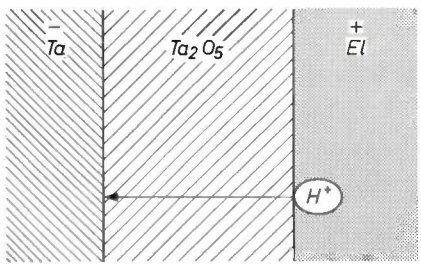


Fig. 6. Current conduction in the forward direction with Ta/Ta_2O_5 electrodes. This conduction is due to the migration of hydrogen ions through the Ta_2O_5. The same thing happens with Al/Al_2O_3 electrodes.

advantage of using tantalum in these investigations is that the oxide Ta_2O_5 is chemically much more stable than Al_2O_3, and this gave us much more freedom in the choice of electrolyte than when we used aluminium. Our investigations showed that the initial current of hydrogen ions, which we have called the primary forward current, is closely dependent on the thickness of the oxide film at a given voltage. After some time the oxide film is forced away from the metal locally by the pressure of the accumulated hydrogen which has been formed, giving a short-circuit between electrode and solution. This explains why the "secondary forward current" which occurs after some time is much stronger and practically independent of the thickness of the oxide film.

On the other hand, a well-chosen electrolyte contains no *negative* ions which could migrate through the oxide film when the polarity of the capacitor is reversed. When a voltage is applied in this *reverse direction*, therefore, little or no current flows and the capacitor can be used as such.

In order to verify the above theory of the rectifying action of the electrolytic capacitor we have devoted considerable attention to the question of whether the primary forward current was a current of hydrogen ions or of electrons (or perhaps holes). The possibility that other charge carriers might cause this current through the oxide film could fairly easily be ruled out for the following reasons:

a) The metal and oxygen ions forming part of the oxide film are mobile only at a field strength much higher than that applied in the forward direction (see fig. 4). Moreover, the displacement of these ions in the forward direction would cause the dissolution of the oxide film and thus make it thinner. This has never been observed.

b) The possibility that other ions than hydrogen ions in the electrolyte might cause the primary forward current was ruled out by varying the composition of the electrolyte: this had no effect on the rectifying action.

In order to determine whether the conduction is due to hydrogen ions or to electrons (or to both) we have used electrochemical methods, since the electrolytic capacitor can also be regarded as an electrochemical cell. If a forward current is passed through this cell, electrochemical reactions take place. For the oxidized electrode the following possibilities exist:

a) If the oxide film is an electron conductor, a redox reaction takes place at the surface of the film so that, for example, in the presence of Fe^{3+} ions we have:

$$Fe^{3+} + e \rightarrow Fe^{2+}.$$

The potential difference between the electrode and the

solution then depends upon the redox potential of the electrolyte. This redox potential is a measure of the ease with which the electrochemical reduction reaction takes place. To measure it we used a standard electrode as counter electrode; this is an electrode whose potential difference with the electrolyte is independent of the electrolyte composition. In our investigations we used a calomel standard electrode. If the conduction through the oxide film was purely electronic, the potential difference between the oxidized electrode and the calomel electrode would be equal to the redox potential plus a constant related to the film thickness and the magnitude of the forward current. By varying the electrolyte composition, and hence the redox potential, at a constant forward current it is possible to establish whether there is electronic conduction through the oxide film, and if so to what extent.

b) If the oxide film conducts hydrogen ions, this means that the electrolyte-oxide interface is selectively permeable to hydrogen ions. The electrode does not in that case differ essentially from other selectively permeable membrane electrodes, such as the glass electrode. In this case the potential difference between the oxidized electrode and the calomel electrode will depend upon the concentration of hydrogen ions in the electrolyte and will increase by 0.059 V per unit of change in pH just as for a glass electrode. This also can be experimentally verified by varying the pH of the electrolyte at constant current and redox composition.

In the experiments, special precautions had to be taken on account of the instability of the oxide film when carrying a current in the forward direction. After every measurement at constant forward current, the oxide film was kept for a few seconds at a reverse voltage low enough to prevent the oxide film from growing, but high enough for any breakdown spots to heal. This procedure was followed during the entire period of the measurements. Furthermore, the forward voltage was measured as the difference between the forward voltages across two electrolytic capacitors in which the forward current was the same, and which differed only in the composition of their electrolyte. The composition was kept constant in one capacitor and in the other it was varied both in pH and redox potential. As the oxidized electrodes were identical, the time-dependence of the forward voltage was eliminated (this time-dependence is related to polarization effects in the oxide film). A platinum electrode was used for the current supply, and in each measurement the forward voltage was measured by means of a calomel electrode. In spite of all these precautions, reproducible results were obtained only with thin oxide films (<40 nm) and low forward currents (<30 $\mu A/cm^2$).

The results of these measurements showed that the forward voltage depended only on the pH and not on the redox potential [7], thus proving that the primary forward current through the Ta_2O_5 layer consisted of hydrogen ions.

Investigation of electrolytes containing no hydrogen ions

It appeared from the work described above that a study should be made of the behaviour of electrolytic capacitors with electrolytes which contain no hydrogen ions. The solvents would have to be polar (i.e. the molecules must have a permanent dipole moment) to give solutions of good conductivity. An example of such a solvent, which we have used in many experiments, is pyridine (C_5H_5N). Pyridine has no hydrogen ions, is a highly stable compound and readily dissolves many salts, which are partly ionized in it, giving a conducting solution. We were particularly interested in the nitrates because we found that aluminium, and also tantalum, could be anodically oxidized in such solutions. The properties of the resultant oxide films in the reverse direction proved to be about the same as in conventional electrolytes.

The electrical behaviour in the other direction (i.e. with the oxidized electrode negative) now proved, however, to be strongly dependent upon the nature of the cation. We can distinguish here between the following cases, bearing in mind, however, that many nitrates cannot be prepared in an anhydrous form, or are insoluble in pyridine, and were not therefore investigated any further:

a) Nitrates of Li and Na give a high leakage current when the voltage on the oxidized electrode is negative. By analogy with the behaviour of Li^+ and Na^+ ions in glass we assumed that these ions migrate through the oxide film. The fact that this has not been observed in electrolytes containing hydrogen ions may be explained by the much greater mobility of the hydrogen ion: because of this greater mobility the voltage across the oxide film is too low to allow Li and Na to migrate.

b) Nitrates of heavy metals such as Ag and Pb also give a high leakage current when the voltage on the oxidized electrode is negative. In this case, however, the leakage is associated with deposition of the metals at discrete places on the oxide film. The ions of such heavy metals (with the exception of Ag) show no appreciable mobility in glass, and for this reason it is assumed that they do not migrate through the oxide film. The occurrence of a high leakage current is here probably attributable to breakdown at weak spots, which is followed by deposition of the metal.

c) Nitrates of large organic ions, e.g. tetraethylammonium nitrate $[N(C_2H_5)_4]^+NO_3^-$. In the first few

[7] For this purpose the redox potential was varied over an interval of not less than 1 volt.

minutes the leakage current is very low, but it then gradually increases. The explanation for this is that the very large tetraethylammonium ion is unable to migrate through the undamaged oxide film. Here also breakdown occurs after some time at weak spots in the oxide film, causing short-circuiting between metal and electrolyte at an increasing number of points. However, unlike the heavy metals, the organic compounds resulting from the discharge of tetraethylammonium ions are not good conductors, so that initially the leakage current stays low.

d) As already remarked, one of the reasons why oxide films are able to withstand such high field strengths when the oxidized electrode is at a *positive* voltage is the self-healing action of the electrolytic capacitor. In the event of breakdown the film is repaired by the further anodic oxidation of the underlying metal. In cases (a), (b) and (c) there is no such self-healing electrode reaction when the voltage at the oxidized electrode is *negative*.

However, with this voltage direction a self-healing action is obtained when calcium nitrate is used. It was found that any metal cathode in a solution of calcium nitrate in pyridine becomes coated with an insulating film of calcium oxide during electrolysis. The same electrode reaction can be used for protecting an oxidized aluminium electrode against breakdown when the voltage at the oxidized electrode is negative.

In a later phase of the investigation, using solvents other than pyridine, we found that nitrates of Mg, Sr, Ba, etc., are also capable of the same self-healing electrode reaction.

Practical realization of a non-polar electrolytic capacitor

An electrolyte of the type mentioned above can be used to make a non-polar electrolytic capacitor, i.e. a capacitor that can be connected up with either polarity and is relatively much smaller than the bipolar type.

In this application, the following properties were necessary, and in most cases sufficient:

a) The absence of water and other solvents containing hydrogen ions, such as alcohols.

b) Good electrical conductivity: this means that solvents must be polar.

c) The absence of all cations except Mg, Ca, Sr, Ba, Al and rare earths, which are needed for the repair of the dielectric when the oxidized electrode is negative.

d) Presence of oxidizing ions—particularly NO_3^- ions — for repair of the dielectric when the oxidized electrode is positive.

In addition, a high boiling point, a low melting point and chemical stability are desirable.

Examples of electrolytes meeting these requirements, which give sufficiently low leakage currents for both directions of voltage and which have a low resistivity are solutions of calcium nitrate or magnesium nitrate in dimethylacetamide or in dimethylsulphoxide. Aluminium has proved to be by far the best metal for the oxidized electrode. In a suitable electrolyte, aluminium oxide films withstand practically the same field strength at a negative voltage as at a positive voltage. The field strength which tantalum oxide films can withstand at a negative voltage is only 30% of that which they can withstand at a positive voltage. The reason for this is not yet understood.

In the manufacture of non-polar capacitors the same procedures can be used as in the manufacture of conventional electrolytic capacitors, the conventional electrolyte being replaced only in the last stage of manufacture by one of the new electrolytes described here.

The other characteristics of non-polar electrolytic capacitors are much the same as those of conventional types.

Summary. Non-polar electrolytic capacitors share with conventional types the advantage of a large capacitance per unit volume and they have the added advantage that they can be connected either way round in the circuit. It is shown that the high leakage current in the forward direction of the conventional electrolytic capacitor is caused by the migration of hydrogen ions of the electrolyte through the oxide film on the electrode. A study was made of the behaviour of electrolytic capacitors in which electrolytes which did not contain hydrogen ions were used. With a suitable solvent and salt a capacitor is produced which has a low leakage current and is self-healing for both directions of the voltage. Nitrates are suitable for the self-healing action in one direction, and ions of Mg, Ca, Sr, Ba, Al or rare earths are required for self-healing in the other direction.

A universal vectorcardiograph

In conventional electrocardiography the potential difference which appears between two points on the surface of the human body during the heartbeat is recorded as a function of time. A number of such electrocardiograms recorded for various pairs of points, which are referred to as "leads", provide the cardiologist with information about abnormalities in the action of the heart.

The discovery that the phase difference between the time functions obtained from different leads is also of importance led to the development of vectorcardiography [1]. If the potential differences from *two* leads are plotted one against the other, the result is a kind of Lissajous figure. If the time scale is marked on this figure, we then have another way of displaying information about the action of the heart. To a first approximation the Lissajous figures can be regarded as representing, for all possible pairs of points, projections of one and the same curve described in space and characterizing the heart under examination. The radius vector of this three-dimensional curve is called the instantaneous heart vector; the curve displays the variation of the heart vector during the heartbeat, hence the term "vectorcardiogram".

The vectorcardiogram has proved to be a valuable aid in diagnosis, as it is closely correlated with the action of the human heart, and will show if any abnormal condition is present. Individual loops in the vectorcardiogram correspond to the different phases in the heartbeat, each of which gives a peak in a conventional electrocardiogram; see *fig. 1*. The non-specialized physician can learn to interpret the vectorcardiogram more easily than a conventional electrocardiogram. The vectorcardiogram has specific advantages for the recognition of certain disorders, whereas the conventional electrocardiogram remains the best means of investigating rhythmic disturbances. The two methods complement one another.

The Philips Medical Service at Eindhoven is now using an experimental vectorcardiograph, developed at the Philips Research Laboratories, which makes it very easy for the doctor to display and study the vectorcardiogram in a variety of forms (*fig. 2*). A brief description of this equipment now follows. (We should add that it is not commercially available.)

A number of electrodes are applied at appropriate positions to the patient under investigation, and three or more independent leads are taken from the electrodes. Lissajous figures can be produced from these leads in the way described above. In order to interpret these figures, it is desirable, however, that they should be reduced to projections of the three-dimensional curve described by the heart vector on the three planes of a rectangular co-ordinate system (X, Y, Z). This is done by means of a linear transformation, i.e. appropriate fractions of the potential differences from all the leads are added together in a mixing circuit so as to obtain signals which are proportional to the components X, Y and Z of the heart vector. The correct fractions of each of the original signals for a selected lead system are determined empirically, and provide the data for adjusting the values of the resistors in the mixing circuit into which the signals are fed after amplification. The pairs of component XY, XZ and YZ are then applied to the deflection plates of three cathode-ray tubes, one tube for each pair, each tube thus indicating one of the required projections of the vectorcardiogram.

There are at present various internationally standardized lead systems in use. The number of leads depends on the system employed. Our vectorcardiograph is designed to be used with up to six leads. With this arrangement the cardiologist is able to choose between two standard lead systems incorporated in the instrument (the Frank system and the Burger-Wilson system) and a third which he can decide upon for himself. If all the electrodes have been applied previously

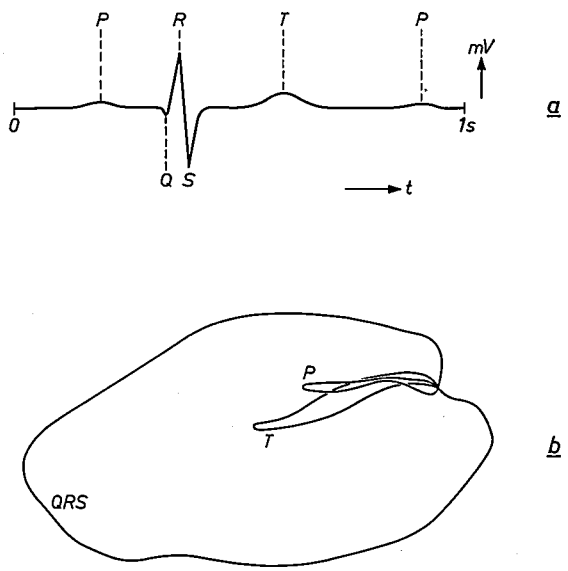


Fig. 1. *a*) Conventional electrocardiogram. *b*) Perspective sketch of a vectorcardiogram. The peaks in the conventional electrocardiogram are indicated in the same way as the corresponding loops in the vectorcardiogram.

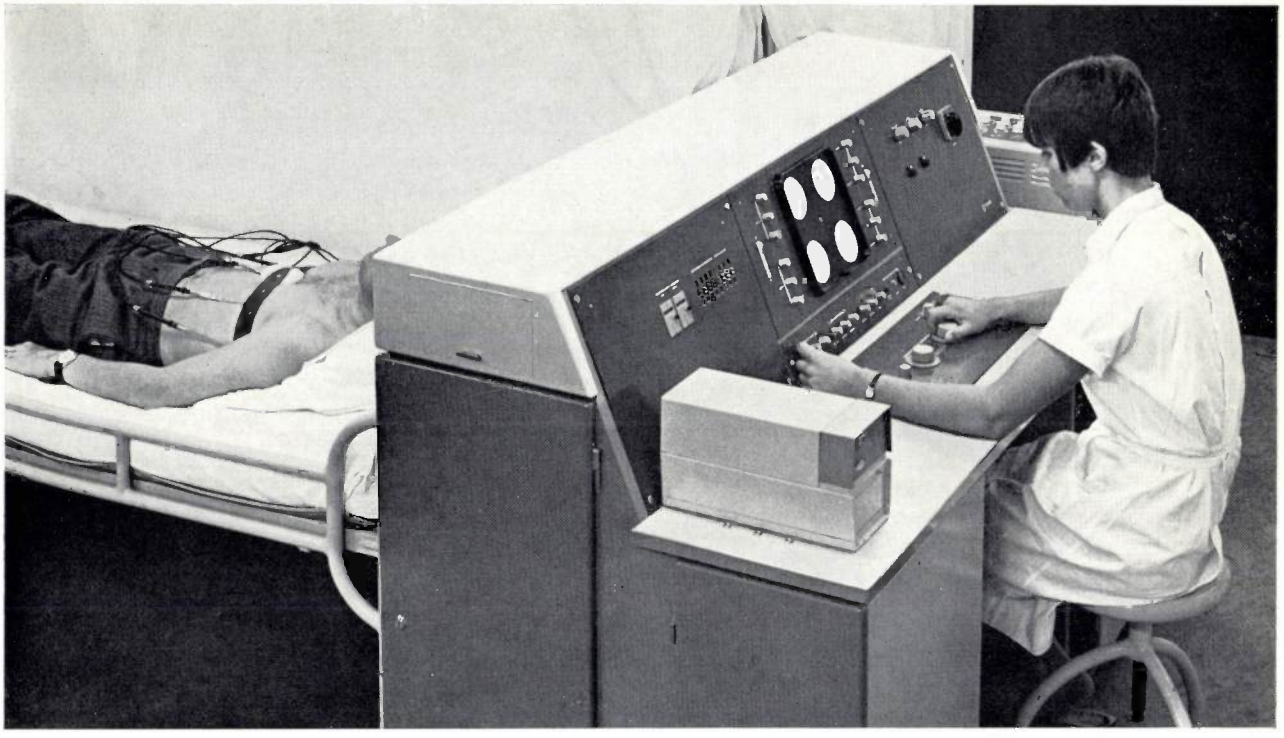


Fig. 2. The experimental vectorcardiograph, developed at Philips Research Laboratories and in use at the Philips Medical Service at Eindhoven.

he can switch over from one system to the other. This offers the facility of comparing the merits of the different lead systems.

A very useful feature is a circuit device for rotating the system of axes around the Y axis and the Z axis. A loop can thus be viewed from the most suitable angle and a good impression obtained of its spatial orientation.

In addition to the three cathode-ray tubes for displaying the projections of the vectorcardiogram there is a fourth tube, the "monitor". Any one of the three components X, Y or Z can be displayed on the monitor screen as a function of time, the curve thus obtained corresponding to a conventional electrocardiogram (derived however from a composite rather than a bipolar lead). The QRS peak of this electrocardiogram (see fig. 1*a*) is used in the instrument for supplying control pulses to various ancillary circuits.

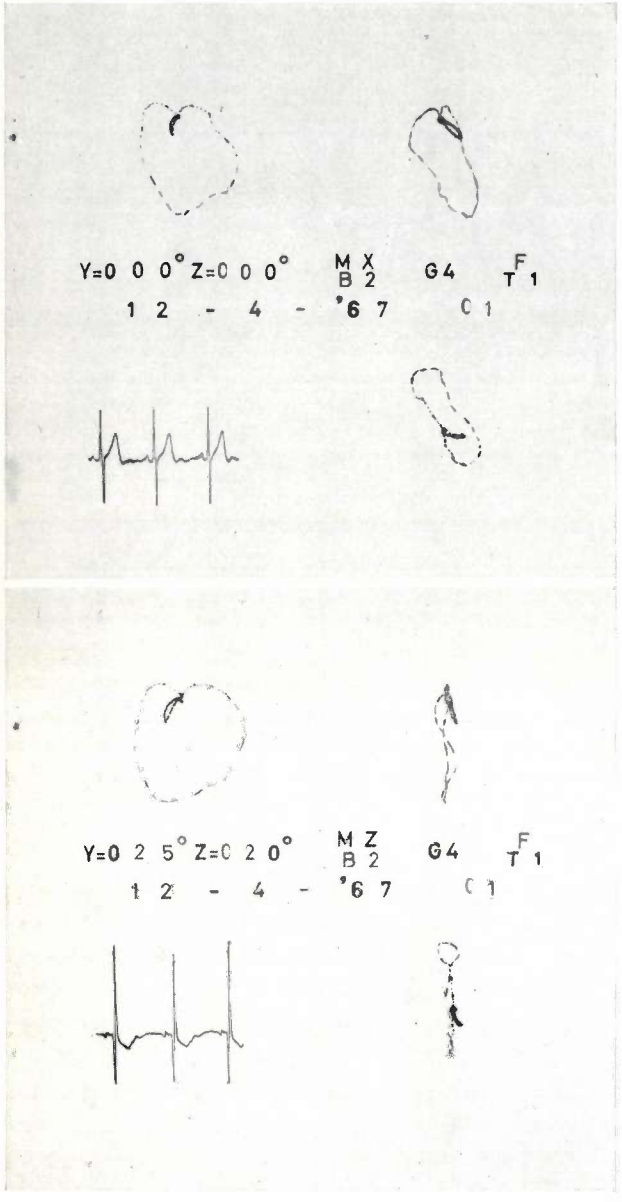
The amplification of the signals can be varied in eight steps of $\sqrt{2}$, and the bandwidth in six steps from 50 Hz to 5000 Hz. There is a built-in oscillator for testing the complete amplifier and mixing circuit. The use of special circuits [2] ensures that when the cable connected to the patient has a length of 6 m the combined level of the hum and noise signals is no greater than about 1 % of the signal from a normal QRS loop of a vectorcardiogram. All the circuits between the patient and the cathode-ray tubes are direct-coupled; this has the great advantage that there are no long blocking times if a circuit should be overloaded.

In the design of the vectorcardiograph considerable attention was given to easily operated devices for displaying either a complete cycle or, for closer study, a selected part of a cycle. The selection of a particular part of the cycle for photographing is done by connecting the signal through only during a particular time interval and suppressing it for the rest of the cycle. The selected portion is indicated on the monitor by extra brightness. Special attention was also paid to arrangements for automatically photographing the screen displays. This is done with the aid of four other cathode-ray tubes (the phototubes), which are connected in parallel with the four display tubes shown in fig. 2; the displays on the phototubes are simultaneously photographed on 70 mm film.

During the writing of the vectorcardiogram the electron beam is periodically suppressed for a few milliseconds to obtain a time marking. Since in addition the beam current and thus the brightness of the picture

[1] G. C. E. Burger and G. Klein, Philips tech. Rev. **21**, 24-37, 1959/60.
For a recent survey, see S. Boutkan, Vectorcardiography, physical bases and clinical practice, Centrex Publishing Co., Eindhoven 1965.

[2] Circuits like those used here are described in: G. Klein and J. J. Zaalberg van Zelst, Precision electronics, Centrex Publishing Co., Eindhoven 1966.



a

b

Fig. 3. Photographs of the displays on the screens of the four cathode-ray tubes. The coding shows that in *a*) the co-ordinate system has not been rotated ($Y = 000°$, $Z = 000°$) and that the X component is displayed on the monitor (MX). In *b*) the co-ordinate system has been rotated 25° about the Y axis and 20° about the Z axis. Here the Z component is displayed on the monitor. Since the whole cycle is displayed, the area around the origin is very dark, obscuring the smallest loop. This drawback can be overcome by displaying only part of the cycle.

is modulated so that a) each point where the trace is interrupted has the form of an arrow, and b) one in every four line sections is given extra brightness, it is possible to determine the direction in which the heart vector curve is described, the total time it takes to describe it, and also the corresponding points in the three projections. Modulation of the electron beam is also used to ensure that variations in the writing speed of the beam as it traces out the picture do not cause excessive brightness differences in the display.

When the camera control button is operated the camera shutter opens. The screens of the phototubes for the three projections, which are normally dark, then display the part selected from the cycle, which is immediately photographed, together with the picture on the photomonitor and data on the rotation of the co-ordinate system, the lead system employed, amplification, date, the patient's serial number, and so on. Typical photographic recordings are to be seen in *fig. 3*. After 4 seconds the shutter again closes and the film is automatically transported. The system contains a variety of precautionary devices, for preventing double exposures, etc.

The instrument is suitable both for fundamental research and for routine examinations. To assist routine work it is fitted with two inputs, which permit changeover from one patient to another without loss of time.

F. Douma
J. Guldemond

F. Douma and J. Guldemond are with Philips Research Laboratories, Eindhoven.

Continuous hot pressing

G. J. Oudemans

Hot pressing is a recently developed process which can be used for making ceramic materials with special properties. The process is costly and has therefore been used so far only on a modest scale, for example in making nose cones for space vehicles. Until now it has been confined to piece production, which is normally used in ceramic manufacture. A continuous hot-pressing process has now been developed at Philips Research Laboratories, Eindhoven; this process is very suitable for automatic operation and offers prospects of more economic production.

Introduction: hot pressing

In the course of the centuries the potter's craft has grown into a versatile technology which can be used for processing a great variety of raw materials into products possessing widely divergent properties. Today, ceramic products are very extensively used on account of their hardness, resistance to heat or to chemical attack; because of special electrical properties (use in insulators, capacitors, etc.), or for their magnetic properties (as in ferroxcube, ferroxdure and ferroxplana).

Although widely different basic raw materials are used, the manufacturing process is broadly the same. As a rule it consists of the following three stages:

1) The raw materials are ground and mixed (clay, of course, does not have to be ground).
2) The powder is then mixed with water or with organic binders and moulded or pressed to the required shape, or extruded with the required profile. The product thus formed is described as "green ware". It contains a very large number of pores (filled with air, water or binder) which together constitute a third to a half of the volume.
3) The green ware is dried and fired at high temperature, which gives it strong cohesion as a result of sintering processes. This considerably reduces the porosity and at the same time the product shrinks.

Some years ago a ceramic manufacturing process was developed in which the traditionally separate operations of forming and firing are carried out in one step. In this process the ceramic powder is pressed in a die which is heated to a temperature high enough for sintering of the compressed powder in the die. The pressure applied is 100 bar and higher, the temperature is about 1000 °C or higher. This process is known as "hot pressing" [1].

The process has several important advantages:

1) The sintering temperature can be lowered and the sintering time shortened. This is important, for example, where sintering under normal conditions would cause volatilization or decomposition of the material.
2) The microstructure, i.e. the distribution of crystallites and pores, which has a considerable effect on the properties of the ceramic product, can be controlled by varying the pressure, the temperature and the length of time in the die. Given an appropriate choice of these process parameters the porosity can be reduced in many cases almost to zero; in this respect the method is as a rule superior to the conventional one. It is also possible in this way to obtain a microstructure which has small crystallites with low porosity, which is virtually impossible with the conventional process.
3) Another advantage which is in principle to be expected is that of a better dimensional accuracy than that obtained with the conventional process, since sintering shrinkage takes place during the forming of the product and is thus accounted for at this stage.

The second point calls for some explanation. In the

Dr. G. J. Oudemans is with Philips Research Laboratories, Eindhoven.

[1] P. Murray, D. T. Livey and J. Williams, The hot pressing of ceramics, in: W. D. Kingery (ed.), Ceramic fabrication processes, Technology Press M.I.T., Cambridge, Mass., U.S.A., 1958, pp. 147-171; T. Vasilos and R. M. Spriggs, Proc. Brit. Ceramic Soc. 3, 195, 1965.

In modern powder metallurgy, where the processes are very similar to ceramic processes, the hot-pressing method has been in use for much longer. An important difference, however, is that much lower temperatures can often be used, so that various problems typical of the ceramic process do not arise.

conventional ceramic process, particularly when a low porosity is required, the necessary sintering temperature is usually so high and the sintering time so long that some of the grains of the ceramic powder are able to grow at the expense of the others. The result is that very large crystallites are formed (sometimes up to a few millimetres in diameter) which are frequently surrounded by smaller ones. A microstructure of this kind may often be undesirable, and is moreover not readily reproducible. With the hot-pressing process it is possible to obtain at a lower sintering temperature and a shorter sintering time a product whose crystallites are not much larger than the grains of the ceramic powder and are homogeneous in size (e.g. a few μm in diameter).

The practical importance of dimensional accuracy, low porosity and small crystallites will be illustrated at the end of this article with a few experimental results.

In spite of these attractive features, hot pressing is intrinsically more complicated and more expensive than conventional processes. Up to now it has therefore only been used where expense is no object, e.g. for making nose cones for space vehicles.

The chief difficulty in using hot pressing for materials which are of interest in the electronics industry was the short life of the die. An investigation at Philips Research Laboratories, Eindhoven, was therefore started, with the aim of making a die of sufficiently long life, to be used in a continuous process. For an important class of products a continuous process can be cheaper to operate than the normal piece production process.

Principles of continuous hot pressing

As a result of these investigations, a continuous process has been developed in which the ceramic material is hot-pressed in the form of a continuous cylindrical rod [2]. So far we have carried out tests only with the production of a circular cylindrical rod, but in principle a rod of any required profile can be made. We shall describe the principle with reference to *fig. 1*. *Fig. 2* shows the equipment used.

A small quantity of powder is poured on to the lower punch *B* while it is in the cylindrical opening of the die *A*. The powder is then compressed by the upper punch *C* (fig. 1*a*). The lower punch is moved continuously downwards at the rate of about 5 cm/h. After the powder has been compressed to the required volume, the upper punch is raised again, freeing the opening at the top of the die, and a new charge of powder is injected. The upper punch is then lowered once more into the die, and the whole cycle is repeated until the compacted rod has the desired length (fig. 1*b*). Under the conditions which we have used the cycle time can be set to any value between 15 s and 5 min.

The lower punch, except during the very first stage, serves only to support the hot-pressed material, which itself takes over the function of the lower punch, in the die.

The temperature distribution in the die should be such that only the bottom layer of each charge of powder is completely sintered in one cycle. The top layer should remain unsintered in order to ensure a continuous transition with the next charge of powder. An internal examination of the rods made by this process shows no trace at all of the layers in which the powder was fed into the die.

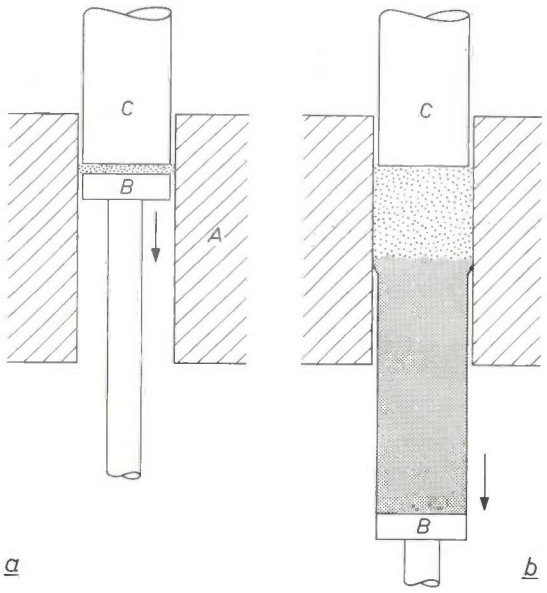


Fig. 1. Cross-section of the die for continuous hot pressing. The die *A* has a cylindrical opening in which the ceramic powder is compacted at a pressure of about 1000 bar between the lower punch *B* and the reciprocating upper punch *C*. Every time the upper punch reaches its uppermost position a new charge of powder is fed into the opening of the die. The lower punch moves downwards continuously at a constant speed, adjustable from a few cm/h to 60 cm/h. *a*) Starting position and *b*) later position of the lower punch.

The equipment

The first requirement to be met by the material of the die is a mechanical strength permitting the use of high pressures at temperatures above 1000 °C. Another important requirement is that the ceramic material should not sinter to the wall of the die. Finally — and this applies particularly to certain applications for the electronics industry — the material of the die should be proof against oxidation at the high temperature employed. In such applications the sintering will often have to take place in an oxidizing atmosphere, since many materials of importance in electronics, such as

[2] The idea was first proposed by G. S. Gruintjes of this laboratory, who together with W. C. P. M. Meerman helped to bring it into practical use.

ferrites and titanates, are susceptible to reduction at these temperatures. This means in fact that the only materials which can be used for the die are *oxides*. The most suitable material, after the other requirements have been taken into account, is pure and very densely sintered aluminium oxide (alumina).

A weak point of this material is its low tensile strength. Since the compression strength of Al_2O_3 is about ten times greater than its tensile strength, we avoided this difficulty by mounting the die under high compression as shown in the arrangement of *fig. 3*.

Six hydraulic cylinders E in a heavy metal ring D around the die A exert identical forces on the six sides of the die, which is ground to the shape of a regular hexagonal prism. The ball joints F on the plungers G take up the play which arises in transmitting the compressive forces via blocks H_1 and H_2. The blocks H_1 are made partly of Al_2O_3 and partly of asbestos, which

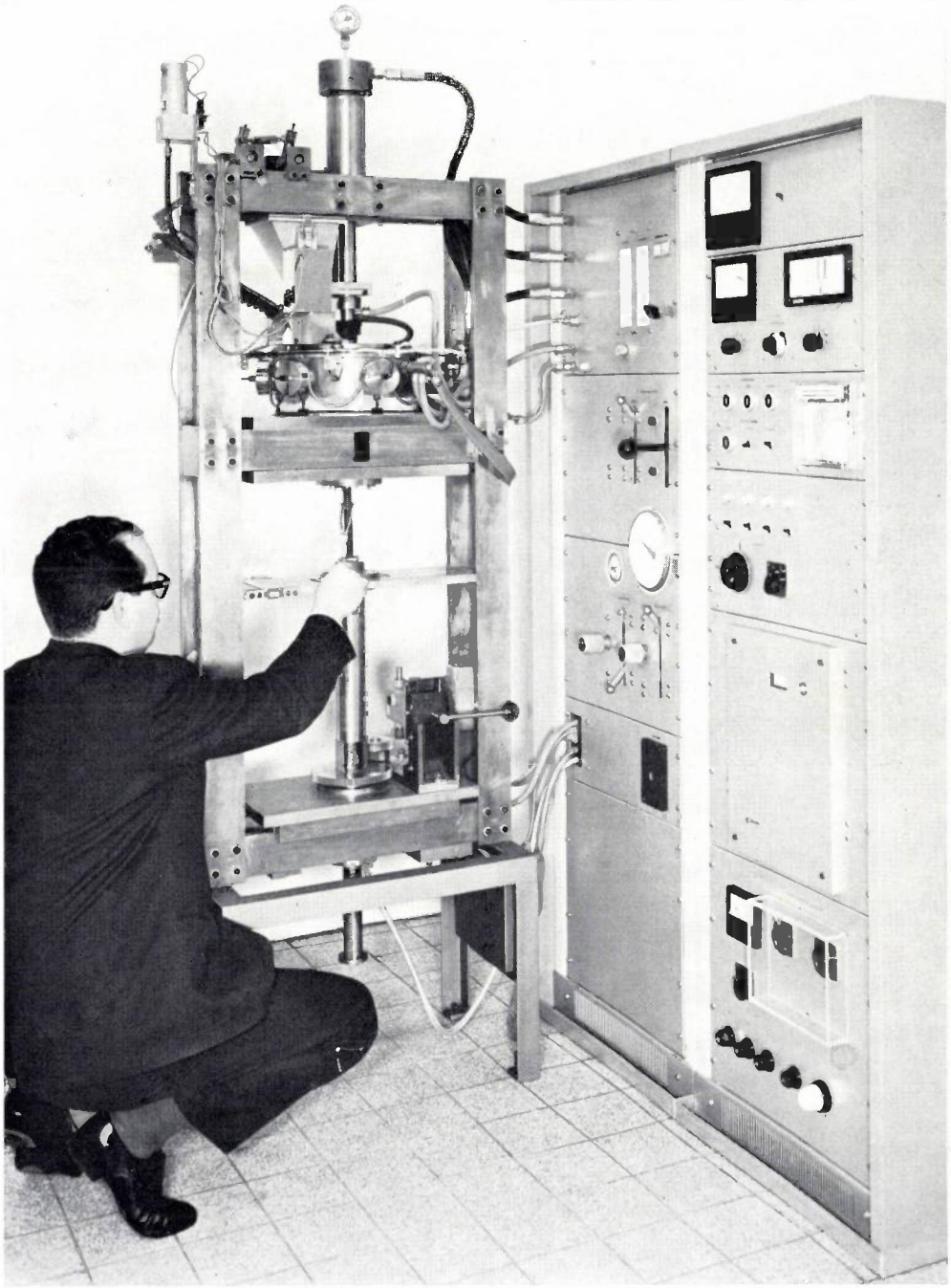
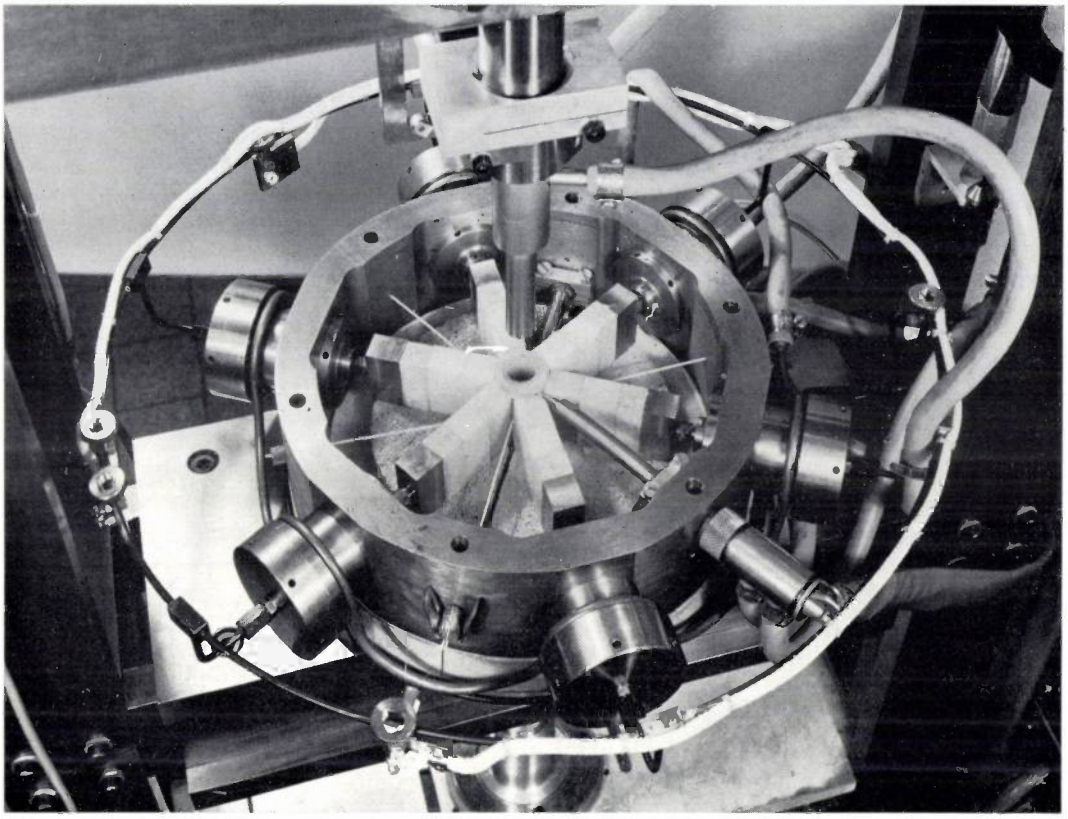


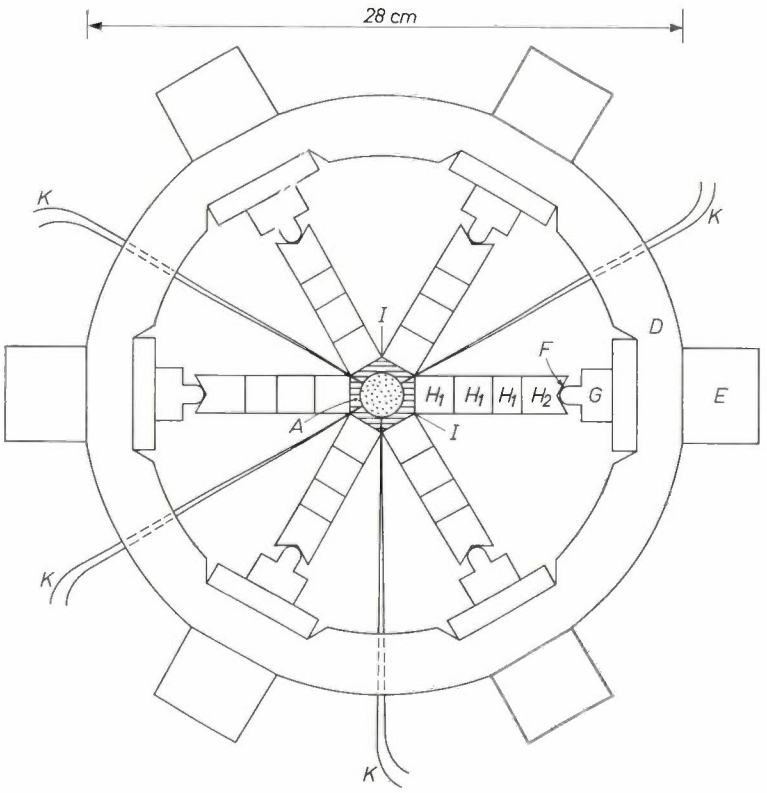
Fig. 2. The equipment at Philips Research Laboratories, Eindhoven, which is used for making ceramic rods by continuous hot pressing.

are both able to withstand the high temperature in the region of the die. The blocks H_2 are made of stainless nickel-chromium steel which, unlike Al_2O_3 or asbestos, is capable of withstanding the concentrated load exerted by the plungers.

The die, thus encapsulated, is heated to the required temperature by a resistance wire wound in a helical groove ground in the outer surface of the die wall. In our arrangement we use a platinum-rhodium wire, which dissipates 1 kW at a voltage of about 20 V.



a



b

Fig. 3. *a*) Plan view of the die, and *b*) cross-section perpendicular to the centre-line of the punch. Hydraulic pressure is exerted upon the die *A* by means of six cylinders *E* mounted in a robust steel frame *D*. The ball joints *F* on the plungers *G* take up the play in the transmission of the forces via the blocks H_1 and H_2. A helical groove ground in the outside wall of the die contains a heater wire with terminals *I*. The four thermocouples *K* are fitted in holes drilled in the die.

To monitor the axial temperature distribution, Pt/Pt-Rh thermocouples are placed in holes drilled at different heights in the side edges of the die, which extend to a depth of a few millimetres from the inside wall of the die. The thermocouples are pressed against the body of the die by a spring mounted outside the metal ring. In this way the temperature can be measured to an accuracy of about 5 °C.

The upper punch is made of steel and is water-cooled to prevent the top layer of the powder in the die from sintering. (As we noted above, this layer should not be allowed to sinter if a continuous transition with the next charge of powder is desired.) A drawback is that the reciprocating movement of the cooled punch subjects the material of the die to considerable fluctuations in the temperature. The Al_2O_3 has poor resistance to thermal shock and the effect of this is the appearance of irregular hair cracks. The usefulness of the die, however, is in no way affected, since the individual pieces formed by the cracks remain pressed together and undergo no relative displacements. We have had a die in operation at 1200 °C and 1000 bar nearly every working day since September 1966, and it is still giving excellent service.

The speed of the production process

A high rate of production is obviously desirable for economic reasons. However, since it is one of the parameters which determine the properties of the product, the speed of the process has to be held within certain limits. In the experiments discussed in the next section the production rate varied from 2.5 cm/h to 15 cm/h at a maximum sintering temperature of 1200 °C.

With some materials the rate of production can be increased, while maintaining the desired properties, if the sintering temperature is raised. Although the mechanical properties of Al_2O_3 are seriously impaired above 1200 °C, a few preliminary experiments have shown that heating to at least 1400 °C is possible. In some cases we have succeeded in reaching 99.9% of the theoretical density at a production rate of 60 cm/h. Since the sintering zone is about 1 cm long, this means that the material is sintered in one minute.

Process supervision

Useful information about the process is obtained by measuring the force exerted on the lower punch. This force is equal to the force on the top punch minus the frictional force exerted by the sintering powder on the wall of the die. It is found that the movement of the top punch during compaction of the powder is to some extent discontinuous. A few times in every cycle the friction rises slowly to about 90% of the force exerted on the powder by the upper punch, and then it drops abruptly to about 60% of this value. The number of times this happens depends partly on the rate of travel of the lower punch. The recording of these changes tells the man at the press how the process is running.

The movements of upper and lower punches and the feeding of the die can easily be made fully automatic. *Fig. 4* shows the filling mechanism of the completely automatic equipment (shown in fig. 2). Once the process has been started, it requires hardly any supervision. This means that one man can be left to supervise the production of numerous automatic machines, each of which could be equipped with a multiple die. We have thus found an economic method of producing ceramic materials possessing properties which cannot be obtained by the conventional process.

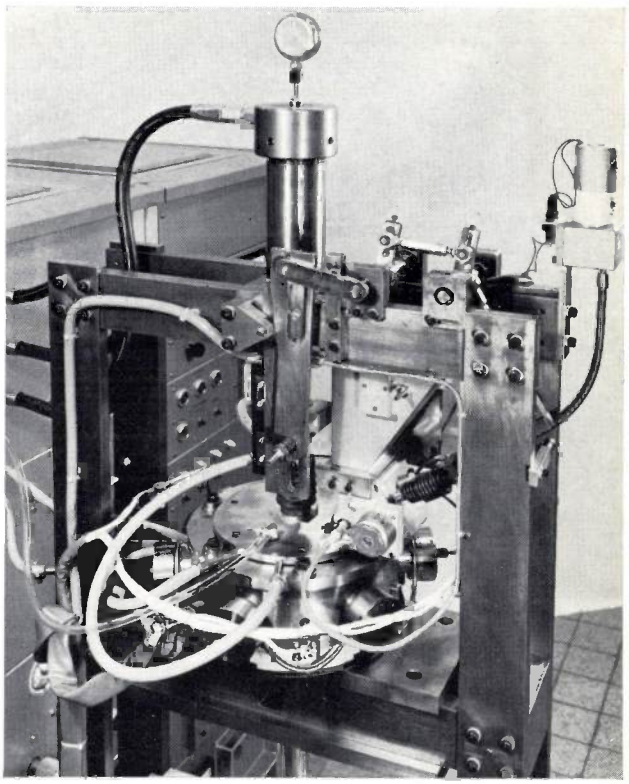


Fig. 4. The feed mechanism.

Some results

We have mentioned in this article three fundamental advantages of the hot-pressing technique: it gives products with accurate dimensions, low porosity and small grain size. We shall now illustrate the practical importance of each of these properties with a few examples.

Dimensional accuracy

In certain nuclear reactors ceramic rods of uranium oxide are used, which have to conform very strictly to specified dimensions. These rods are therefore built up

from small cylinders, each one of which has to be ground to size. A rod made by our continuous hot-pressing method would be a much better alternative. In a preliminary experiment we have hot-pressed a uranium-oxide rod of the required dimensional accuracy to a bulk density of 98%.

The exceptional dimensional accuracy achieved with our method was demonstrated by our experiments on the production of nickel-zinc ferrite [3]. The diameter of rods produced by this method over a period of six months was found to remain within a tolerance of 5 μm.

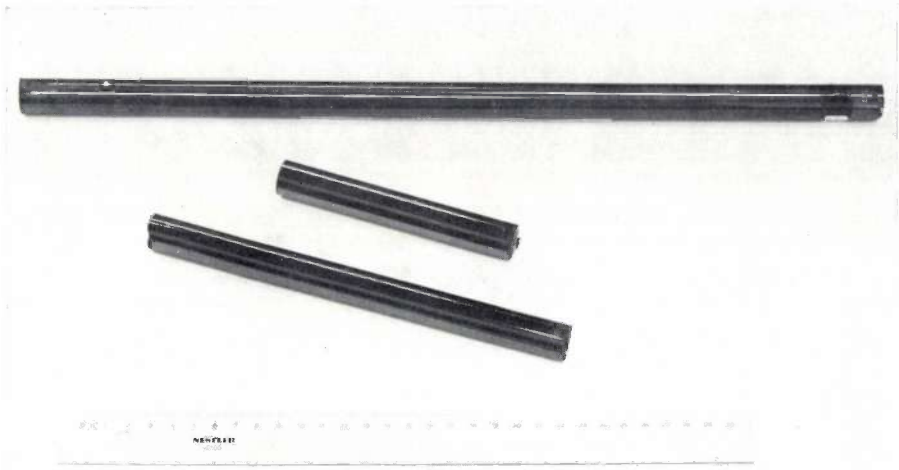


Fig. 5. Some ceramic rods made by continuous hot pressing.

Another feature is that the rods all had a shiny, smooth surface (*fig. 5*). Tiny grooves in the surface, running parallel with the axis of the rod, were visible only under a microscope.

Low porosity

We have also experimented with a group of lithium-zinc ferrites $(Li_{0.5}Fe_{0.5})_{1-x}Zn_xFe_2O_4$, where x can be between 0.2 and 0.7. These ferrites were of interest in our laboratories because of possible applications in microwave ferrite devices, such as circulators and waveguide phase shifters. For this kind of application the porosity of the ferrite should be extremely low, no more than a few tenths of one per cent, since the demagnetizing fields of the pores disturb the required homogeneity of the magnetic field in the ferrite. It had already proved possible to make sufficiently dense lithium-zinc ferrites by conventional means, but the temperature required for sintering was higher than 900 °C. Above this temperature limit part of the lithium is lost in the gas phase in the form of volatile compounds. Owing to the disappearance of the lithium ions some of the ferric ions in the ferrite are converted into ferrous ions. The existence of both types of iron ions side by side in the crystal lattice is associated with an unacceptable fall in resistivity. The continuous hot-pressing method enables lithium-zinc ferrites to be sintered to a very low porosity at 900 °C. The ferrites which we have made had a porosity between 0.2 and 0.3% and a resistivity between 10^8 and 10^9 Ωcm. They therefore meet the primary requirements for circulator materials.

Small grain size with low porosity

Apart from the low porosity obtained by this technique, small grain size is important in microwave applications of lithium-zinc ferrite (and certain other ferrites).

In these magnetic materials the energy loss increases sharply when the amplitude of the microwave field exceeds a critical value h_{crit}. This critical value therefore sets a limit to the power which can be handled by the ferrite. *Table I* gives the h_{crit} values for a lithium-zinc ferrite and for two other ferrites, made both by the conventional method and by hot-pressing. It can be seen from the table that, particularly for the magnesium-manganese ferrite, the value of h_{crit} is considerably increased by hot pressing.

The energy loss of microwaves whose magnetic field has an amplitude greater than the critical value h_{crit} is related to the parametric excitation of spin waves [4]. High attenuation of the spin waves entails a high h_{crit}, and the material can be used for higher powers. An important contribution to this attenuation comes from the grain boundaries. Since the surface area of the grain boundary increases sharply with decreasing grain size, fine-grained materials are more suitable for high-power applications.

[3] All ferrites mentioned in this article have been prepared in close co-operation with Ir. J. G. M. de Lau of this laboratory.

[4] See M. Sparks, Ferromagnetic-relaxation theory, McGraw-Hill, New York 1964.

Table I. Critical amplitude h_{crit} of the microwave magnetic field for ferrites made by continuous hot pressing and by conventional methods. The porosity and average grain size are also shown.

Ferrite	Method of preparation	Porosity %	Average grain size µm	h_{crit} oersted
$Li_{0.4}Zn_{0.2}Fe_{2.4}O_4$	hot-pressed	0.7	0.5	12
$Ni_{0.36}Zn_{0.64}Fe_2O_4$	conventional	0.3	35	5
$Ni_{0.36}Zn_{0.64}Fe_2O_4$	hot-pressed	0.2	0.5	16
$Mg_{0.78}Cu_{0.20}Mn_{0.05}Fe_{1.97}O_4$ [*]	conventional	2.7	5.0	12
$Mg_{0.98}Mn_{0.05}Fe_{1.97}O_4$	hot-pressed	0.9	0.5	>120

[*] Cu has to be added to this ferrite to obtain an acceptable porosity when the conventional method is used.

For magnetic cores or material for ferrite aerials for frequencies higher than 100 kHz, nickel-zinc ferrites are used. These should preferably be densely sintered and are doped with small amounts of cobalt. The function of the cobalt ions in the crystal lattice is to suppress the movement of the magnetic domain walls which is associated with the magnetic loss. Another way to suppress the movement of the domain walls is to reduce the size of the grains in the nickel-zinc ferrite to a diameter of 1 µm or less. The few domain walls then remaining in a grain are thus kept virtually stationary by the grain boundaries. Although pinning the domain walls in this way reduces the magnetic permeability, the overall magnetic quality is increased.

The grains of nickel-zinc ferrites sintered in the conventional way are generally too large for this purpose; this is certainly the case if these ferrites undergo prolonged sintering at high temperature in order to obtain the low porosity.

Fig. 6a and *b* shows the microstructure of two nickel-zinc ferrites obtained from the same powder; the first was prepared by conventional methods, and the second

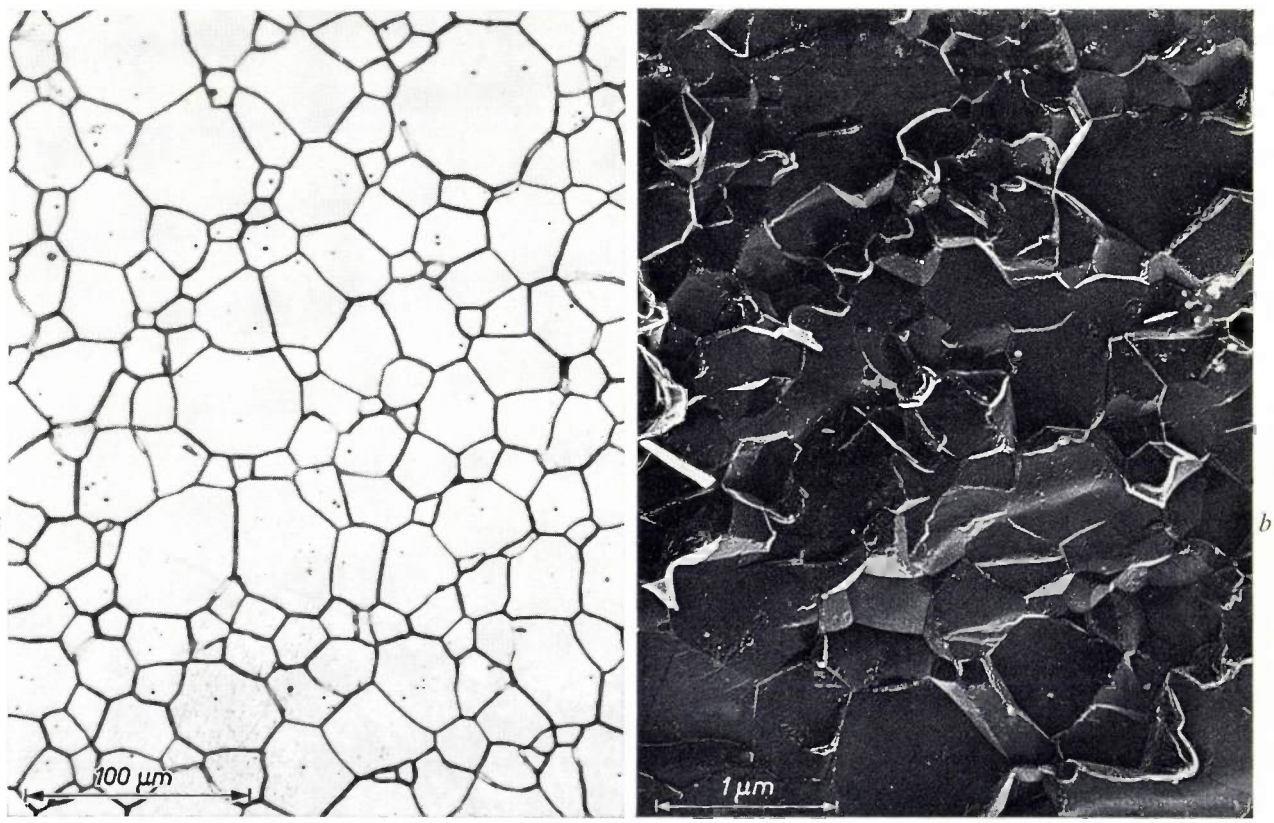


Fig. 6. Specimens of $Ni_{0.36}Zn_{0.64}Fe_2O_4$ of the same porosity (0.3 %) and the same resistivity (10^7 Ωcm), prepared in various ways: *a*) by the conventional method (20 h at 1275 °C), *b*) by continuous hot pressing (30 min at 1025 °C and 1000 bar); *a*) is a polished and thermally etched cross-section photographed with the aid of a microscope; *b*) is a fracture surface as shown by the EM 200 electron microscope.

by our continuous hot pressing process. Both samples had the same resistivity (10^7 Ωcm) and porosity (0.3 %). The conventionally prepared ferrite was sintered for 20 hours at 1275 °C, and the average diameter of the grains is 35 μm. The hot-pressed ferrite was sintered at 1025 °C for about 30 min at a pressure of 1000 bar. The diameter of the grains is between 0.5 and 1 μm.

Fig. 7 shows some results of measurements of the magnetic behaviour of both ferrites [5]. The quantity μ'', a measure of the loss, is plotted in fig. 7*a* as a function of frequency. In the conventionally prepared ferrite this quantity begins to increase at a frequency of about 0.1 MHz, whereas in the hot-pressed ferrite it begins to increase only at frequencies in the region of 1 MHz, where the other ferrite has its maximum loss. It can be seen in fig. 7*b* that the magnetic permeability μ' of the hot-pressed ferrite in the same frequency region is lower than that of the other one. The ratio $\mu''/(\mu')^2$ is used in these applications as a figure of merit: the smaller the ratio the more useful the material. In fig. 7*c* this ratio is plotted against frequency for the materials investigated. The lower values of $\mu''/(\mu')^2$ of the hot-pressed ferrite at the same frequency clearly demonstrate the superiority of this ferrite.

The value of $\mu''/(\mu')^2$ in cobalt-containing ferrites can also be substantially reduced by making the ferrites by the hot-pressing process (*fig. 8*) [5], and the useful frequency limit is almost doubled.

Finally, we should like to mention some experiments with lead zirconate-titanate, a ceramic material possessing piezo-electric properties [6]. The object of the experiments was to make the material suitable for use as a transducer for delay lines at a frequency higher than 10 MHz. In this application the material is used in the form of very thin wafers, 50 to 100 μm thick, which are cut from a larger piece of material and then ground. Metal electrodes are deposited on both faces of the wafer. In practice these vacuum-evaporated or chemically-deposited electrodes often become short-circuited. This is caused by the presence in the material of a pore

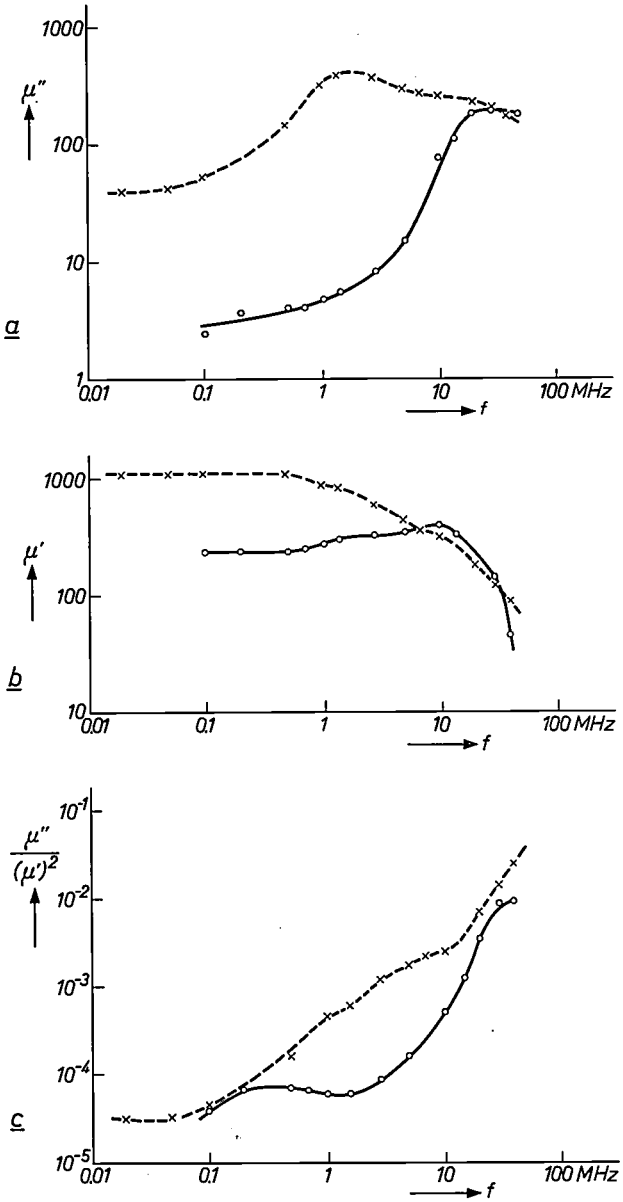


Fig. 7. *a*) The imaginary part μ'' of the permeability, which is a measure of the magnetic loss, *b*) the real part μ' of the permeability, and *c*) the ratio $\mu''/(\mu')^2$, which is a figure of merit for the magnetic quality, all plotted as a function of frequency f for the NiZn ferrite of fig. 6. The dashed curves relate to the conventionally prepared ceramic and the solid curve to the hot-pressed product [5].

[5] These investigations were carried out in this laboratory by Ir. J. G. M. de Lau. The complete results will be reported in due course in Proc. Brit. Ceramic Soc. and elsewhere.

[6] These experiments were performed by Drs. R. R. P. Varekamp, of Philips Electronic Components and Materials Division (Elcoma), Eindhoven.

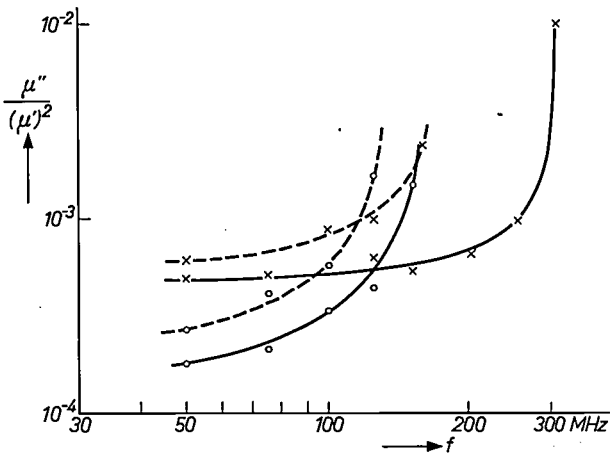


Fig. 8. The figure of merit $\mu''/(\mu')^2$ as a function of frequency f for conventionally prepared (dashed curve) and continuously hot-pressed (solid curve) cobalt-containing nickel ferrite and nickel-zinc ferrite [5]. The crosses and circles relate to measurements on $Ni_{1.02}Co_{0.02}Fe_{2.4+\delta}$ and $(Ni_{0.8}Zn_{0.2})_{1.01}Co_{0.03}Fe_2O_{4+\delta}$ respectively (where δ is very small).

or a hole left when a large grain was dislodged during the grinding process.

On this account, if a completely solid material is not attainable, it is at least desirable that the material should have extremely small pores, e.g. an order of magnitude smaller than the thickness of the wafer. The average grain size should also be such that there are at least three to four grains across the thickness of the wafer.

With conventional processes a temperature of nearly 1300 °C is needed to sinter the lead zirconate-titanate to the required low porosity. At this temperature marked grain growth occurs.

By using the continuous hot-pressing process we have succeeded in making a lead zirconate-titanate ceramic of very low porosity and with the required grain size for use in the form of these extremely thin wafers. *Table II* gives some data on the grain size of the material obtained at the various temperatures between 900 °C and 1150 °C.

Table II. The bulk density and average grain size of continuously hot-pressed lead zirconate-titanate ceramic ($PbZr_{0.54}Ti_{0.46}O_3$) for various temperatures and production rates at a pressure of 400 bar. The diameter of the rods was 17 mm, except for the experiment mentioned in the last row, where the diameter was 20 mm.

Temperature °C	Production rate cm/h	Density g/cm³	Average grain size µm
900	3.5	7.95	1-2
900-1050	3.5	7.95	3.0
1050	5.5	7.85	3.6
1100	3.5	7.82	4.4
1125	3.5	7.89	5.3
1150	5.5	7.72	5.3
1150	3.5	7.94	12.5

Summary. Some years ago the ceramics industry became interested in a process known as hot pressing. In this process the powder material is formed and sintered simultaneously under high pressure. The method has the important advantage of promoting good control of the microstructure. However, it did have disadvantages of an economic nature. These have now been overcome by using a continuous process, developed at the Philips Research Laboratories at Eindhoven. This process can be used to produce rods which in principle can have any desired profile. A die made of pure, densely sintered alumina is used. A description is given of the production equipment which has now been in operation for 16 months producing various materials at 1200 °C and 1000 bar without any significant wear of the die. A production rate of up to 60 cm/h is possible. The equipment is completely automatic. The properties and possible applications of various materials made by this technique (uranium dioxide, ferrites and lead zirconate-titanate) are discussed.

Photoemission of semiconductors

J. van Laar and J. J. Scheer

The phenomenon of photoemission has played an important role in the history of fundamental physics. Einstein took the empirical laws underlying photoemission as his starting point for postulating, as an extension of Planck's quantum hypothesis, that light consists of photons. Photoemission also has a considerable practical significance; photocathodes are used in many radiation detectors and, in particular, in most image intensifiers. Present photocathodes which are reasonably sensitive to visible light, were nearly all evolved more or less empirically. The article below shows that it is now possible to arrive at a good photocathode for visible light with the aid of more sophisticated methods.

Introduction

Consider a circuit (*fig. 1*) consisting of a vacuum tube with two electrodes, cathode and anode, connected to a battery and a sensitive ammeter. The cathode is cold; there is no thermionic emission and therefore no current flows in the circuit. If the cathode is now illuminated, the result — under favourable conditions — is a measurable current: the light releases electrons

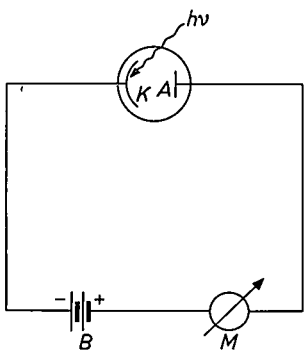


Fig. 1. Principle of photoemission. *K* cathode, *A* anode, *B* battery, *M* microammeter. A measurable current flows in the circuit only when the cathode is exposed to light radiation.

from the cathode. This phenomenon, called *photoelectric emission* or the *external photoelectric effect*, was first discovered in metals by Elster and Geitel in 1850.

This effect was found to obey simple laws which were not compatible with the classical wave theory of light. These laws are easily understood, however, if it is assumed that monochromatic light of frequency ν consists of a current of light quanta, each with an energy $E = h\nu$ [1]. An electron *inside* the cathode has a lower energy than in the vacuum outside the cathode. Upon absorbing a light quantum [2], an electron is excited to a state of higher energy. If the energy $h\nu$ of the absorbed light quantum is sufficiently high the excited electron is able to surmount the potential barrier at the surface.

This picture immediately leads to the rules found by experiment. 1) No photoemission occurs for photon energies below a certain value E_d, called the photoelectric work function or long wavelength threshold. 2) The number of electrons emitted per second is proportional to the number of photons incident per second on the cathode; in other words, the photoelectric saturation current is proportional to the intensity of the incident beam of light. 3) The maximum E_{max} of the energy of the emitted electrons is linearly related to the photon energy: $E_{max} = h\nu - E_d$.

It is important in what follows that the emitted electrons originate from a thin layer at the surface of the cathode. Electrons excited at greater depths lose too much energy on their way to the surface, by interaction with other electrons and with the atomic lattice, to be able to reach the vacuum. We assume that of the electrons excited at a distance x from the surface, only a fraction proportional to $\exp(-x/d)$ are emitted, where d is the "escape depth". Roughly speaking, only the photoelectrons from a surface layer of thickness d can escape.

In this article we are concerned with photoemission of *semiconductors*, and in particular with the research on this subject which has been done in recent years at the Philips laboratories in Eindhoven. There are two aspects of this research: a) The photoelectric effect can be used as a tool for studying the solid state, in our case semiconductor materials. b) As photoemission finds wide application in light detectors, an attempt can be made to design a good light detector using the knowledge obtained from research under (a). Before entering

Drs. J. van Laar and Dr. J. J. Scheer are with Philips Research Laboratories, Eindhoven.

on a detailed exposition of the subject, we should like to say a little more about these two aspects of the research, and also quote a few of the important results.

a) In research on semiconductors, photoemission provides information on the energy levels which have a significant electron population. The photoelectric work function gives directly the distance from the upper of these levels to the vacuum level; and because of the small escape depth, information is obtained on the situation at the surface. These are the main points of interest in our research. Other investigators [3] have studied the relative density of states in the occupied levels, i.e. the energy band structure, by analysing the wavelength-dependence of the photoemission. Additional information can be obtained along these lines by measuring the energy distribution of the photoelectrons. One might also try to learn something about the escape mechanism of the excited electrons, i.e. about the transfer of energy to the lattice and possible reflections of electrons at the surface. This is a subject which has so far received relatively little attention.

In addition to the above-mentioned threshold energy for photoemission, E_d, the *thermionic work function* φ is a quantity which is directly accessible by experiment, for example by contact potential measurements; φ is the difference between the vacuum level and the Fermi level in the semiconductor.

According to the band theory of semiconductors a zone of forbidden energies exists for the electrons in the material. In general, however, electronic states with energies in this forbidden zone will occur at the surface (because of impurities, lattice irregularities or other reasons). These surface states have a considerable influence on the properties of the semiconductor, in particular on φ and E_d as functions of the doping content of the semiconductor. As will be shown, the band model leads to the following results for two extreme cases:

1) If there are many surface states, all at roughly the same energy, then φ is independent of the doping content and E_d does depend on it (*fig. 2a*).

2) If there are no surface states, then φ is dependent upon the doping content and E_d independent of it (fig. 2*b*).

In the present article we shall discuss the results of our work on *silicon* and *gallium arsenide*. Our measurements show that Si is a typical representative of the first case. This situation is so common that the relevant model has been assumed to possess general validity. The results of our measurements make it clear, however, that GaAs definitely does not conform to this model, but on the contrary shows fairly good agreement with the second case.

b) If we now attempt to make a good photocathode from a conventional semiconductor, and in particular, a photocathode which is sensitive to visible light, we require in the first place that the threshold wavelength should lie in the infra-red. In other words, the photoelectric threshold energy must be sufficiently low. Conventional semiconductors do not meet this requirement. There is, however, an empirical method by which the photoelectric threshold energy can be shifted to a lower energy — *the adsorption of metal atoms on the surface*. The most effective metal appears to be caesium, which, given the right conditions, makes it possible to reduce the threshold energy to 1.4 eV (which is the ionization energy of a caesium atom in the adsorbed state; 1.4 eV corresponds to 0.90 μm wavelength). We shall discuss this method, again on the basis of the energy band model.

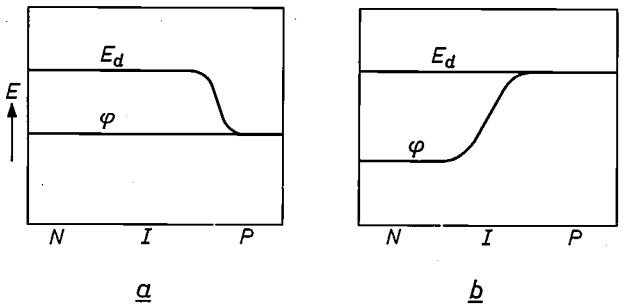


Fig. 2. The threshold energy E_d for photoemission and the thermionic work function φ as a function of the doping content of a semiconductor, *a*) in the presence of many surface states at an energy in the forbidden zone, *b*) in the absence of surface states. From left to right the doping goes from strong *N* via weak *N* and weak *P* to strong *P*; in the centre (*I*) the semiconductor is intrinsic.

Another requirement of a good photocathode is a high quantum yield. To achieve this, a substantial fraction of the photons has to be absorbed in the escape layer. Heavily-doped GaAs, coated with caesium, looks particularly promising in this repect. In this system the energy of an electron *in vacuo* lies roughly at the same height as the bottom of the conduction band. This means that an electron, once it has been excited to the conduction band, needs hardly any kinetic energy to escape from the semiconductor. The escape depth is then of the order of the diffusion-recombination length, that is to say many times greater than the escape depths normally encountered. These considera-

[1] A. Einstein, Ann. Physik **17**, 132, 1905.

[2] It may be assumed in general that an electron absorbs at the most one light quantum. Excitation processes in which an electron absorbs two quanta are also known, but they call for very special conditions. See H. Sonnenberg, H. Heffner and W. Spicer, Appl. Phys. Letters **5**, 95, 1964. The threshold energy for photoemission is then lower by a factor of 2 and the relation between the numbers of incident photons and emitted electrons is parabolic.

[3] D. Brust, Phys. Rev. **139**, A 489, 1965; M. L. Cohen and J. C. Phillips, Phys. Rev. **139**, A 912, 1965; N. B. Kindig and W. E. Spicer, Phys. Rev. **138**, A 561, 1965.

tions prompted us to make a photocathode of caesium-coated GaAs. We found that it gave the quite remarkable yield of about 0.30 electrons per incident photon on the blue side of the visible spectrum.

The first photoemission research on semiconductors, reported twenty years ago, was done with films deposited by vacuum evaporation [4]. In order to obtain well-defined materials and surfaces, we confined our experiments to vacuum-cleaved single crystals. We shall not go further into experimental detail here.

thermionic work function φ is equal to the threshold energy E_d. There are other methods of determining φ, for example contact potential measurements, and the results can be compared.

Let us now consider an analogous simple model of a semiconductor, shown in fig. 3*b*. Here, the Fermi level lies in the forbidden zone. The energy band below it (the valence band, with upper edge E_V) is completely occupied by electrons, and the band above it (the conduction band, with lower edge E_C) is empty. In addi-

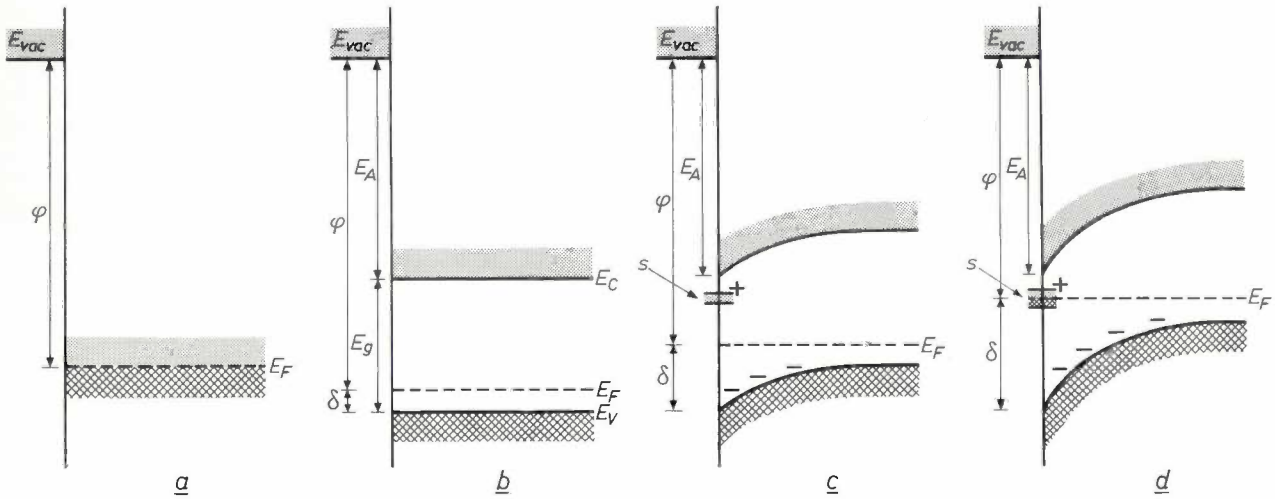


Fig. 3. Energy bands at the surface of a material. E_{vac} is the potential energy of an electron *in vacuo*, E_F the Fermi level in the material; $\varphi = E_{vac} - E_F$ is the thermionic work function.
a) Metal. The threshold for photoemission E_d (long wavelength threshold) is here equal to φ.
b) Semiconductor without donors or acceptors at the surface. E_C bottom of conduction band, E_V top of valence band, $E_g = E_C - E_V$ energy gap. $E_A = E_{vac} - E_C$ is the electron affinity. $\delta = E_F - E_V$ gives the height of the Fermi level with respect to the bands. The threshold for photoemission is $E_d = \varphi + \delta = E_A + E_g$.
c) Semiconductor with surface states *s* in a narrow energy range. The donors at the surface give up electrons to the acceptors in the material. Because of the resultant negative space charge the energy bands are bent.
d) Semiconductor with *stabilized Fermi level*. When an excess of surface donors exists, they give up their electrons only until the Fermi level coincides with the surface levels. In this case the Fermi level at the surface is independent of the doping content of the semiconductor.

Description of photoemission in terms of the band theory of solids

Using the band theory of solids [5] we shall now consider what may be expected in various cases for the photoemissive threshold energy E_d and the thermionic work function φ.

We consider first the situation in a metal (*fig. 3a*). The conduction band — at least at low temperature — is filled with electrons up to the Fermi level E_F. All the higher energy levels are practically empty. At the surface between metal and vacuum there is a potential barrier: the potential energy of an electron *in vacuo* (E_{vac}) lies above the Fermi level by an amount φ. Photoemission may be expected from the metal if for the incident light, $h\nu \geqslant \varphi$. In a metal, therefore, the

tion to the work function $\varphi = E_{vac} - E_F$ we introduce the *electron affinity* $E_A = E_{vac} - E_C$, the *energy gap* $E_g = E_C - E_V$ and the quantity $\delta = E_F - E_V$, which defines the position of the Fermi level in relation to the bands. In photoemission, the electrons originate from the valence band [6]; we may therefore expect for the threshold energy $E_d = E_A + E_g = \varphi + \delta$.

Let us now consider the effect on E_d and φ of variation of the bulk doping. To a first approximation bulk doping leaves the position of the energy bands unaffected, but it does affect the occupation of the levels and hence the position of the Fermi level; in this way the quantity δ can be varied between 0 and E_g. From fig. 3*b* it can then be seen at once that φ can be varied between E_A and $E_A + E_g$ by bulk doping, whereas

$E_d = E_A + E_g$ is independent of bulk doping. For any bulk doping the value of δ can be determined from the bulk properties of the semiconductor (conductivity, Hall coefficient). Thus, although not so directly as in the case of a metal, the thermionic work function $\varphi = E_d - \delta$ can again be derived from photoemission measurements.

In the model outlined above it was assumed that the surface causes no complications, in other words that the energy levels are not dependent upon location with respect to the surface. Let us consider an energy band diagram in which the horizontal co-ordinate represents the spatial co-ordinate perpendicular to the surface. The assumption which we have just mentioned implies that in such a diagram the energy bands are represented by straight horizontal lines. It is known, however, that at the surface of a semiconductor there is nearly always some *bending of the energy bands* owing to the presence at the surface of additional donors or acceptors with energy levels in the forbidden zone [7]. Suppose, for example, that there exists at the surface of a *P*-type semiconductor a large concentration of donor levels situated in relation to the bands in such a position that they would be above the Fermi level if the bands were straight. Now this is not a state of equilibrium; the donors at the surface will give up electrons to the acceptors in the bulk of the material, giving rise to a negative space charge region near the surface combined with a positive surface charge. As a result the potential at the surface drops and the bands at the surface are bent downwards (fig. 3*c*).

The potential distribution at the surface can be calculated by making the assumption that the acceptor levels in a layer of a given thickness are all filled, and that the acceptor levels outside that layer are all empty [8]. In this layer, the "Schottky layer", there is thus a constant and finite space charge density, and outside it a space charge density equal to zero. Solving Poisson's equation for this charge distribution (in a rationalized system) one finds that the potential V in the layer is parabolically related to the distance x from the surface:

$$V = Ne(x - x_0)^2/2\varepsilon, \quad \ldots\ldots \quad (1)$$

where N is the concentration of the acceptors, e the elementary charge, ε the dielectric constant and x_0 the thickness of the space charge layer. The total band bending ΔV (the potential difference between surface and bulk) is thus

$$\Delta V = Nex_0^2/2\varepsilon. \quad \ldots\ldots \quad (2)$$

In other words, for a given total band bending, the space charge layer is thicker, and the band bending therefore more gradual, the lower the bulk doping. This is of great importance in what follows.

If the number of surface donors per unit area of surface is very large, and if they all lie at roughly the same energy, then they will give up electrons to the acceptors in the bulk material only until the Fermi level at the surface just coincides with this donor level (fig. 3*d*). The Fermi level at the surface, and hence the thermionic work function φ, is therefore independent of the doping; in this case we speak of a *stabilized Fermi level*. The band bending is now determined by the position of this donor level at the surface and by the position of the bands in the bulk of the material in relation to the Fermi level; the latter position is in turn determined by the doping.

We now consider the effect of variation of bulk doping on the threshold for photoemission E_d in the case of a stabilized Fermi level. *Figure 4* shows the valence band at a stabilized Fermi level for *P*-type doping of various concentrations and for one kind of *N*-type doping (the conduction band is shown for only one case).

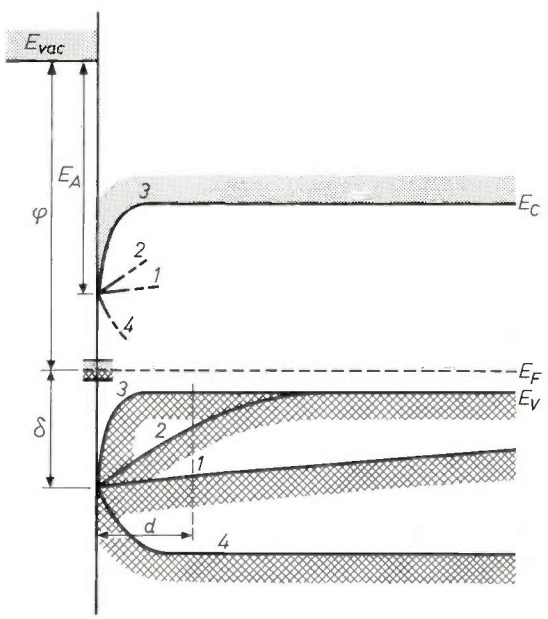


Fig. 4. Relation between volume doping and valence band emission when the Fermi level is stabilized. *d* escape depth. Other symbols as in fig. 3. *1* weak *P*-type doping, *2* stronger, *3* very strong *P*-type doping, *4* strong *N*-type doping. The conduction band is drawn only for case *3*. Case *1* corresponds to the "straight-band case" of fig. 3*b*, $E_d \approx \varphi + \delta$. In case *3* the photoelectrons mainly originate from the level $E_V \approx E_F$, so that $E_d \approx \varphi$. In case *4* the threshold is $E_d = \varphi + \delta$, but at this threshold energy the yield for photoemission is appreciably smaller than in the straight-band case (*1*).

[4] L. Apker, E. Taft and J. Dickey, Phys. Rev. **74**, 1462, 1948.
[5] For further details of this model, see L. Heijne, Philips tech. Rev. **25**, 120, 1963/64.
[6] This applies to *P*-type semiconductors and to fairly weakly doped *N*-type semiconductors. With very strong *N*-type doping some emission from the conduction band is to be expected. This has never been established with certainty, however.
[7] J. Bardeen, Phys. Rev. **71**, 717, 1947.
[8] See W. Schottky, Z. Physik **118**, 539, 1941/42.

As we have seen, the bending occurs in a thinner layer the greater the doping concentration. If the doping is very weak (fig. 4, curve *1*) the part of the band bending taking place inside the escape depth d is negligible. In that case the threshold E_d is equal to $E_g + E_A$, as it was with the straight bands in fig. 3*b*. Stronger *P*-type doping lowers the threshold energy E_d, the bands in the bulk of the material being raised with respect to the energy of an electron *in vacuo*. The resulting effect on emission becomes more noticeable as the band bending occurs to a greater extent within the escape depth (fig. 4, curves *2* and *3*). With very strong *P*-type doping, $E_d \approx \varphi$, since the Fermi level in the volume then coincides approximately with the upper edge of the valence band and the thickness of the space charge region is smaller than the escape depth.

Fig. 4, curve *4*, shows the valence band for strong *N*-type doping, with stabilized Fermi level. In this case the same threshold is to be expected as with straight bands, but the yield will be lower for photon energies close to the threshold.

Table I summarizes the behaviour of φ and E_d in both models when the doping is varied (see also fig. 2).

Table I

Model	$\varphi =$	$E_d =$
Straight bands	E_A for strong *N*-type doping; $E_A + E_g$ for strong *P*-type doping	$E_A + E_g$, independent of doping
Stabilized Fermi level	$E_A + E_g - \delta$, independent of doping	$E_A + E_g$ for *N*-type to weak *P*-type doping; $E_A + E_g - \delta$, thus $= \varphi$, for strong *P*-type doping

In reality, of course, intermediate cases can occur where δ at the surface is neither equal to the value in the bulk material, as it is in the case of straight bands, nor independent of doping, as it is with a stabilized Fermi level.

When analysing the results of the measurements one should bear in mind that the surface levels themselves can emit photoelectrons, which may mask in particular the weak beginning of the valence band emission. For this reason *N*-type material is generally less suitable for research on valence band emission, since the emission is exceptionally low near the threshold in this material.

Experimental results for silicon

With reference to the foregoing model, we shall now discuss the results of our measurements on single crystals of silicon [9]. We begin with the spectral distribution of the emission in the case of weakly doped materials, i.e. both *P*-type and *N*-type containing 10^{16} charge carriers per cm^3 (*fig. 5*). We see in the first place that the spectral distribution is the same for both *P*-type and *N*-type material, indicating that the doping in each case must be weak enough to allow us to consider the bands as straight (the space charge layer being very much greater than the escape depth for photoelectrons).

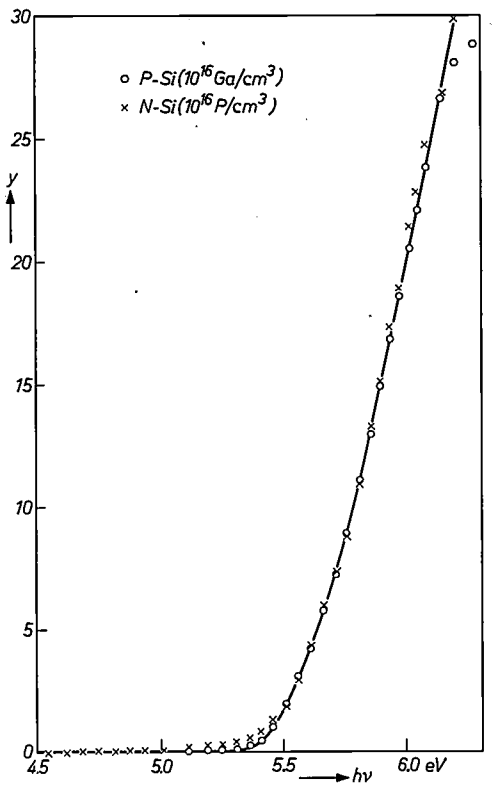


Fig. 5. Yield y of the photoemission for *P*-type and *N*-type silicon (1×10^{16} charge carriers/cm^3) as a function of photon energy $h\nu$. The solid curve represents the relation $y = 42(h\nu - 5.4)^{3/2}$. The scale for y is arbitrary.

Secondly, we see that the curve consists of two parts, i.e. a "tail" with a very low yield followed by a steeply rising part beginning at 5.4 eV. We interpret the tail as emission from surface levels. In the following we shall put forward two arguments in support of this. The steep section of the curve is attributed to the valence band emission. The threshold energy for this emission is thus 5.4 eV. This part of the curve appears to

[9] J. van Laar and J. J. Scheer, Philips Res. Repts. **17**, 101, 1962, and Rep. Int. Conf. on the Physics of Semiconductors, Exeter 1962, p. 827 (publ. Inst. Phys./Phys. Soc., London 1962).

[10] G. W. Gobeli and F. G. Allen, Phys. Rev. **127**, 141, 1962.

[11] The values of $\varphi - 4.85$ eV for weakly doped and 4.90 eV for strongly doped *P*-type material — indicate that the stabilization of the Fermi level is not complete.

be well described by the purely empirical expression $y \propto (h\nu - 5.4)^{3/2}$, where $h\nu$ is the energy of the photons in electron volts.

Let us now consider the emission from more strongly doped *P*-type material (1×10^{19}, 6×10^{19} and 1×10^{20} charge carriers per cm^3). The theory predicts that the stronger the doping the lower the energy at which emission will begin, owing to the above-mentioned band bending caused by the presence of surface levels (fig. 4, curves *2* and *3*). This is in fact clearly evident from the spectral distributions measured (*fig. 6*). The influence of doping can be quantitatively predicted if we assume that the Fermi level at the surface of silicon is stabilized and that the band emission for straight bands is given by $y = C(h\nu - 5.4)^{3/2}$. This calculation is given in the appendix of this article. The fixed position of the Fermi level in relation to the energy of an electron *in vacuo* (given by φ) and the escape depth d can be fitted to the experimental results. For $d = 2.0$ nm and $\varphi = 4.90$ eV a reasonable agreement is found between calculation and experiment (*fig. 7a, b, c*). The value found here, $\varphi = 4.90$ eV, is consistent with the value

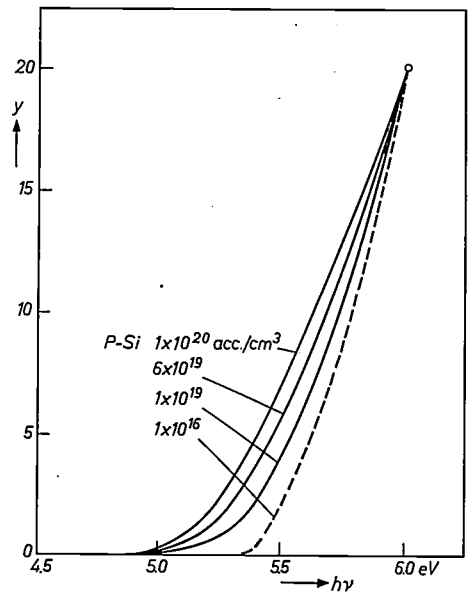


Fig. 6. Influence of doping on the photoemission of silicon. The yield y is plotted against the photon energy $h\nu$, for three volume doping contents: 1×10^{19}, 6×10^{19}, 1×10^{20} acceptors/cm^3. For comparison the yield for weakly doped material (1×10^{16} acceptors/cm^3) is also shown (dashed line, taken from fig. 5). The curves are reduced to the same y-value at $h\nu = 6.0$ eV.

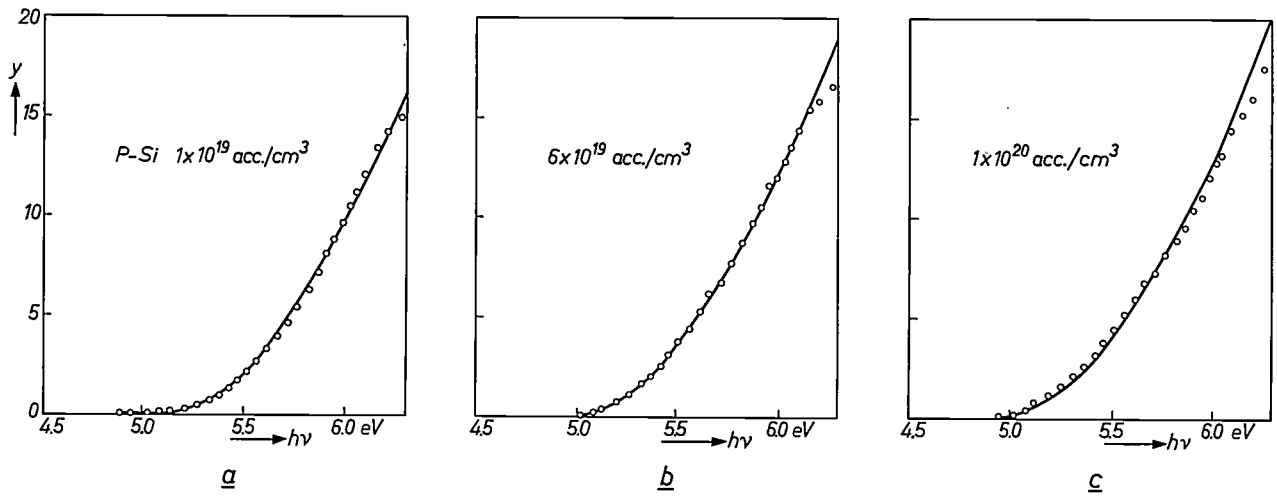


Fig. 7. Comparison of measured and calculated spectral distribution of the photoemission for *P*-type silicon. *a*) 1×10^{19} acceptors/cm^3, *b*) 6×10^{19} acceptors/cm^3, *c*) 1×10^{20} acceptors/cm^3. The calculation (see Appendix) was carried out for a stabilized Fermi level with $\varphi = 4.9$ eV and an escape depth of $d = 2.0$ nm. These values give the best fit with the experimental results.

derived by other authors from contact potential measurements [10].

This brings us to the first argument for the above-mentioned interpretation of the tail emission of weakly doped material as being emission from surface levels. Turning to fig. 7, we see that the curves for strong doping no longer exhibit any extra tail in the emission. This is precisely what we would expect for emission from surface levels: as it does not follow the shift to lower energy with stronger *P*-type doping, it disappears entirely in the stronger band emission.

The second argument in support of our interpretation is that highly accurate measurements on weakly doped material show that the starting point of the tail emission lies at 4.85 eV (*fig. 8*), and this, according to the above-mentioned contact potential measurements by Gobeli and Allen [10] is precisely the position of the Fermi level [11]. (These authors, however, give

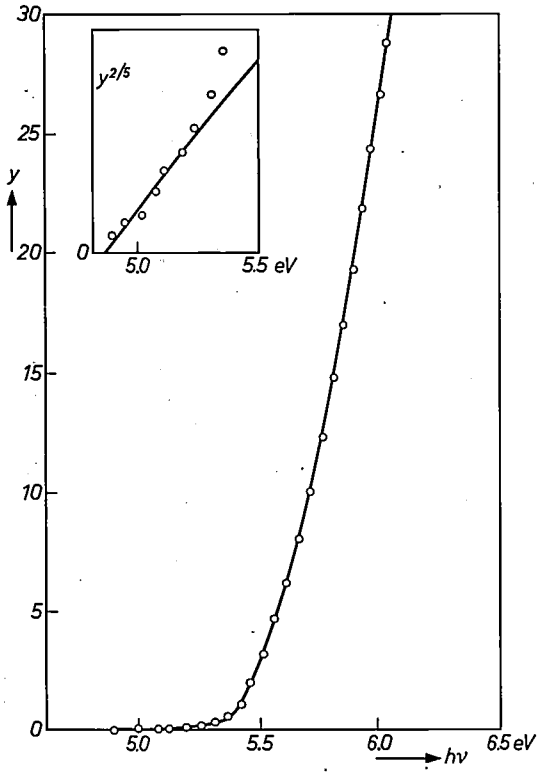


Fig. 8. Photoemission of silicon with weak *P*-type doping (1.6×10^{16} Ga/cm³): the yield y as a function of photon energy $h\nu$. In the inset, $y^{2/5}$ is plotted against $h\nu$. The solid curve represents the relation

$$y = (h\nu - 4.85)^{5/2} + 28.3(h\nu - 5.36)^{3/2}.$$

The tail emission ($h\nu < 5.4$ eV) is attributed to emission from surface levels.

a different interpretation of the tail emission.) This part of the emission appears to satisfy the empirical expression $y \propto (h\nu - 4.85)^{5/2}$. Summarizing, the entire spectral distribution for weakly doped material is therefore given by:

$$y = C_1(h\nu - 4.85)^{5/2} + C_0(h\nu - 5.4)^{3/2},$$

where C_0/C_1 is found to be approximately equal to 30.

The photoemission of silicon can therefore be described by assuming that the Fermi level at the surface is stabilized at 4.9 eV below vacuum and that E_A has the value $5.4 - 1.1 = 4.3$ eV ($E_g = 1.1$ eV). Thus, at the surface, the Fermi level lies 0.6 eV below the bottom of the conduction band.

This stabilization of the Fermi level occurs for *P*-type volume doping up to 10^{20} cm^{-3}. Using equation (2) we can calculate the number of filled acceptors per unit surface of the space charge layer. Taking $N = 10^{20}$ cm^{-3}, $\Delta V = 0.5$ volt, $\varepsilon = \varepsilon_0\varepsilon_r$, $\varepsilon_r = 11.8$, $\varepsilon_0 = 8.9\times10^{-12}$ F/m and $e = 1.6\times10^{-19}$ C, we find this number to be $Nx_0 = (2\Delta V\varepsilon N/e)^{1/2} \approx 10^{13}$ cm^{-2}. To supply this charge there must be at least 10^{13} surface levels per cm². These levels must be localized within an energy region of 0.1 eV, as follows from the accuracy with which the stabilization of the Fermi level was determined.

The 5/2 power law found for the emission from surface levels proves to be quite common, also in cases where the surface levels are due to impurities [12]. The 5/2 exponent was found, for example, for the tail emission of cadmium telluride surfaces with impurities due to cleavage in air or baking *in vacuo*. The tail emission here must definitely be attributed to the impurities, because a clean surface, obtained by vacuum cleavage does not give this emission [13].

Experimental results for gallium arsenide

In the case of silicon we found a very nearly stabilized Fermi level at the surface. Our results for gallium arsenide [14] clearly show that the Fermi level here is *not* stabilized and that the situation in fact conforms fairly well to the simple straight-band model (see fig. 3*b*).

The first indication of the lack of surface levels is the spectral distribution of the photoemission from weakly doped *P*-type material (*fig. 9*). The greater part of this distribution can be described by the empirical relation $y \propto (h\nu - 5.57)^{3/2}$, and only a small additional tail occurs at low photon energies. In the case of silicon the 3/2 power was characteristic of valence-band emission. If this applies to gallium arsenide too, the small tail emission means that the surface levels are present only in a low concentration, so that there may well be no stabilization of the Fermi level here. This is in fact clearly evident from the very limited influence of bulk doping on the photoemission (*fig. 10*).

For a stabilized Fermi level this influence can be calculated as in the case of silicon. Assuming that the escape depth in GaAs is also 2.0 nm, we have made the calculation for two positions of the Fermi level, given respectively by $\delta = 0.1$ eV and $\delta = 0.7$ eV at the surface. From *figs. 11* and *12* it appears that the first choice gives only a poor fit with the experimental results and that the second does not fit at all. We conclude therefore that the Fermi level is not stabilized.

This conclusion is confirmed by the results of measurements of the effect of bulk doping on the thermionic work function φ. To obtain information on this point, contact potential measurements were carried out to determine the difference in φ between *N*-type and *P*-type material [14]. The results of these experiments showed that, even in the case of weakly doped material, the difference in φ roughly corresponded to the differ-

[12] J. J. Scheer and J. van Laar, Physics Letters **3**, 246, 1963.
[13] J. J. Scheer and J. van Laar, Philips Res. Repts. **16**, 323, 1961.
[14] J. van Laar and J. J. Scheer, Surface Sci. **8**, 342, 1967.
[15] W. E. Spicer, J. appl. Phys. **31**, 2077, 1960.

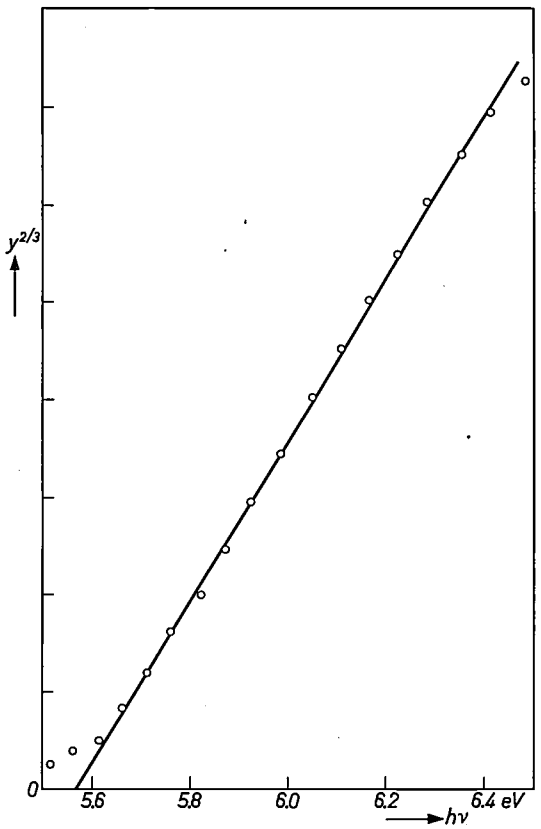


Fig. 9. Photoemission of GaAs with weak *P*-type doping (2.5×10^{17} acceptors/cm^3). The 2/3 power of the yield y is given as a function of photon energy $h\nu$. The tail emission is of little significance.

ence in the position of the Fermi level in the volume. This ties up with the assumption that the bands run more or less straight at the surface.

Photocathodes for visible light; lowering of the work function by adsorption

We shall now deal at somewhat greater length with the problem, mentioned in the introduction, of making a satisfactory photocathode for visible light. First, it should be stated that good photocathodes in use at the present time can have a very high maximum yield, with values up to 0.30 (expressed in electrons per incident photon). A yield as high as this can scarcely be improved upon. The Na_2KSb-Cs photocathode [15] has its threshold wavelength at 0.85 μm; this is a sufficiently long wavelength for the detection of visible light. Existing cathodes therefore meet the requirements for visible-light detectors reasonably well.

All the good photocathodes were evolved more or less empirically. Later it turned out that all efficient cathodes were *P*-type semiconductors. The question that now arises is whether more scientific methods could be applied to the development of photocathodes, using known semiconductors.

A high yield can only be obtained with valence band emission. In the example of silicon given above, the threshold for photoemission with strong *P*-type doping

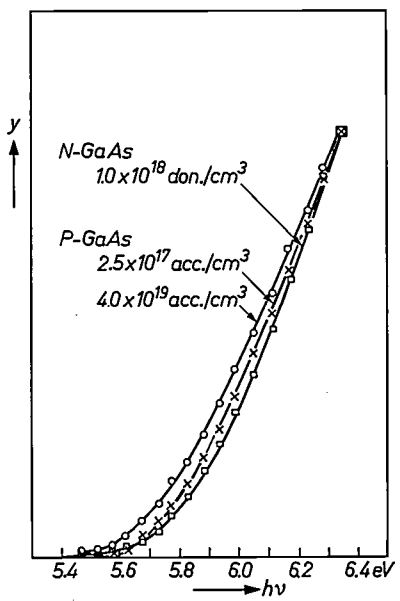


Fig. 10. Influence of volume doping on the photoemission of GaAs. The yield y is given as a function of photon energy $h\nu$ for *P*-type material, 4.0×10^{19} acceptors/cm^3 and 2.5×10^{17} acceptors/cm^3 and for *N*-type, 1.0×10^{18} donors/cm^3. The curves have been reduced to the same y-value at $h\nu = 6.35$ eV.

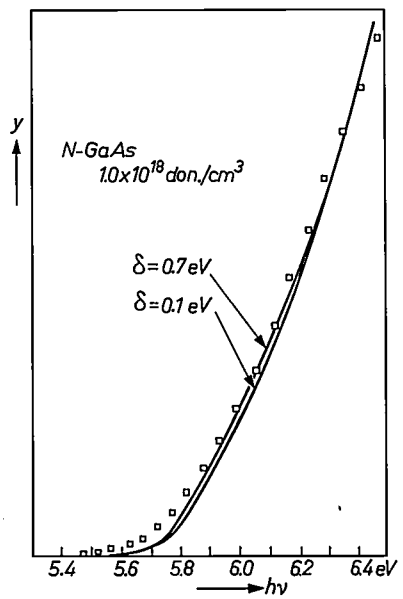


Fig. 11. The yield y as a function of $h\nu$ for *N*-GaAs, 1×10^{18} donors/cm^3, as found theoretically for two values of δ and an escape depth d of 2.0 nm. The experimental points (squares) are the same as in fig. 10. The curves have again been reduced to the same y-value at $h\nu = 6.35$ eV.

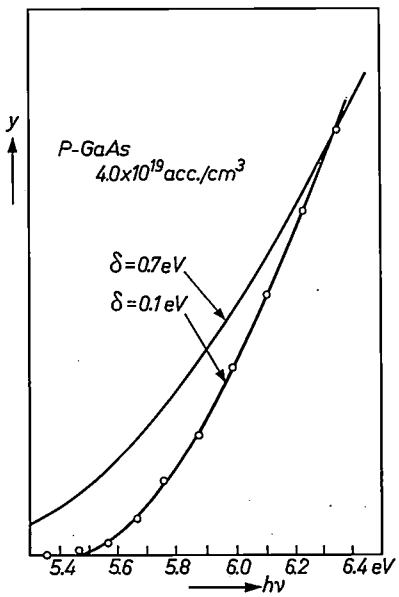


Fig. 12. As fig. 11, but now for *P*-GaAs, 4×10^{19} acceptors/cm^3. Only a few of the corresponding experimental points of fig. 10 are shown (the circles). The zero point of the theoretical curve (i.e. the threshold energy E_d) for $\delta = 0.7$ eV is at $E_A + E_g - \delta = 5.57 - 0.7 = 4.87$ eV.

was 4.9 eV, that is to say in the far ultraviolet (0.25 μm). In the case of GaAs it is not possible to shift the emission to lower energy by means of band bending, and E_d here is even higher (5.6 eV). Values as high as this are found with most semiconductors.

As we mentioned in the introduction, however, a method does exist by which the threshold for photoemission can be shifted to a lower energy, i.e. by the adsorption of electropositive metal atoms on the surface of the semiconductor. We shall now consider this method.

charge. The negative space charge causes band bending: the potential inside the semiconductor rises in relation to the vacuum level. Upon the further adsorption of atoms the band bending increases. This process continues until the Fermi level has risen to a level slightly above the bottom of the conduction band at the surface (fig. 13*b*). As soon as this has happened, the electrons of the adsorbed atoms enter the conduction band, in which the number of available states is very large; there is no further increase of the space charge and therefore no further band bending. The metal ions, to-

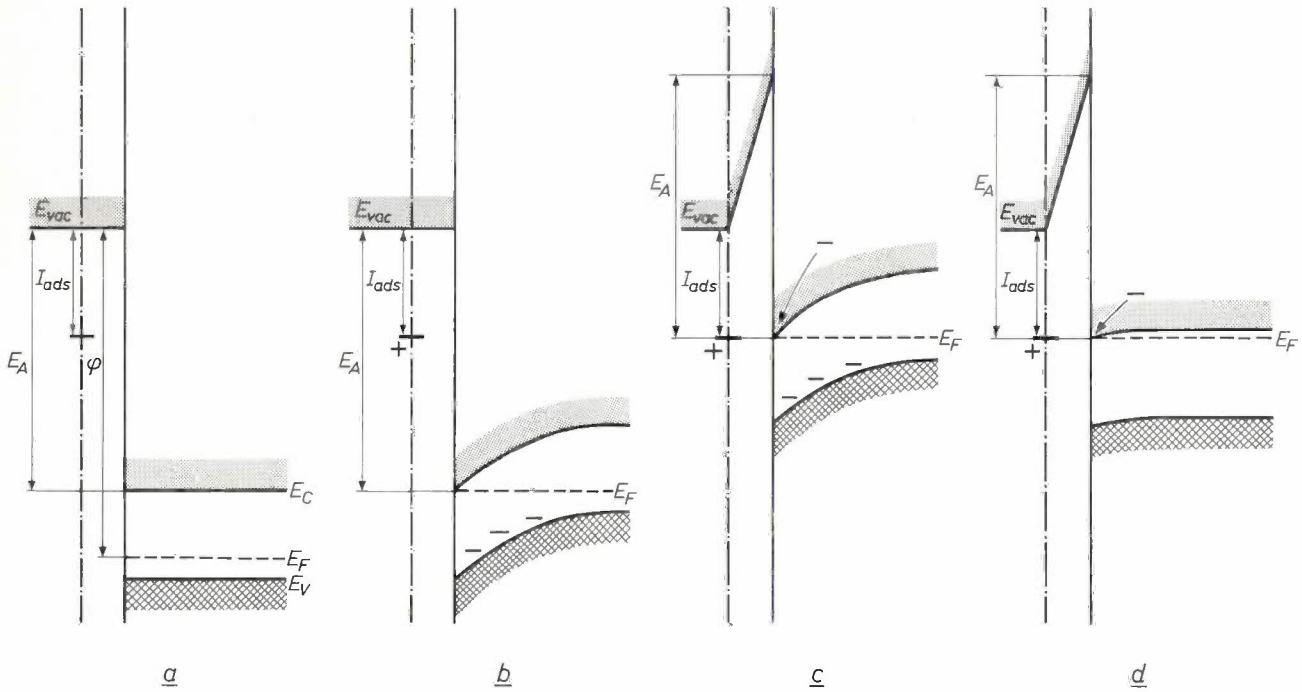


Fig. 13. Lowering of the threshold energy for photoemission by the adsorption of metal atoms. *a*) If the ionization energy I_{ads} of an adsorbed metal atom is lower than the work function of the clean surface of *P*-type material, the metal atoms give up electrons to the acceptors in the bulk. This gives rise to band bending, which continues until the Fermi level coincides with the bottom of the conduction band at the surface (*b*). Upon further adsorption the donated electrons enter the conduction band. They form together with the positive metal ions a dipole layer, as a result of which the levels in the semiconductor are raised with respect to the vacuum level. This continues until the optimum situation *c* is reached, where $\varphi = I_{ads}$. In *N*-type material (*d*) there is virtually no band bending; almost right from the beginning a dipole layer is formed in this material by the positive metal atoms and the donated electrons in the conduction band.

Suppose we have a *P*-type semiconductor whose Fermi level at the surface is not stabilized (see *fig. 13*). Metal atoms are now adsorbed on this surface (fig. 13*a*). Let the ionization energy of these atoms in the adsorbed state be I_{ads} (this need not be equal to the ionization energy of the free atom). If I_{ads} is smaller than the thermionic work function φ of the semiconductor, the metal atoms will give up an electron to the acceptors in the bulk of the semiconductor. As a result, a negative space charge arises in the semiconductor, and the metal ions form an opposite surface

gether with the electrons in the conduction band, form a dipole layer at the surface. These electrons can be regarded as the image charge of the ions: the surface has now acquired a metallic character. The Fermi level rises with respect to the vacuum level as a result of the increasing potential drop across the dipole layer. This continues until the Fermi level coincides with the donor level of the adsorbed atom (I_{ads} below vacuum level). In this situation $\varphi = I_{ads}$ (fig. 13*c*). Upon further adsorption the semiconductor accepts no more electrons.

Since the electrons are able to tunnel through the

potential barrier of the dipole layer, the threshold energy for photoemission E_d finally falls to I_{ads} in strongly *P*-type doped material. In material with weak *P*-type doping, after adsorption of metal atoms, ("straight bands") the threshold energy is $E_d = I_{ads} + E_g$, as in *N*-type material, in which hardly any band bending occurs (fig. 13*d*).

We shall now examine what happens when metal atoms are adsorbed on a semiconductor surface with a stabilized Fermi level (*fig. 14*). If I_{ads} is sufficiently low, the metal atoms will give up their electron to the surface states at the Fermi level (fig. 14*a*). The adsorbed ions now form together with the resultant negative surface charge a dipole layer which causes the potential in the semiconductor to rise with respect to vacuum. As adsorption continues a maximum lowering of the threshold is reached when the Fermi level has risen to I_{ads} below vacuum level (fig. 14*b*). Thereafter the adsorbed atoms have no further reason to give up an electron. In this situation the thermionic work function is again $\varphi = I_{ads}$. For weakly doped material the long wavelength threshold is now $E_d = I_{ad} + \delta$. For strong *P*-type doping we again find $E_d \approx I_{ads}$.

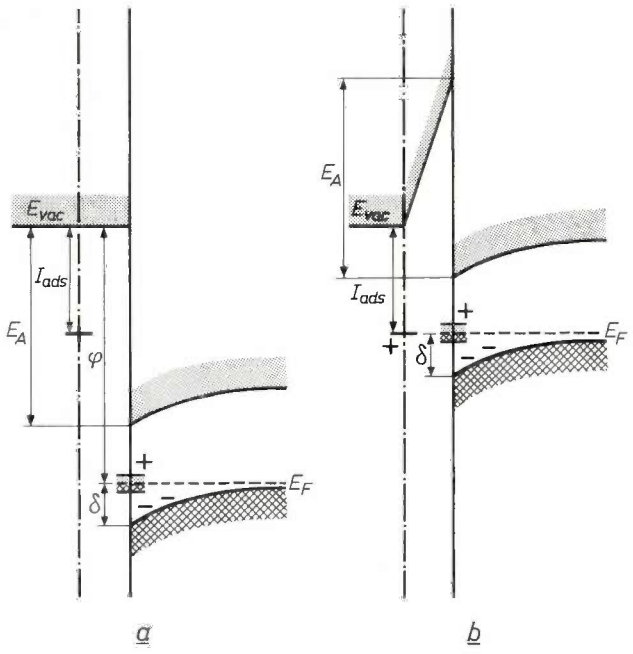


Fig. 14. Adsorption of metal on a *P*-type semiconductor with stabilized Fermi level. If the ionization energy I_{ads} of an adsorbed atom is lower than the work function φ of the semiconductor without metal (see *a*), the metal atoms give up electrons to the surface levels. The dipole layer formed raises the levels in the semiconductor with respect to the vacuum level. In the optimum situation (*b*), φ is again equal to I_{ads}.

[16] J. J. Scheer and J. van Laar, Solid State Comm. **5**, 303, 1967.

The behaviour of some semiconductors is not entirely in accordance with this model. If the surface of these materials is clean, the Fermi level is not stabilized, but the surface, upon adsorbing metal atoms, behaves as if the Fermi level were in fact stabilized. In this case the adsorption of foreign atoms evidently gives rise to excess surface levels. Gallium arsenide is a typical example of such a semiconductor [16]. No satisfactory explanation has yet been found for this phenomenon.

If one is looking for a photocathode with a low threshold energy, the above considerations show that one should take a strongly *P*-type doped semiconductor and cover it with metal atoms for which I_{ads} is small; whether or not the Fermi level is stabilized is then immaterial as far as the threshold energy is concerned, this being equal to I_{ads}.

The lowest value of I_{ads} is found for caesium, at about 1.4 eV (0.90 μm wavelength), virtually independent of the substrate.

For *P*-type silicon coated with Cs this value has in fact been found as the threshold for valence band emission [9]. By varying the volume doping it is possible to estimate the escape depth for electrons in the relevant energy region. The result is $d \approx 20$ nm, a value also found for practical photocathodes [15].

The yield of *P*-type silicon with Cs is very low, however (<1%). This is due to the fact that the electron transitions in silicon, caused by light with a photon energy equal to or slightly higher than 1.4 eV, are *indirect transitions*. These are transitions between states of *different momentum*, the electron exchanging momentum with the lattice vibrations. In a diagram giving the energy *E* against momentum *k* (see *fig. 15*) these are represented by an oblique arrow. The yield is low with indirect transitions for two reasons:

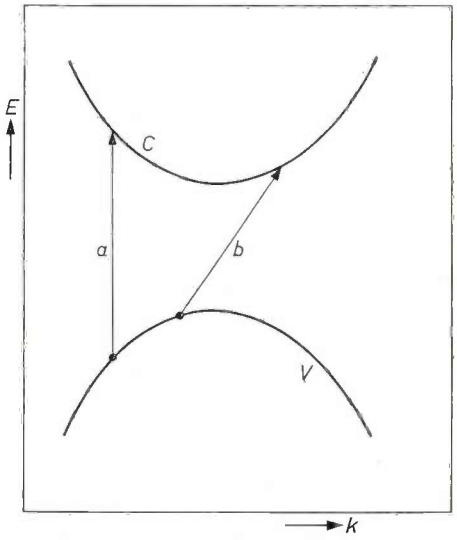


Fig. 15. Energy *E* as a function of momentum (*k*) for electrons in a semiconductor. *C* is the conduction band, *V* the valence band. *a* is a direct transition, *b* an indirect transition.

1) The transition probability is very small for an indirect transition. This is the opposite to the case of a *direct transition*, where the momentum is preserved, so that no lattice vibrations are needed to maintain the balance of momentum. Because of this the absorption of the relevant light is weak and only a small fraction of the light is absorbed in the escape layer. In silicon the penetration depth is roughly 1 μm, which is very much greater than the escape depth found of 20 nm. Thus, the majority of the electrons excited do not reach the surface, and therefore do not contribute to the photoemission.

2) Even the light that is absorbed in the escape layer is to a great extent ineffective, for in principle, at a given $h\nu$, an electron can go from any state in the valence band to any state in the conduction band, provided the energy difference agrees. In a large percentage of these transitions the final state of the electron is a state in the conduction band below E_{vac} (see *fig. 16*). These are known as *competing transitions.*

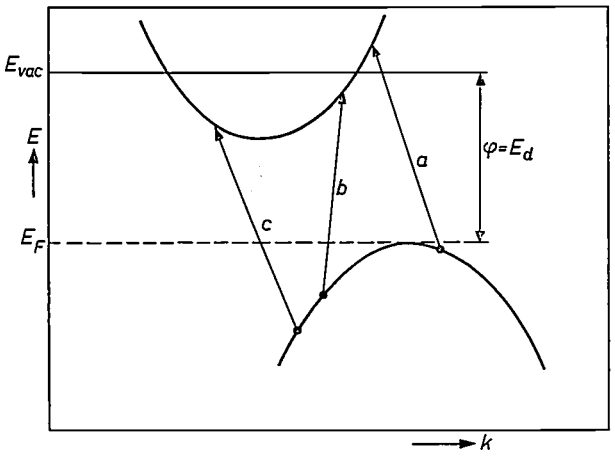


Fig. 16. Indirect transitions in silicon with strong *P*-type doping (E_F lies at the top of the valence band), coated with Cs ($\varphi = E_{vac} - E_F = 1.4$ eV), *a* is a transition where the electron receives an energy higher than E_{vac}; this electron may be emitted; *b* and *c*, competing transitions where the excited electron cannot be emitted. The energy difference $h\nu$ for *a*, *b* and *c* is the same, and greater than $E_d = 1.4$ eV.

This being said, it is obviously of interest to consider semiconductors in which the relevant transitions are direct ones. The disadvantage mentioned under item (2) is then not encountered, for at a given $h\nu$ there is generally only one set of states in the valence and conduction bands possessing the same momentum; there are no competing (direct) transitions. As far as item (1) is concerned, with direct transitions stronger light absorption is usually found, because the probability of a direct transition is so much greater than for an indirect transition. On the other hand it could well be that the number of possible indirect transitions is much greater than the number of possible direct ones.

Let us confine ourselves to the simple situation in which the extremes of the two bands lie at the same k value. We again consider semiconductors with strong *P*-type doping, so that E_F coincides with the top of the valence band, and which are coated with Cs, so that $E_{vac} - E_F = \varphi = 1.4$ eV. *Fig. 17* shows three cases: a) $E_g > \varphi$, b) $E_g = \varphi$, c) $E_g < \varphi$. Case (a) is unfavourable since the electron must first enter the conduction band before it can be emitted; the threshold for photoemission is thus $E_d = E_g > \varphi$. Again, in case (c) the threshold for direct transitions is shifted towards the blue, because the point where E_{vac} intersects the conduction band corresponds to a valence band level below E_F. Only in case (b) is E_d not greater than but equal to φ. It follows, then, that we should choose semiconductors in which the energy gap relevant to direct transitions, i.e. the *optical* energy gap, is roughly equal to 1.4 eV. (It may also be deduced from fig. 17*c* that, at a given $E_g < \varphi$, the value of E_d approaches closer to φ as the valence band becomes flatter and the conduction band sharper, in other words as the effective mass of the holes in the valence band becomes greater and that of the electrons in the conduction band smaller.)

Semiconductors whose energy gap does not differ too widely from 1.4 eV are CdTe ($E_g = 1.5$ eV), GaAs ($E_g = 1.35$ eV) and InP ($E_g = 1.29$ eV). CdTe is immediately ruled out because it cannot be given the necessary high *P*-type doping content. The hope that the absorption of the light might be stronger than for silicon, since direct transitions are involved, is not fulfilled by the other two materials; the penetration depth for light with $h\nu$ values greater than E_g is of the order of 1 μm, as in silicon.

This penetration depth is much too great as long as we are concerned with escape depths of 20 nm. This does not however apply to the cases of fig. 17*a* and *b*. In these cases the bottom of the conduction band in the bulk is above the vacuum level (fig. 17*a*) or coincides with it (fig. 17*b*). This means that electrons in the conduction band possessing thermal energy are already capable of escaping into the vacuum. In other words, excited electrons can still be emitted after they have entirely lost their kinetic energy by collisions in the conduction band. Here, then, the escape depth is no longer determined by the mean free path of high-energy electrons but by the diffusion-recombination length. This lies roughly between 0.1 and 1 μm. In these cases, therefore, a penetration depth of 1 μm for light is very reasonable. GaAs is a material that closely approximates to the situation in fig. 17*b*.

These considerations led us to make an experimental

photocathode of GaAs, coated with Cs, which gave surprisingly good results. Single crystals of GaAs were used, with 3×10^{19} acceptors/cm^3, cleaved *in vacuo* [17]. The measured spectral distribution (*fig. 18*) shows that the threshold for photoemission is 1.35 eV (0.90 μm), which is near the expected value. On the blue side of the visible spectrum (0.4 μm) the yield is about 0.30 electrons per incident photon; at 3.5 eV (0.35 μm) it rises to as much as 0.35. After correction for reflected light, a yield is found of roughly 0.50 electrons per *absorbed* quantum, which is the maximum that can be expected (since half the excited electrons go away from the surface and into the bulk of the material). The sensitivity to light with a colour temperature of 2850 °K is 500 μA/lm. This photocathode is therefore even better than the Na_2KSb-Cs cathode, which has a long wavelength threshold of 0.82 μm and a sensitivity of 300 μA/lm.

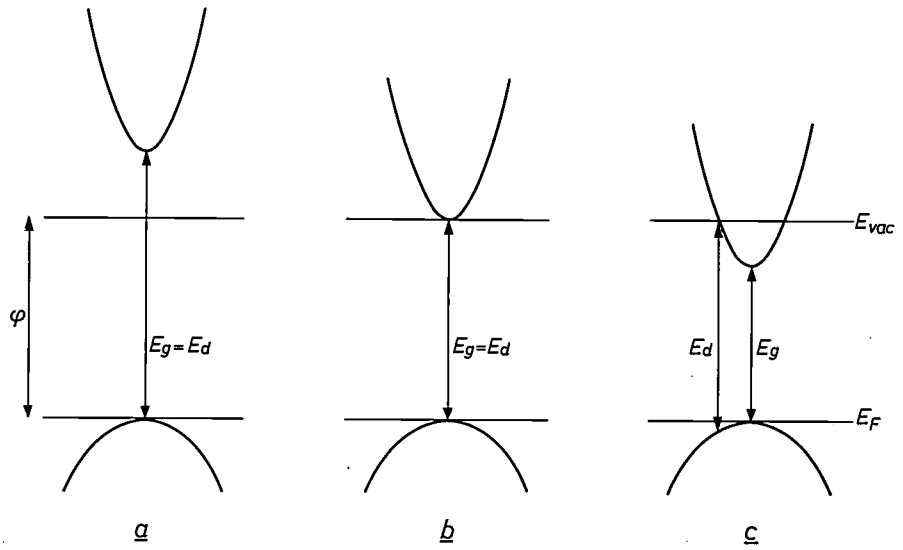


Fig. 17. *P*-doped semiconductors with a simple band structure, with the same value of φ (given by the deposited metal) but with different optical energy gap E_g. *a*) $E_g > \varphi$, *b*) $E_g = \varphi$, *c*) $E_g < \varphi$. Only in case (*b*) is the threshold for direct transitions equal to φ; both in (*a*) and (*c*) the threshold E_d is greater than φ.

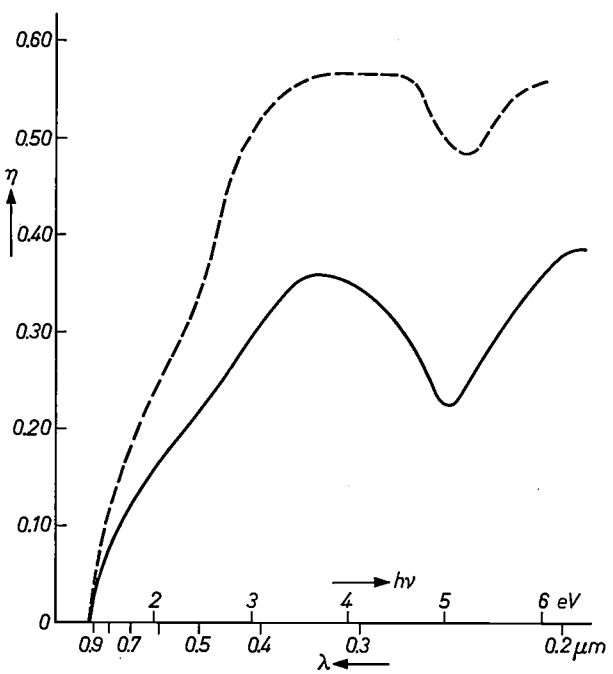


Fig. 18. Spectral distribution of η, the yield per photon, of the photoemission of *P*-type GaAs (3×10^{19} acceptors/cm^3) coated with Cs. The solid curve gives the yield in electrons per incident photon, the dashed curve the yield in electrons per absorbed photon.

Appendix: Calculation of the effect of bulk doping on the photoemission of silicon

We take a *P*-type semiconductor and choose a co-ordinate system in which the x-axis is perpendicular to the surface of the semiconductor; $x = 0$ at the surface, and $x > 0$ in the bulk.

We make the following assumptions:

1) The escape probability of an electron excited at a distance x from the surface has the form $\exp(-x/d)$, the escape depth d being independent of the energy of the excited electron.

2) The intensity of the light at a distance x from the surface is given by $I_0 \exp(-x/m)$, the penetration depth m being independent of the photon energy.

3) The photoemission of a crystal with straight bands is given by $y = C(h\nu - E_0)^{3/2}$, as we found empirically.

The contribution Δy to the photoemission originating from a layer of thickness Δx at a distance x from the surface is given for straight bands by:

$$\Delta y(x) = A(h\nu - E_0)^{3/2}\, e^{-x/L}\, \Delta x,$$

where $L^{-1} = d^{-1} + m^{-1}$ and A is an arbitrary constant.

[17] J. J. Scheer and J. van Laar, Solid State Comm. 3, 189, 1965.

Suppose now, that as a result of band bending the valence band at position x is shifted to higher energy by an amount $\delta E(x)$ with respect to its position at the surface. We now make the additional assumption that the contribution to the photoemission is independent of the state in the conduction band occupied by the excited electron. We then have for bent bands:

$$\Delta y(x) = A\{h\nu + \delta E(x) - E_0\}^{3/2} e^{-x/L} \Delta x. \quad . \quad . \quad . \quad (3)$$

According to eqs. (1) and (2), for a Schottky layer:

$$\delta E(x) = e\{V(x_0) - V(x)\} = \Delta E\left\{1 - \left(\frac{x - x_0}{x_0}\right)^2\right\}, \quad (4)$$

where $\Delta E = e\Delta V$ is the total band bending. Substitution of (4) in (3) and integration with respect to x gives:

$$y(h\nu) = A \int_0^{x_0} \left[h\nu - E_0 + \Delta E\left\{1 - \left(\frac{x - x_0}{x_0}\right)^2\right\}\right]^{3/2} e^{-x/L}\, dx + \\ + AL(h\nu - E_0 + \Delta E)^{3/2} e^{-x_0/L}.$$

The last term in this equation originates from the part of the crystal outside the space charge region.

The value of the integral for a given x_0, ΔE and L has to be determined numerically.

Summary. The article discusses a study made of the surface properties of semiconductors by investigating their photoemission. The spectral distribution of the photoemission was measured on vacuum-cleaved single crystals of silicon and gallium arsenide with various amounts of *P*-type and *N*-type doping. The results are interpreted in terms of the band theory of solids, where the bands can be bent at the surface by a space charge due to surface states and where, in the presence of excess surface states, the Fermi level may be stabilized in the surface levels. The emitted electrons come from a surface layer of thickness *d*, the escape depth. In weakly doped silicon the main emission comes from the valence band and "tail emission" from the surface states. In Si the Fermi level is stabilized; with increasing impurity content, the layer where band bending occurs finally becomes thinner than the escape depth; this explains the observed change in the spectral distribution of the emission. In the case of GaAs it is concluded that there are virtually no surface levels, no band bending and no stabilization of the Fermi level. For photoemission in the visible region the threshold energy of nearly all semiconductors is too high (~ 5 eV); it can be lowered by the adsorption of metal atoms, e.g. of Cs (to 1.4 eV). The yield of *P*-Si coated with Cs is small due to a) a too small absorption coefficient and b) "competing transitions" that do not result in emission. Strongly *P*-doped GaAs coated with Cs may be expected to be a good photoemitter in the visible region since the competing transitions are of little significance and the escape depth is unusually large owing to the vacuum level and the bottom of the conduction band being coincident. This has been confirmed by experiment. An experimental photocathode of this composition has a long wavelength threshold at 0.9 μm and an outstanding photoemmission: the yield is 0.35 electrons per incident photon at 0.35 μm.

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, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, 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.

G. A. Acket: Determination of the Hall mobility of hot electrons in gallium arsenide using 8 mm microwaves.
Physics Letters **25A**, 374-376, 1967 (No. 5). *E*

L. K. H. van Beek, J. Helfferich, H. Jonker & Th. P. G. W. Thijssens: Properties of diazosulfonates, Part II. The rate of the reaction between 2-methoxybenzenediazonium and sulfite ions;
L. K. H. van Beek, J. Helfferich, H. J. Houtman & H. Jonker: idem, Part III. The absorption spectra of *p*-substituted-benzene-*trans*-diazosulfonates; Part IV. The absorption spectra of *o*- and *m*-substituted benzenediazosulfonates.
Rec. Trav. chim. Pays-Bas **86**, 749-754, 975-980, 981-985, 1967 (Nos. 8, 10). *E*

G. Blasse & A. Bril: A new phosphor for flying-spot cathode-ray tubes for color television: yellow-emitting $Y_3Al_5O_{12}$-Ce^{3+}.
Appl. Phys. Letters **11**, 53-55, 1967 (No. 2). *E*

G. Blasse & A. Bril: Structure and Eu^{3+}-fluorescence of lithium and sodium lanthanide silicates and germanates.
J. inorg. nucl. Chem. **29**, 2231-2241, 1967 (No. 9). *E*

G. Blasse & A. Bril: Investigations of Tb^{3+}-activated phosphors.
Philips Res. Repts. **22**, 481-504, 1967 (No. 5). *E*

P. Blume & A. Stecker: Physikalische Eigenschaften von Lochstreifenpapier, I, II, III.
Feinwerktechnik **71**, 262-271, 325-334, 518-527, 1967 (Nos. 6, 7, 11). *H*

A. J. van Bommel & F. Meyer: A low energy electron diffraction study of the PH_3 adsorption on the Si (111) surface.
Surface Sci. **8**, 381-398, 1967 (No. 4). *E*

A. J. van Bommel & F. Meyer: LEED study of a nickel induced surface structure on silicon (111).
Surface Sci. **8**, 467-472, 1967 (No. 4). *E*

J. C. Brice: Diffusive and kinetic processes in growth from solution.
J. Crystal Growth **1**, 161-163, 1967 (No. 3). *M*

A. Broese van Groenou: Permeability of some oxides and metals at low temperatures.
J. appl. Phys. **38**, 3317-3320, 1967 (No. 8). *E*

A. Broese van Groenou: Magnetic after-effects in high-amplitude fields.
Philips Res. Repts. **22**, 463-480, 1967 (No. 5). *E, M*

K. H. J. Buschow & J. H. N. van Vucht: Comments on "The aluminium-rich parts of the Al-Sm and Al-Dy systems".
J. less-common Met. **13**, 369-370, 1967 (No. 3). *E*

H. B. G. Casimir: Van der Waals-Wechselwirkungen.
Naturwiss. **54**, 435-438, 1967 (No. 15/16). *E*

A. Claassen & L. Bastings: The determination of aluminium with 8-hydroxyquinoline, Part I. Precipitation in acetate-buffered solution;
A. Claassen, L. Bastings & J. Visser: idem, Part II. Precipitation in ammoniacal cyanide-EDTA solution.
Analyst **92**, 614-617, 618-621, 1967 (No. 1099). *E*

J. B. Davies: Propagation in rectangular waveguide filled with skew uniaxial dielectric.
IEEE Trans. **MTT-15**, 372-376, 1967 (No. 6). *M*

H. Dormont, J. Frey, J. Salmon & M. Valton: Contribution à l'établissement d'une nouvelle équation cinétique.
C. R. Acad. Sci. Paris **264A**, 978-980, 1967 (No. 22). *L*

W. F. Druyvesteyn: The resistivity of hard superconductors subjected to an increasing field.
Physics Letters **25A**, 31-32, 1967 (No. 1). *E*

P. Eckerlin, A. Rabenau & H. Nortmann: Zur Kenntnis des Systems Be_3N_2-Si_3N_4, III. Darstellung und Eigenschaften von $BeSiN_2$;
P. Eckerlin: idem, IV. Die Kristallstruktur von $BeSiN_2$.
Z. anorg. allgem. Chemie **353**, 113-121, 225-235, 1967 (Nos. 3/4, 5/6). *A*

I. Flinn, G. Bew & F. Berz: Low frequency noise in MOS field effect transistors.
Solid-State Electronics **10**, 833-845, 1967 (No. 8). *M*

L. Fraiture & J. Neirynck: Optimum elliptic-function filters for distributed constant systems.
IEEE Trans. **MTT-15**, 482-483, 1967 (No. 8). *B*

A. H. Gomes de Mesquita: Refinement of the crystal structure of SiC type 6*H*.
Acta cryst. **23**, 610-617, 1967 (No. 4). *E*

A. H. Gomes de Mesquita: The crystal structure of 2,2′-diaminodiphenyl disulphide.
Acta cryst. **23**, 671-672, 1967 (No. 4). *E*

G. E. G. Hardeman, G. B. Gerritsen & R. P. van Stapele: Effect of random crystal fields on the electron-nuclear double-resonance spectrum of cobalt-doped ZnSe.
Phys. Rev. **160**, 281-286, 1967 (No. 2). *E*

E. Himmelbauer & J. C. Francken: On the sensitivity of oscilloscope tubes.
Philips Res. Repts. **22**, 515-540, 1967 (No. 5).

F. M. Klaassen: High-frequency noise of the junction field-effect transistor.
IEEE Trans. **ED-14**, 368-373, 1967 (No. 7). *E*

F. M. Klaassen & J. Prins: Thermal noise of MOS transistors.
Philips Res. Repts. **22**, 505-514, 1967 (No. 5). *E*

G. Klein & H. Koelmans: Active thin film devices.
Festkörperprobleme **7**, 183-199, 1967. *E*

E. Kooi: The surface properties of thermally oxidized silicon.
Festkörperprobleme **7**, 132-157, 1967. *E*

W. Kwestroo & P. H. G. M. Vromans: Preparation of three modifications of pure tin (II) oxide.
J. inorg. nucl. Chem. **29**, 2187-2190, 1967 (No. 9). *E*

J. van Laar & J. J. Scheer: Influence of volume dope on Fermi level position at gallium arsenide surfaces.
Surface Sci. **8**, 342-356, 1967 (No. 3). *E*

A. J. Lambell: Experimental information-storage filter.
Proc. Instn. Electr. Engrs. **114**, 1185-1192, 1967 (No. 9). *M*

H. de Lang & G. Bouwhuis: Measurements on the anomalous non-linear preference for circular mode polarization in a 1.523 μ He-Ne laser.
Physics Letters **25A**, 406-407, 1967 (No. 5). *E*

B. Lersmacher, H. Lydtin & W. F. Knippenberg: Zur Technologie der pyrolytischen Graphit-Herstellung.
Chemie-Ing.-Technik **39**, 833-842, 1967 (No. 14). *A, E*

F. E. Maranzana: Contributions to the theory of the anomalous Hall effect in ferro- and antiferromagnetic materials.
Phys. Rev. **160**, 421-429, 1967 (No. 2). *E*

P. Massini & G. Voorn: The effect of ferredoxin and ferrous ion on the chlorophyll sensitized photoreduction of dinitrophenol.
Photochem. Photobiol. **6**, 851-856, 1967 (No. 11). *E*

R. Memming & G. Schwandt: Potential and charge distribution at semiconductor-electrolyte interfaces.
Angew. Chemie: Int. Edit. in English **6**, 851-861, 1967 (No. 10); German Edit. **79**, 833-844, 1967 (No. 19). *H*

J.-P. Morel: Etude par l'effet Mössbauer de l'aimantation spontanée des deux sous-réseaux de fer dans le ferrite $NiFe_2O_4$.
J. Phys. Chem. Solids **28**, 629-634, 1967 (No. 4). *E*

W. C. Nieuwpoort & G. Blasse: Het ionogene bindingsmodel.
Chem. Weekblad **63**, 497-501, 1967 (No. 44). *E*

J. M. Noothoven van Goor: Hall coefficients of tellurium-doped bismuth.
Physics Letters **25A**, 442-443, 1967 (No. 6). *E*

A. van Oostrom: Adsorption of nitrogen on single-crystal faces of tungsten.
J. chem. Phys. **47**, 761-769, 1967 (No. 2). *E*

A. van Oostrom: Totaldruckmessung nach dem Ionisationsprinzip und ihre Störeffekte.
Vakuum-Technik **16**, 159-167, 1967 (No. 7). *E*

F. A. Staas, A. K. Niessen & W. F. Druyvesteyn: Some experiments on the distribution of a transport current in sheets of superconductors of the second kind.
Philips Res. Repts. **22**, 445-462, 1967 (No. 5). *E*

W. L. Wanmaker & D. Radielović: Luminescentie van fosfaten.
Chem. Weekblad **63**, 486-494, 1967 (No. 43).

J. te Winkel & B. C. Bouma: An investigation into transistor cross-modulation at VHF under AGC conditions.
IEEE Trans. **ED-14**, 374-381, 1967 (No. 7). *E*

H. J. de Wit & C. Crevecoeur: *n*-type Hall effect in MnO.
Physics Letters **25A**, 393-394, 1967 (No. 5). *E*

L. E. Zegers: The reduction of systematic jitter in a transmission chain with digital regenerators.
IEEE Trans. **COM-15**, 542-551, 1967 (No. 4). *E*

Volume 29, 1968, No. 2 pages 37-68 Published 14th February

PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 3/4

The RAMP inertial navigation system

F. Hector

Ever since travel began man has been faced with the problem of finding his location. Simple alignment with respect to known landmarks such as those of a coastline cannot be used in darkness and fog, and is useless for crossing the oceans. Many attempts have been made to find a navigation system which is reliable under all circumstances. At the beginning of this century navigation at sea was based upon sextant, magnetic compass and chronometer. In the first decades of this century the gyro-compass and radio-navigation systems were introduced. Still more recent are sophisticated navigation systems based on the application of inertial forces and making use of gyros and accelerometers. In this paper the RAMP system, a new inertial navigation system, is described. It uses a pendulum which indicates the change in the direction of the vertical in a moving vehicle.

Introduction

The problem of navigation is as old as man. In the beginning landmarks were used both for land and sea navigation, a method which was hampered by visibility limitations in darkness and fog. Later, navigation was carried out by observation of the sun, the moon and the stars. Although this method enabled mariners to make fairly accurate measurements of *latitude*, the ascertainment of *longitude* remained extremely difficult for a long time. In 1637 the Government of the United Netherlands awarded a prize to Galileo Galilei for his proposal of a navigation system. This was based on observations of the eclipses of the satellites of Jupiter, and the times of these eclipses for the meridian of Venice were given in tables. The system was ingenious, but complicated. The situation became easier when Christiaan Huygens and others invented the chronometer. From the 13th century onwards the magnetic compass was used for determining *direction*. Better results still were obtained with the gyro-compass which first came into use at the beginning of the 20th century.

Radio navigation systems, introduced at the end of the 'twenties, solved many of the problems of visibility and long-range navigation. Although their contribution to the reliability and safety of shipping and air traffic is unquestionable, some of the drawbacks of these systems are nowadays becoming more and more perceptible. Radio navigation is possible only in regions where accurate and reliable measurements of directions to radio transmitters are feasible or where landmarks can be observed by radar. Again, this is not the case for the greater part of the distance ships and aircraft have to cover when crossing the oceans, and there can also be difficulties in war-time when transmitters can be jammed. Higher speeds of travel have also resulted in a heavy demand for navigation techniques which operate automatically. For this purpose more sophisticated systems have been developed, such as Doppler radar and automatic star trackers, but even these have been found inadequate for the aircraft which are now being planned for travel at supersonic speeds.

During the last two decades navigation methods have been evolved which make use of inertial forces and for this reason have been termed "*inertial navigation*". These systems have the remarkable feature of being independent of any external source of reference (except for gravity) and are therefore independent of the vagaries of the weather and of enemy interference in war time.

This paper presents a new inertial navigation system, developed by the Swedish Philips Co., Ltd, called the RAMP system (RAMP standing for Rate and Acceleration Measuring Pendulum). It was developed primarily as a navigator for airborne vehicles. The general specification was written for speeds up to 1000

F. Hector, M. A. Physics, is with the Swedish Philips Co., Ltd, Stockholm.

nautical miles [1] per hour, corresponding to supersonic conditions (Mach 2). Another requirement was that it could also be used during long-duration flights of about 9 or 10 hours with an accuracy corresponding to arrival within the terminal area, i.e. within 15 or 20 nautical miles, after a transoceanic flight.

Before describing RAMP we shall deal with the fundamental principles of inertial navigation and consider the most important characteristics of pendulums and gyroscopes.

The principle of inertial navigation

Most inertial navigation systems are based on measurement and double integration of the linear acceleration of the vehicle. In order to be able to determine the displacement in a horizontal plane, it is necessary to measure the acceleration in two perpendicular horizontal directions. For this purpose two accelerometers are mounted in a frame, generally called a platform, which is kept in a horizontal position. This means that for a vehicle moving on or parallel to the Earth's surface, the platform has to be kept perpendicular to the local vertical (it should therefore follow the curvature of the Earth), or a correction should be made for the deviation.

The accelerometers have to meet very severe requirements for their horizontal alignment. If they are tilted with respect to the horizontal plane, they record an "acceleration" due to gravity which, if no correction is made, is added to the measured horizontal acceleration. As the effect of gravity is continuous and gives a much greater acceleration than the motion of an aircraft (except for take-off and very sharp turns), this misalignment may cause considerable errors. If, for instance, an accelerometer senses 0.1 % of the gravity field, double integration of the corresponding acceleration results in a position error which, after one minute, will amount to 18 m, and after ten minutes to 1800 m. For longer time intervals the errors would be catastrophic.

The local direction of the vertical can be indicated by means of a plumb-bob. This, however, is only possible under the condition of no motion. Linear acceleration and circular motions of the pivot will influence the direction of the plumb-bob. In 1923 M. Schuler [2] proved theoretically that at the surface of the Earth a pendulum device with an oscillation time of 84.4 min will maintain a vertical indication independent of motion. As this is the principle on which the RAMP system is based, we shall return to it later.

Measurement of the change of the vertical

If the direction of the local vertical is ascertained by means of a "Schuler tuned" pendulum and the initial direction of the vertical (at the starting point of the vehicle) can be stored, *the accelerometers mentioned above are no longer necessary*. It is evident that the change of position can be found from the change of the direction of the vertical (*fig. 1*) [3]. Since a pendulum suspended from an axial pivot has only one degree of freedom, full position information should be obtained by measuring the change of the vertical in two orthogonal planes. Corrections for the rotation of the Earth have to be applied, and also the appropriate transformation from great circles to latitude and longitude.

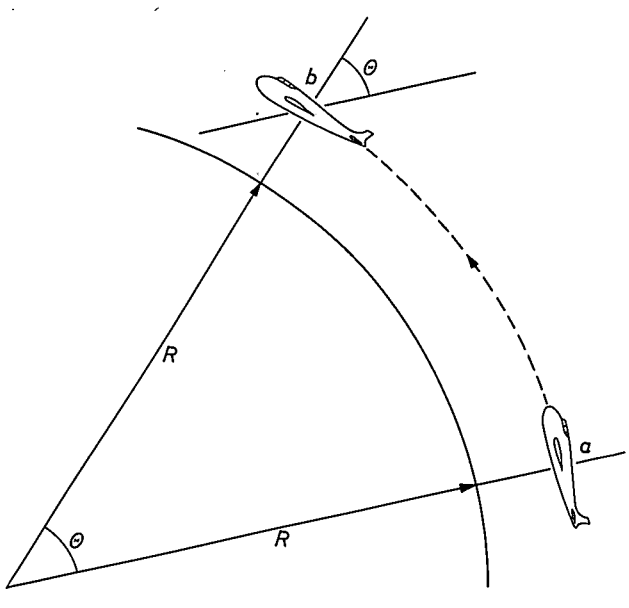


Fig. 1. The change of position of a vehicle moving from *a* to *b* can be found from the change Θ of direction of the local vertical. *R* is the radius of the Earth. The RAMP system is based on this principle.

The vertical deviates from the radial direction because the gravity field is caused not only by the attraction force, but also by the centrifugal force of the rotating Earth. The angle between the vertical and the equatorial plane is known as the *astronomical latitude* (*fig. 2*). For position indication, however, the *geographical latitude* is required; this is the angle between the equatorial plane and the normal to the reference ellipsoid of the Earth. The difference between the two latitudes is called the (local) deflection of the vertical. This deflection varies, but its magnitude seldom exceeds 10 seconds of arc. In this presentation it will be neglected.

Pendulums

A *physical* (or *compound*) *pendulum* is a rigid body with an axial pivot above its centre of gravity. Its oscillation time for small deviations is:

$$T = 2\pi \sqrt{\frac{I}{mgh}}, \quad \ldots \quad (1)$$

where m is the mass of the body, I its moment of inertia

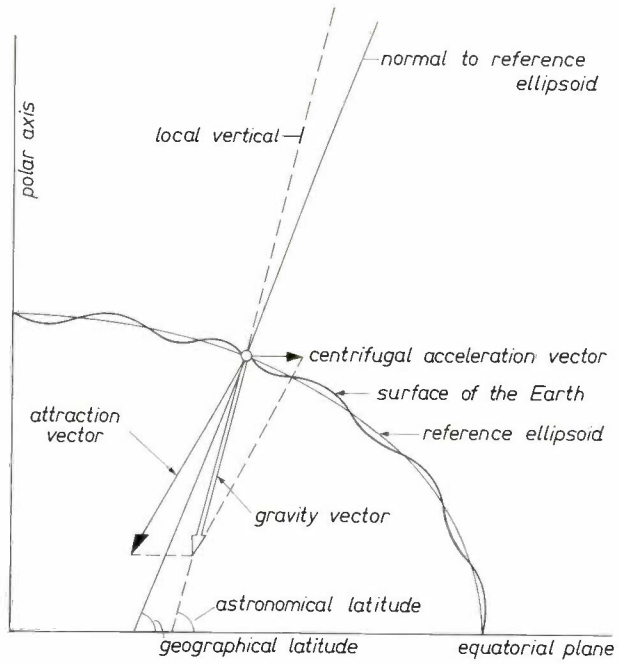


Fig. 2. Relationship between the gravity vector and its components, and geographical and astronomical latitudes.

about the axis of suspension, h is the distance between the pivot and the centre of gravity and g is the acceleration due to gravity.

A special case is the *mathematical* (or *simple*) *pendulum*. This consists of a particle of mass m, suspended from a fixed point by a weightless inextensible string. In this case the distance h between the pivot and the centre of gravity is equal to the length of the string. Therefore $I = mh^2$, and the oscillation time is:

$$T = 2\pi \sqrt{\frac{h}{g}}. \qquad (2)$$

A pendulum which is not in motion with respect to the Earth's surface shows the direction of the local vertical. It can, in fact, be considered as a plumb-bob. When the pivot is part of a vehicle performing an accelerated horizontal motion the direction will deviate from the vertical. However, as we have already noted, a "Schuler pendulum", with an oscillation time of 84.4 minutes, maintains a vertical indication, independent of the motion of the point of suspension. On the other hand, as Schuler himself emphasized, a purely mechanical pendulum fulfilling this requirement cannot be made. A mathematical pendulum of the oscillation time quoted would have a length h equal to the Earth's radius, whereas for a physical pendulum with a feasible mass, the distance between pivot and centre of gravity would be a few nanometers, which is of course quite unrealistic.

This can be shown by considering a pendulum moving in a meridian plane over the surface of the Earth (*fig. 3*). (The Earth is taken to be a perfect sphere.) This pendulum will continue to coincide with a vertical line if the angular acceleration of the pendulum equals the angular acceleration $\ddot{\Theta}$ of the radial line. Under this condition the linear acceleration of the pendulum is $R\ddot{\Theta}$ and therefore the torque acting on the pendulum body, due to this acceleration, is $mhR\ddot{\Theta}$. To obtain an angular acceleration $\ddot{\Theta}$ this torque should be equal to $I\ddot{\Theta}$. The moment of inertia should therefore be:

$$I = mhR,$$

and according to (1) the oscillation time is:

$$T = 2\pi \sqrt{\frac{R}{g}}.$$

A comparison with equation (2) shows that a mathematical pendulum with a length equal to the radius of the Earth would meet the requirement. Putting $R = 6.37 \times 10^6$ m and $g = 9.81$ m/s^2, we get $T \approx 84.4$ min. Taking some realistic values for the physical pendulum, $m = 1$ kg and $I = 10^{-2}$ kg m^2, we obtain $h \approx 1.6 \times 10^{-9}$ m.

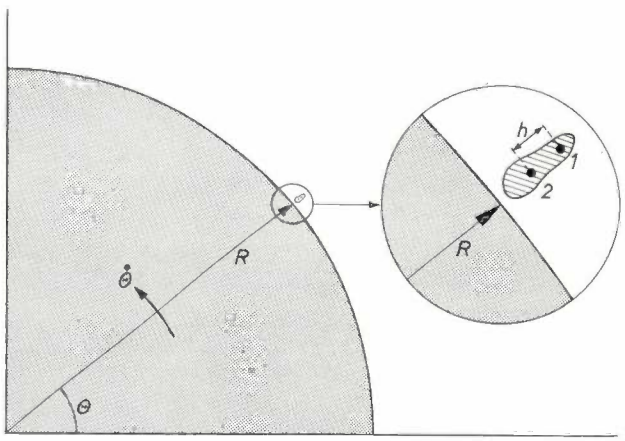


Fig. 3. Indication of the local vertical with a physical pendulum. R Earth's radius; Θ latitude; *1* pivot; *2* centre of gravity; h distance between *1* and *2*.

Although a purely mechanical realization of a Schuler pendulum is impossible, solutions using servotechniques can meet the requirement of a long oscillation time. One of these solutions will now be described.

Electrical feedback

In *fig. 4* a pendulum is shown which is equipped with an electrical feedback system. Attached to the shaft PA there is a generator SG_p, delivering a signal which is fed after amplification to an electrodynamic magnet system TG_p, called a *torque generator* (or *torquer*). This exerts on the shaft a signal which depends on the applied signal e_T.

[1] As usual in navigation the distances in this article are given in nautical miles. 1 nautical mile = 1851.9 m.

[2] M. Schuler, Die Störung von Pendel- und Kreiselapparaten durch die Beschleunigung des Fahrzeuges, Phys. Z. **24**, 344-350, 1923.

[3] This method is fundamentally the same as navigation by celestial observations.

A pendulum with an electrical feedback system as shown is the principle underlying a type of accelerometer called the *rebalanced pendulous accelerometer*. The generator SG_p in this case supplies a signal proportional to the angular deviation of the pendulum from the vertical. If the amplification is sufficiently large, any deviation from this position is then reduced nearly to zero by the torque generator: the feedback system acts as an "artificial spring". The period of oscillation in this case is very small. When the pendulum is accelerated in a horizontal plane, in a direction perpendicular to the shaft PA, the torque-generator voltage e_T required to maintain the vertical position ("rebalancing") is a direct indication of the acceleration a. Neglecting transient phenomena, a follows from the equation:

$$e_T S_T = mha,$$

giving:

$$a = \frac{S_T}{mh} e_T,$$

where S_T is the sensitivity of the torque generator (newton-metres per volt).

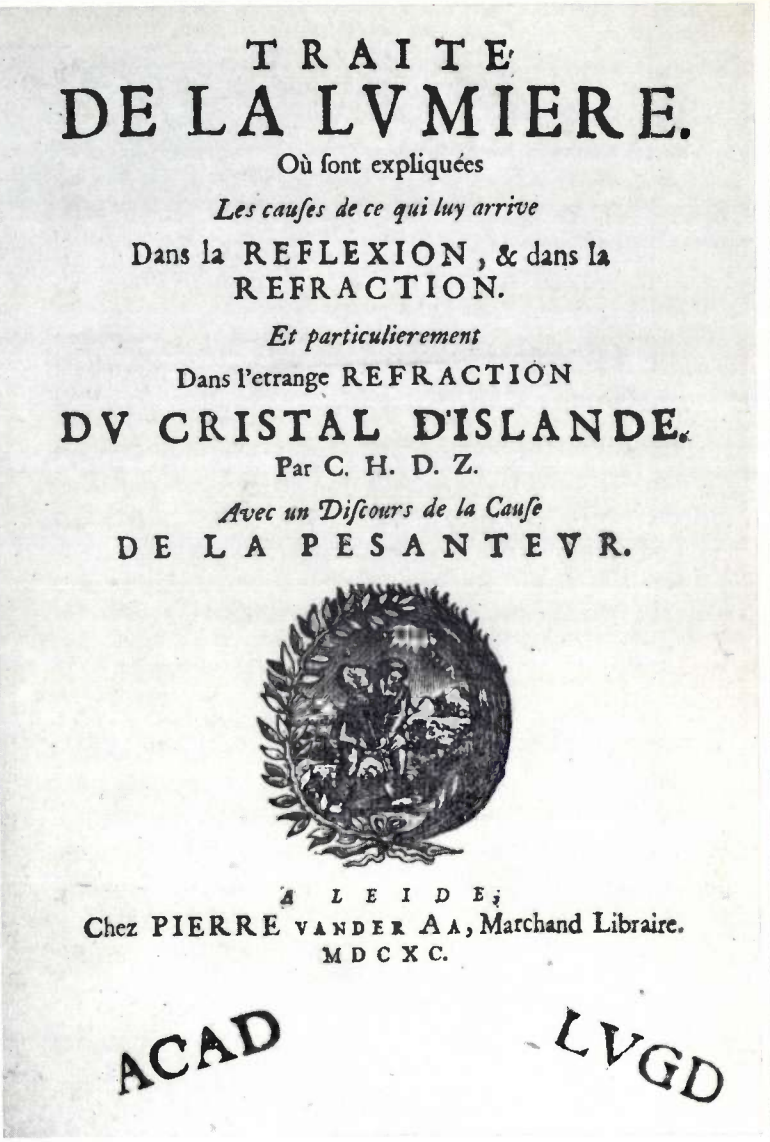
Instead of a generator that delivers a signal proportional to the angular *deviation*, a type may be used that supplies a signal proportional to the angular *acceleration*. In this case the feedback brings about an apparent increase of the moment of inertia of the pendulum, which can be shown in the following way.

The equation of motion of the pendulum is now:

$$I\ddot{\Theta} + mhg\Theta + S_G A S_T \ddot{\Theta} = 0,$$

where Θ is the deviation from the vertical, S_G is the sensitivity of the signal generator (volts per unit of angular acceleration), A the amplification of the amplifier and S_T is again the sensitivity of the torque generator. From this equation we get the oscillation time:

$$T = 2\pi \sqrt{\frac{I + B}{mgh}}. \qquad (3)$$


TRAITE'
DE LA LVMIERE.
Où sont expliquées
Les causes de ce qui luy arrive
Dans la REFLEXION, & dans la REFRACTION.
Et particulierement
Dans l'etrange REFRACTION
DV CRISTAL D'ISLANDE.
Par C. H. D. Z.
Avec un Discours de la Cause
DE LA PESANTEVR.
A LEIDE;
Chez PIERRE VANDER AA, Marchand Libraire.
MDCXC.
ACAD
LVGD

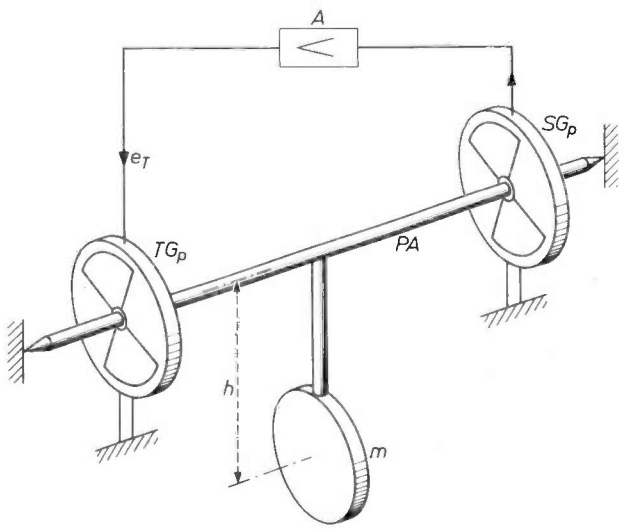
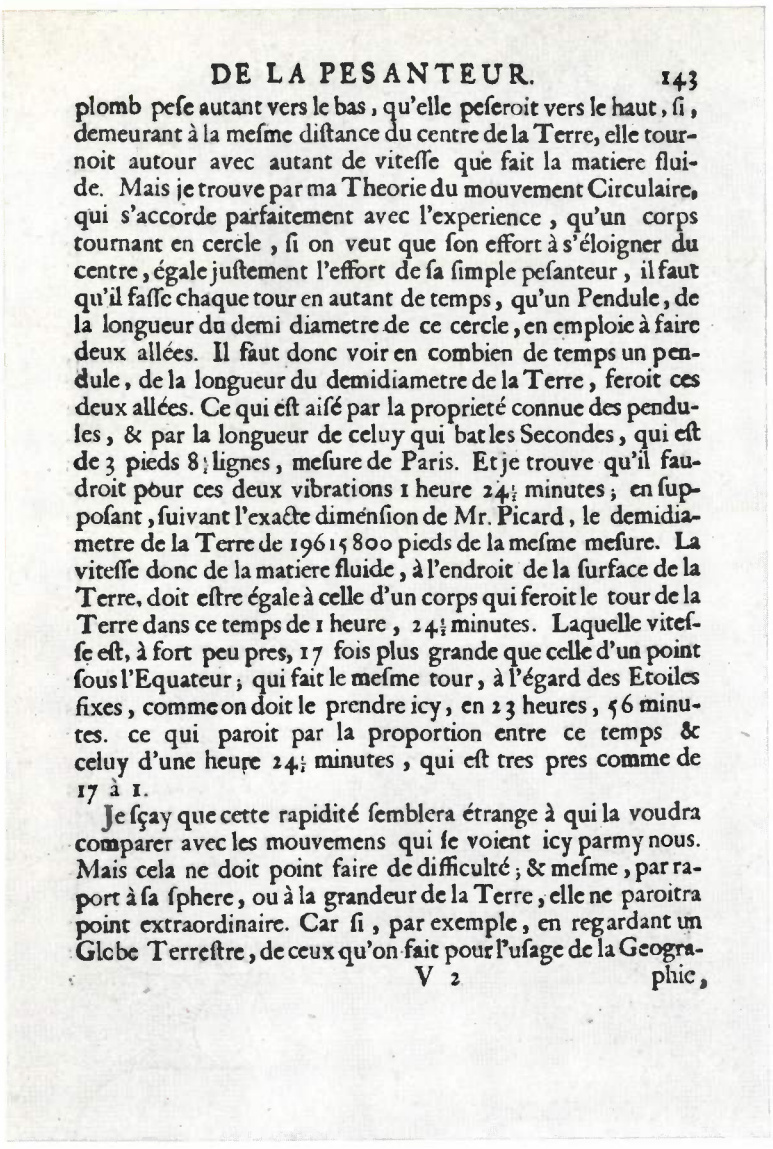


Fig. 4. Schematic diagram of a pendulum with electrical feedback. PA pendulum shaft. m pendulum mass. h length of the pendulum. SG_p signal generator. A amplifier. TG_p torque generator. e_T signal fed to TG_p. If SG_p delivers a signal which is proportional to the angle of rotation, the device will operate as a rebalanced pendulous accelerometer, with "zero" deflection (and a very high natural frequency). However, if SG_p delivers a signal which is proportional to the angular acceleration, the oscillation time compared to that of the free pendulum will be increased, giving in one special case the "Schuler period".



DE LA PESANTEUR. 143

plomb peſe autant vers le bas, qu'elle peſeroit vers le haut, ſi, demeurant à la meſme diſtance du centre de la Terre, elle tournoit autour avec autant de viteſſe que fait la matiere fluide. Mais je trouve par ma Theorie du mouvement Circulaire, qui s'accorde parfaitement avec l'experience, qu'un corps tournant en cercle, ſi on veut que ſon effort à s'éloigner du centre, égale juſtement l'effort de ſa ſimple peſanteur, il faut qu'il faſſe chaque tour en autant de temps, qu'un Pendule, de la longueur du demi diametre de ce cercle, en emploie à faire deux allées. Il faut donc voir en combien de temps un pendule, de la longueur du demidiametre de la Terre, feroit ces deux allées. Ce qui eſt aiſé par la proprieté connue des pendules, & par la longueur de celuy qui bat les Secondes, qui eſt de 3 pieds 8½ lignes, meſure de Paris. Et je trouve qu'il faudroit pour ces deux vibrations 1 heure 24½ minutes; en ſuppoſant, ſuivant l'exacte diménſion de Mr. Picard, le demidiametre de la Terre de 19615800 pieds de la meſme meſure. La viteſſe donc de la matiere fluide, à l'endroit de la ſurface de la Terre, doit eſtre égale à celle d'un corps qui feroit le tour de la Terre dans ce temps de 1 heure, 24½ minutes. Laquelle viteſſe eſt, à fort peu pres, 17 fois plus grande que celle d'un point ſous l'Equateur; qui fait le meſme tour, à l'égard des Etoiles fixes, comme on doit le prendre icy, en 23 heures, 56 minutes. ce qui paroit par la proportion entre ce temps & celuy d'une heure 24½ minutes, qui eſt tres pres comme de 17 à 1.

Je ſçay que cette rapidité ſemblera étrange à qui la voudra comparer avec les mouvemens qui ſe voient icy parmy nous. Mais cela ne doit point faire de difficulté; & meſme, par raport à ſa ſphere, ou à la grandeur de la Terre, elle ne paroitra point extraordinaire. Car ſi, par exemple, en regardant un Globe Terreſtre, de ceux qu'on fait pour l'uſage de la Geogra-

V 2 phie,

The "Schuler period" of 84.4 minutes, which plays such a prominent part in F. Hector's article in this issue, is there shown to be equal to the oscillation time of a simple pendulum whose length is equal to the radius of the Earth. This oscillation time had already been calculated by Christiaan Huygens in the seventeenth century — witness the page reproduced here from his treatise "Discours de la Cause de la Pesanteur" published at Leyden in 1690.

Huygens was interested in the value of this oscillation time because it is equal — as he had shown — to the orbital period of an object moving near the surface of the Earth at a speed such that the centrifugal force balances the gravitational force. This is the well-known condition which holds for Sputnik and the innumerable subsequent satellites that now circle our planet and which, when orbiting at a height of no more than about 200 km (just outside the region of atmospheric friction), have in fact an orbital period of about 1½ hours. It may come as a surprise to many who have followed these satellite developments that this basic orbital period for an object circling the Earth had already been calculated correctly and quite accurately in the seventeenth century.

Huygens can hardly have dreamed of artificial satellites, nor was he concerned with the navigational significance of the Schuler period. He made these particular calculations in an effort to support a curious theory of gravity. This theory (the "vortex theory") was devised by Descartes in an attempt to do away with the assumption of an *action at a distance*, on which Newton's theory of gravity was based and which many scientists of those times found difficult to accept.

---

where $B = S_G A S_T$ . . . . . (4)

is the apparent increase of the moment of inertia. By making B sufficiently large, a pendulum can be made with an oscillation time equal to the "Schuler period" of 84.4 min (about 5000 s). If the period of the physical pendulum proper has a reasonable value, e.g. 0.5 s, this can be done by an apparent increase of the moment of inertia by a factor of 10^8.

The signal generator SG_p, which is required to deliver a signal proportional to the angular acceleration, will be discussed later. It may be noted, however, that a generator could be used which delivers a signal proportional to the angular *velocity* (angular *rate*) or to the angular *displacement* (as noted above). In these cases, in order to obtain the required artificial increase of the moment of inertia, the signal should undergo a single or double differentiation.

Gyroscopes

The main part of a gyroscope is a wheel spinning at a very high angular velocity ω. The angular velocity about axes perpendicular to the "spin axis" must be very low compared to ω. The wheel is mounted on a shaft which is supported by a set of gimbals. By applying Newton's laws of motion the fundamental characteristic of the gyro can be explained, i.e. its reaction against any change of direction of its spin axis. Therefore, if a spinning gyro-wheel is supported by a set of "frictionless" gimbals (*fig. 5*), with axes going through the centre of the wheel, it will maintain its direction in space, irrespective of movements of the supporting frame.

It also follows from Newton's laws that a torque about one of the gimbal axes results in a rotation about

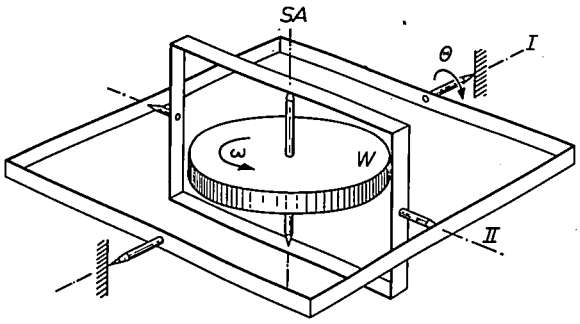


Fig. 5. Schematic diagram of a free gyroscope. *SA* spin axis of the wheel *W*, which has an angular velocity ω. *I* and *II* are the gimbal axes.

the other axis. This phenomenon is called *precession*. The directions of rotation and torque are such that the spin momentum of the gyrowheel attempts to align itself with the torque. The basic equation underlying the precession is:

$$T_0 = H\dot{\Theta}. \quad \ldots\ldots \quad (5)$$

Here H is the spin momentum of the wheel, that is, its moment of inertia I_w multiplied by its angular velocity ω:

$$H = I_w\omega. \quad \ldots\ldots \quad (6)$$

Furthermore Θ is the angular displacement about one of the gimbal axes (e.g. the axis *I*; $\dot{\Theta}$ is thus the angular velocity about this axis), and T_0 is the torque about the other axis (*II*).

Gyro with spring restraint

An angular velocity of the gyro about one of the gimbal axes can be measured by determining the torque about the other axis. This can be done by applying a linear elastic restraint, e.g. a set of springs (*fig. 6*). The angular displacement φ about the latter axis is now proportional to the torque T_0:

$$T_0 = K\varphi, \quad \ldots\ldots \quad (7)$$

where K is the spring constant. The gyro now has only one degree of freedom; the corresponding gimbal axis is generally called the *input axis* (*IA*), the other axis, where the torque is measured, being the *output axis* (*OA*). From (5) and (7) it follows that:

$$\varphi = \frac{H}{K}\dot{\Theta}, \quad \ldots\ldots \quad (8)$$

and the angular displacement about the output axis is thus proportional to the angular velocity (rate) about the input axis. This rate therefore can be read from a calibrated scale about the output axis. A gyro equipped with a spring restraint about one of its axes is therefore called a *rate gyro*.

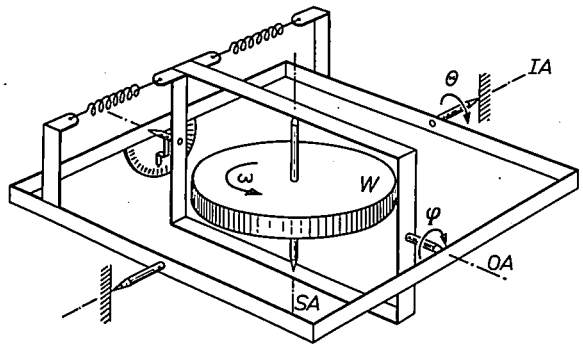


Fig. 6. Schematic diagram of a gyro with a spring restraint about one of the axes. The deviation φ about this axis (the output axis *OA*), which can be read from the scale, is now proportional to the angular *velocity* $\dot{\Theta}$ of the gyro-wheel about the other axis (the input axis *IA*). This device is known as a rate gyro.

In some types of gyro a restraint of another kind is used. The gyro-gimbal is then cylindrical in shape and immersed in a liquid. The torque given by this viscous restraint is proportional to the angular velocity about the appropriate axis. If now the output axis has a viscous restraint, the torque which accompanies an angular velocity about the input axis must be caused by an angular velocity about the output axis. It follows from (5) that in this case these velocities are proportional, so that the total displacement angle about the output axis is proportional to the integral of the angular velocity (rate) about the input axis. For this reason a gyro equipped with a viscous restraint is called a rate-integrating gyro, or simply an *integrating gyro*.

Gyro with electrical feedback

Instead of a spring restraint a gyro can be equipped with an electrical feedback system, consisting of a signal generator, an amplifier and a torque generator. *Fig. 7* shows a gyro which has been provided with such a feedback system about the output axis. The generator SG_0 is required to deliver a signal proportional to the angular displacement φ about this axis. Therefore

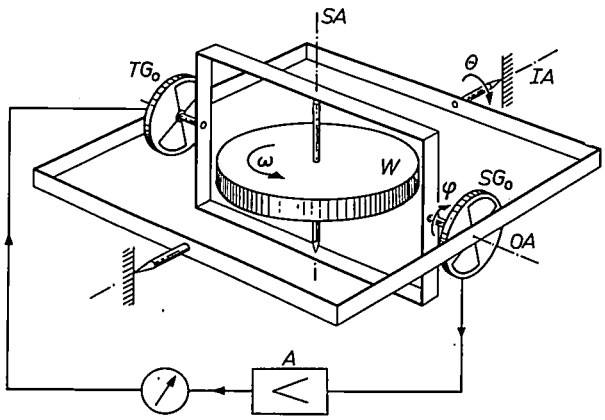


Fig. 7. Schematic diagram of a rate gyro with an electrical feedback system. SG_0 generator, delivering a signal proportional to the rotation angle φ about the output axis *OA*. *A* amplifier. TG_0 torque generator. In this case also φ is proportional to the angular velocity $\dot{\Theta}$ about the input axis *IA*.

the torque due to the torque generator TG_o is also proportional to this angle and the feedback system can be considered as an "artificial spring", the "spring constant" being:

$$K' = S_{Go} A S_{To}, \quad \ldots\ldots \quad (9)$$

where S_{Go} is the sensitivity of the signal generator (in volts per radian), A is the amplification of the amplifier, and S_{To} the sensitivity of the torque generator. The gyro in fig. 7 will have the same characteristics as the one shown in fig. 6: the rotation angle φ about the output axis is proportional to the angular velocity $\dot{\Theta}$ about the input axis. In fig. 7 a scale about the output axis is not needed; the reading can be obtained by measuring the signal fed to the torque generator. If the torque exerted is proportional to this signal, the reading on the signal meter is proportional to the angular velocity about the input axis.

If the electrical circuit is modified a signal can be obtained which is proportional to the angular *acceleration* about the input axis. For this purpose an integrator is inserted in the feedback loop (*fig. 8*). In this case the torque exerted by the torque generator is:

$$T_o = S_{Go} A \frac{1}{\tau} S_{To} \int \varphi \, dt, \quad \ldots\ldots \quad (10)$$

where τ is the time constant of the integrator. From (5) and (10) it follows that:

$$\varphi = \frac{H\tau}{S_{Go} A S_{To}} \ddot{\Theta}. \quad \ldots\ldots \quad (11)$$

Thus the angular displacement about the output axis is proportional to the angular acceleration about the input axis. The signal from the generator can be used as an indication; after amplification this can be read from the meter shown in the figure.

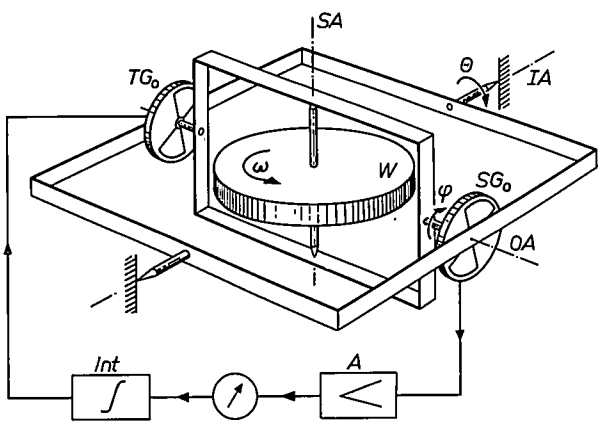


Fig. 8. By inserting an integrator *Int* in the feedback loop the angle of rotation φ about the output axis OA, and thus also the signal read on the meter, becomes proportional to the angular *acceleration* $\ddot{\Theta}$ about the input axis IA.

The RAMP, a combination of a pendulum and a gyro

Since a gyro with an integrating feedback system for one of its axes delivers a signal proportional to the angular acceleration about the other axis, it can be used as a signal generator for artificially increasing the moment of inertia of a pendulum, in accordance with the principle illustrated in fig. 4. A diagram of this device is shown in *fig. 9*. The outer gimbal of the gyro forms the pendulum and is therefore loaded with a mass m. The outer gimbal axis (input axis) serves as pendulum axis, PA. The outer gimbal is equipped with an integrating

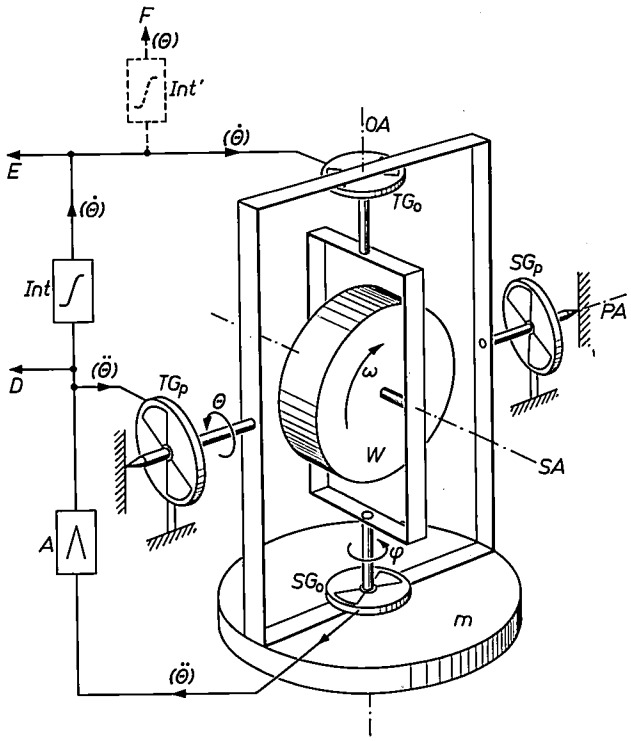


Fig. 9. Schematic diagram of a Rate and Acceleration Measuring Pendulum (RAMP). The pendulum is formed by the outer gimbal of the gyro, loaded with the mass m: its angle of rotation is Θ. The output axis OA of the gyro has an integrating feedback loop as in fig. 8. The signals D, E and F are proportional to $\ddot{\Theta}$, $\dot{\Theta}$ and Θ respectively (indicated by the bracketed terms), i.e. to the acceleration, the velocity and the displacement of the vehicle. SG_p signal generator for the servomotor of the slaved platform.

feedback loop, as in fig. 8. Apart from being used for the feedback of the gyro, the amplified signal D from the generator SG_o, which is proportional to the angular acceleration $\ddot{\Theta}$ about the pendulum axis, is now also used for controlling the torque generator TG_p for this axis. From (4) and (11) it follows that the artificial increase of the moment of inertia of the pendulum in this case is:

$$B = \frac{S_{Tp}}{S_{To}} H\tau, \quad \ldots\ldots \quad (12)$$

where S_{Tp} and S_{To} are the sensitivities of the torque gen-

erators TG_p and TG_o respectively. By making B sufficiently large, the pendulum can be "Schuler tuned": *it indicates the vertical, independently of movements of the pivots.* When mounted in a vehicle moving over the Earth's surface in a direction perpendicular to the pendulum axis PA, the pendulum changes its direction in space, because of the curvature of the Earth's surface. The angular velocity $\dot{\Theta}$ is proportional to the horizontal velocity of the vehicle ($v = R\dot{\Theta}$) and, as indicated in (5), this velocity is also proportional to the torque exerted by the torque generator TG_0. The signal E applied to the torque generator therefore indicates the vehicle's horizontal velocity in the direction perpendicular to the pendulum axis.

As the signals E and D, delivered by the device are proportional to the *velocity* (*rate*) and the *acceleration*, this combination of a pendulum and a gyro is called RAMP (Rate and Acceleration Measuring Pendulum). By the addition of another integrator, *Int'*, a signal F is obtained which is proportional to Θ, and hence to the *displacement* of the vehicle in the direction mentioned. We have thus determined the displacement from the change of direction of the vertical *without storing*

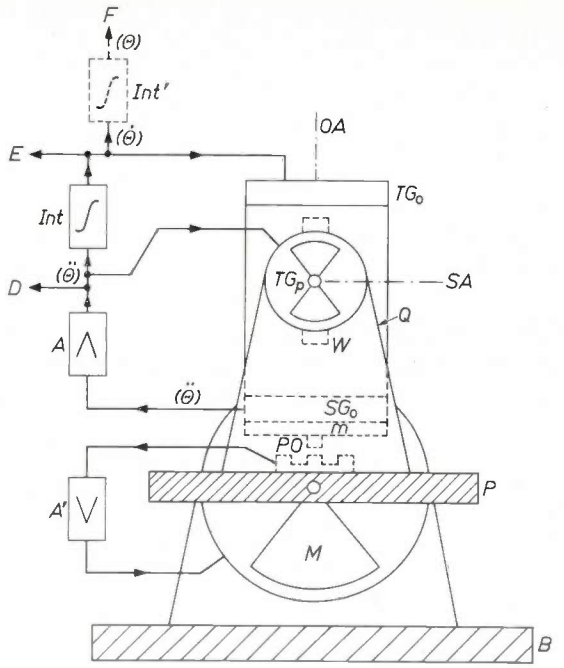


Fig. 10. Layout of a RAMP with slaved platform. The pendulum axis (input axis of the gyro) is perpendicular to the plane of the figure. B base. The pendulum is carried on a frame Q on a platform P, which follows the pendulum with the aid of the signal generator (pick-off) PO, the amplifier A' and the servomotor M.

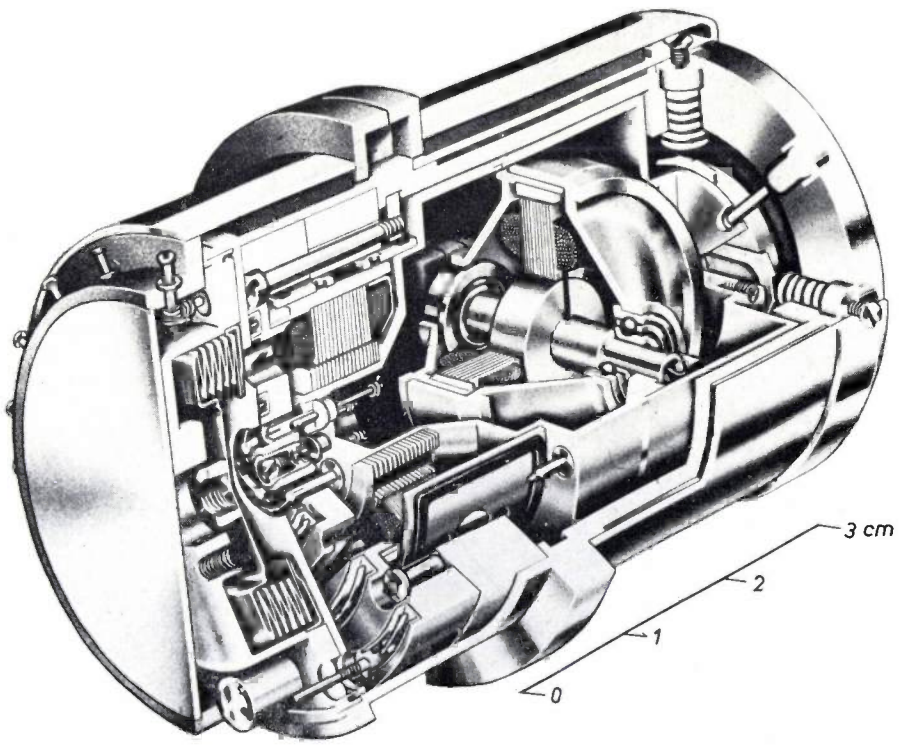


Fig. 11. Cut-away picture of a floating gyro with a single degree of freedom.

the direction of the vertical at the starting point of the vehicle. We measure directly the *change* of direction.

In an equipment the RAMP is mounted on a "platform", which is not rigidly fixed to the vehicle, but follows the movements of the pendulum (it is "slaved" to the pendulum). The pendulum thus only has to make very small movements with respect to this platform, which simplifies the construction. The plat-

form is therefore mounted in gimbals and equipped with a servomotor. This motor is controlled by the amplified signal from a generator SG_p, shown on the input axis in fig. 9, which delivers a signal proportional to the deviation between the directions of the pendulum and the platform.

An approach to a realistic arrangement is shown in *fig. 10*. The signal generator SG_p for controlling the platform has been designed as a differential transformer called a "pick-off" (*PO*). An important characteristic of this arrangement is that the closed feedback loops can be achieved without external connections via slip-rings. By the use of micro-electronics the corresponding electronic equipment can be built into the RAMP unit itself.

The RAMP pendulum bearing is a "compensated spring" bearing and is therefore "stictionfree", thus improving the gyro environment in comparison with a conventional platform. In the present design the gyro is a miniature floating type [4]. The wheel and the spin-axis assembly are enclosed in a sealed cylindrical can, which is immersed in a liquid whose density is equal to the effective density of the can. The angular momentum of the wheel is 10^5 gcm^{2}s^{-1}. A photograph of such a floating gyro is shown in *fig. 11*, and *fig. 12* shows a complete RAMP.

As the pendulum and gyro form one unit, assembly and replacement in the platform is very simple [5]. The alignment of the RAMP is further simplified by the fact that the dimensions are so chosen that clean-room conditions are not necessary.

The dynamics and the stability of the RAMP with its slaved platform can be studied in a signal-flow diagram. The transfer functions are then given by their Laplace transforms. *Fig. 13* (on page 78) shows such a diagram, but we shall not go into further detail here.

The general layout

As navigation requires the determination of displacement in two perpendicular directions, *two* RAMPs, the *X*-RAMP and the *Y*-RAMP, are mounted on a platform (an "inertial platform") which is slaved to both of them. As a reference for the *heading* of the aircraft, a directional gyro, the *Z*-gyro, is also mounted on the platform. (The input axis of this gyro points vertically downwards.) A diagram of the arrangement is shown in *fig. 14*. In practical use the *x*-direction, which corresponds to the direction of the pendulum

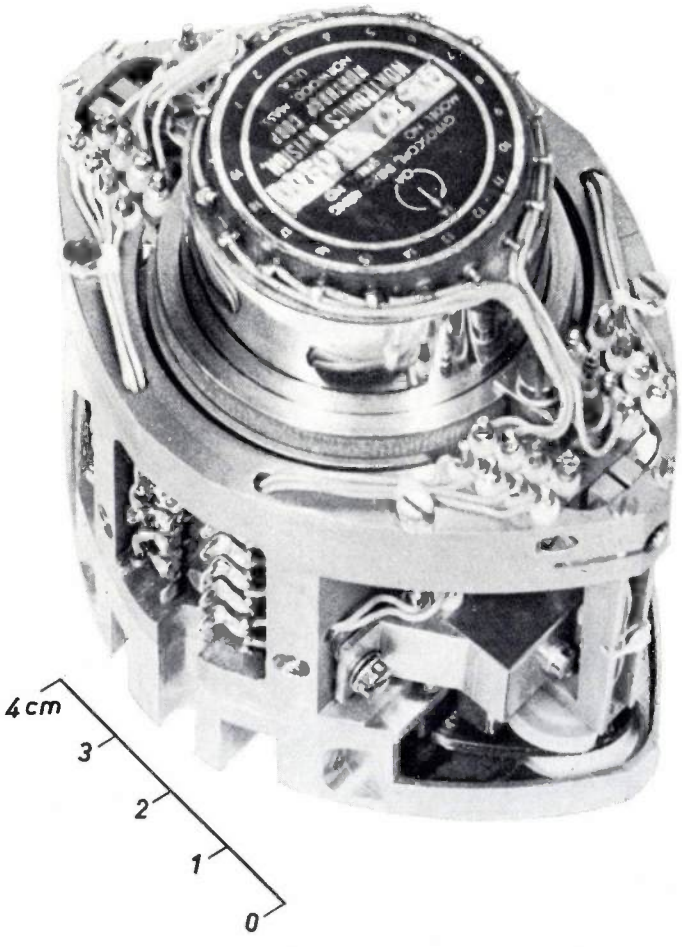


Fig. 12. Complete pendulum for the RAMP navigation system.

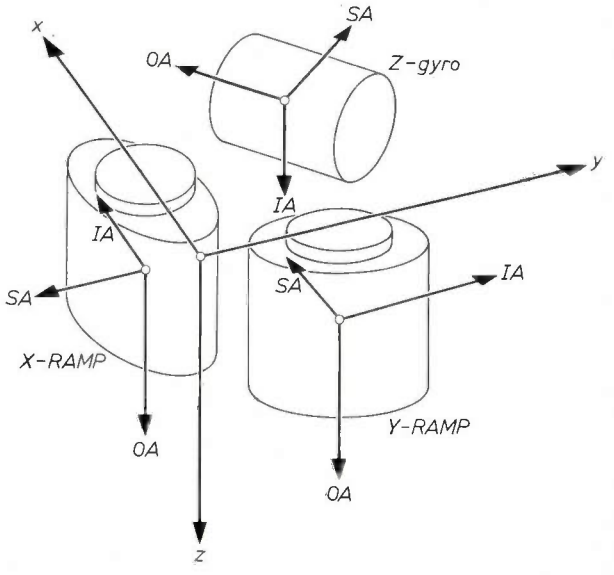


Fig. 14. Arrangement of two RAMP units and a gyro as mounted on the platform in a RAMP navigation system. *SA* spin axes. *IA* input axes. *OA* output axes. The input axes correspond to the co-ordinate system *x*, *y*, *z* (cf. fig. 26).

[4] Introduced by C. S. Draper at the Massachusetts Institute of Technology.

[5] In a conventional inertial navigation system both the accelerometers and the gyros have to be carefully aligned.

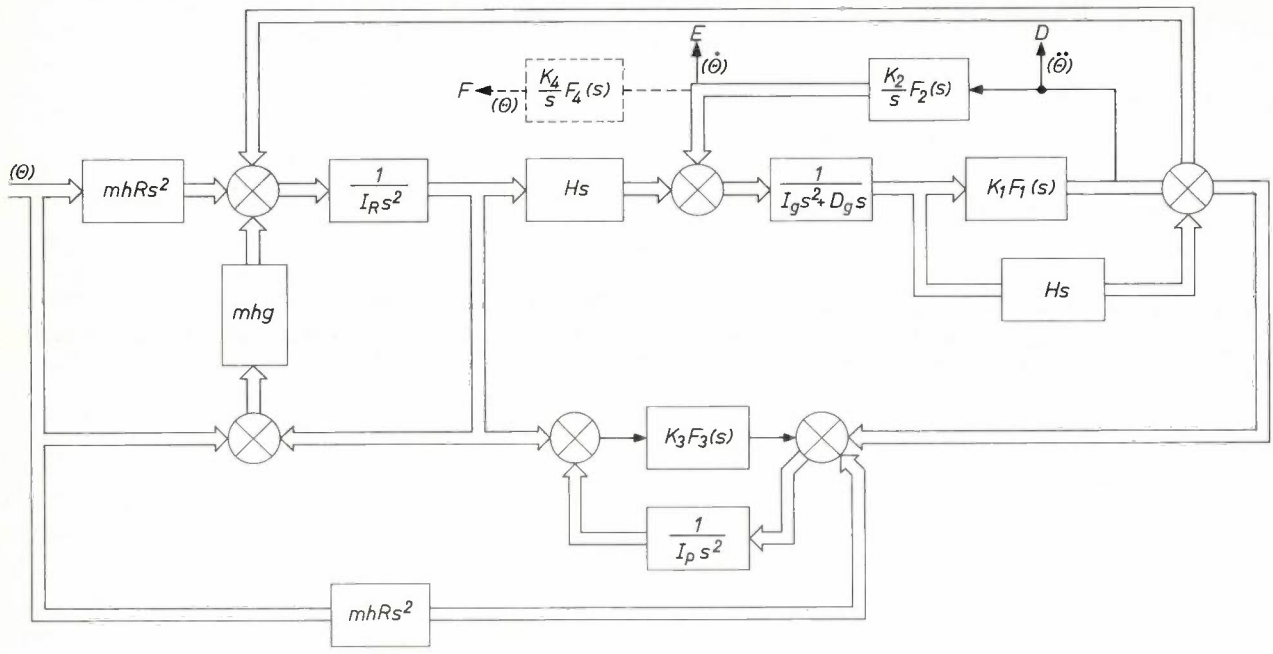


Fig. 13. Signal flow diagram in a RAMP with a slaved platform. The transfer functions are indicated by their Laplace transforms. s Laplace operator. *mh* mass unbalance of the pendulum. *g* gravity acceleration. *R* radius of the Earth. I_R moment of inertia of the RAMP pendulum. *H* gyro angular momentum. I_g moment of inertia of gyro gimbal. D_g gyro viscous damping constant. I_p moment of inertia of slaved platform. $K_1 \ldots K_4$ constants. $F_1 \ldots F_4$ Laplace transforms of transfer functions of correction networks. *D*, *E*, *F* output signals (cf. fig. 9).

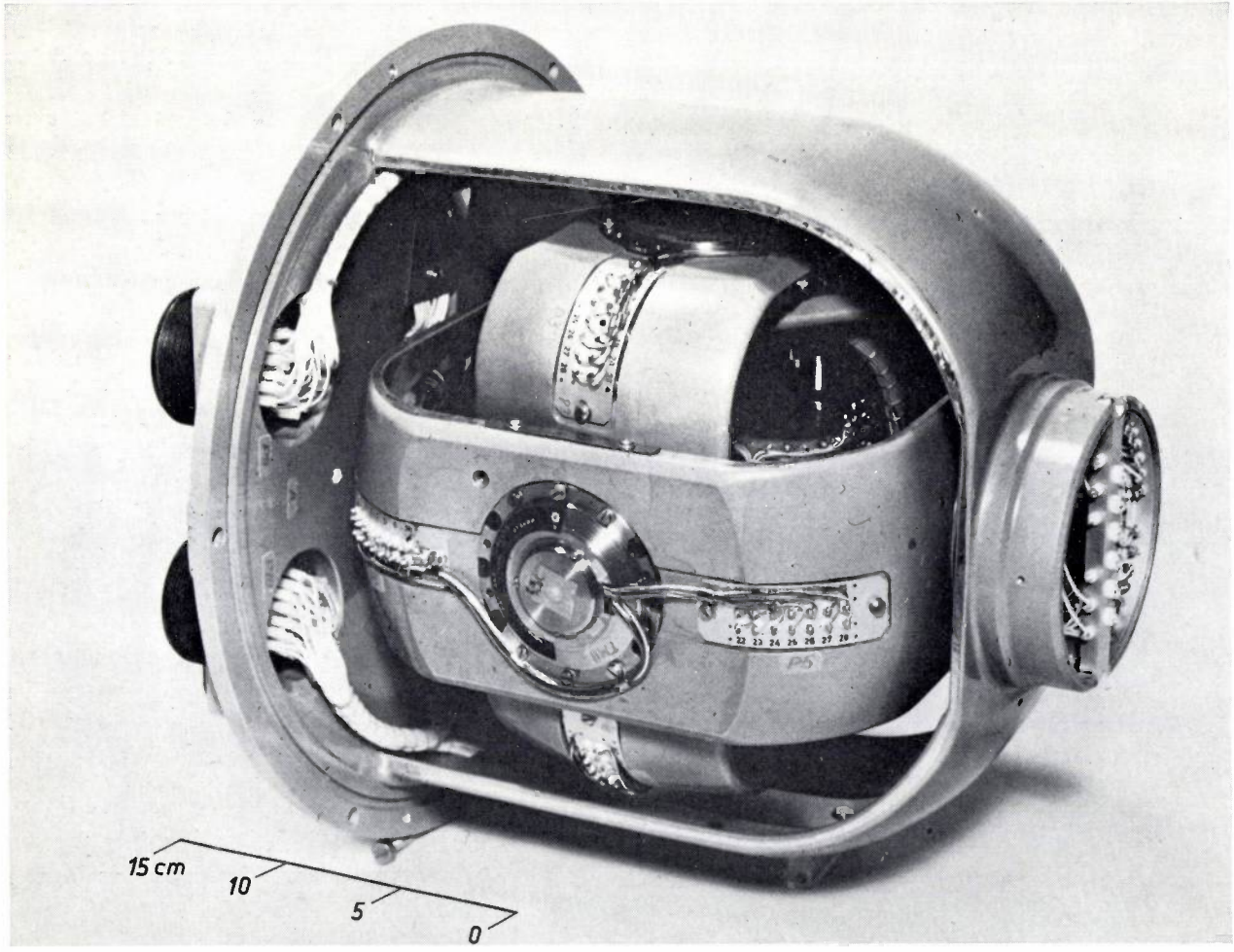


Fig. 15. The complete inertial platform, with two RAMP units and a directional gyro.

axis of the *X*-RAMP, points north and the *y*-direction (pendulum axis of the *Y*-RAMP) points east (cf. fig. 26).

Complete freedom of the platform with respect to the vehicle can be obtained by a three-gimbal system. However, if the attitude of the vehicle with respect to the platform is such that one gimbal axis coincides with one of the other axes, one degree of angular freedom is lost. This is known as "gimbal-lock". It can be avoided by introducing a fourth gimbal, which is equipped with a torque generator, controlled by a signal from a generator on the second gimbal. This is a known technique and will not be explained here.

A complete RAMP platform is shown in *fig. 15*. A block diagram of the complete navigation equipment is given in *fig. 16*. The output signals from the RAMP platform are proportional to the acceleration and the velocity in two perpendicular directions (*D* and *E* in

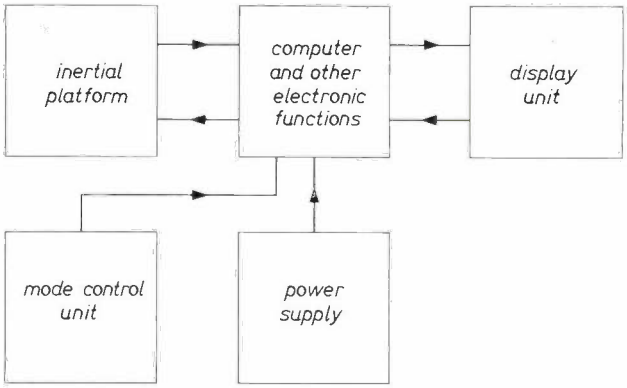


Fig. 16. Block diagram of the RAMP navigation equipment.

figs. 9, 10 and 13); signals giving information about the attitude of the aircraft (heading, roll and pitch) are also obtained. These signals are fed to the *computer and electronic unit*. This unit comprises an analogue computer which covers a number of functions: co-ordinate transformation (great-circle angular velocities to latitude and longitude velocities), integration of velocity to position (in the integrator *Int'*, see figs. 9 and 10), correction-term computations (cf. Appendix) and computations of the velocity vector from its two components. This unit also comprises signal converters and circuits for bias-setting for the Earth's rotation and compensation of the gyro drift.

The computer and electronic unit is connected to the *display unit*, which presents latitude, longitude, heading, roll and pitch. A photograph of this unit is shown in *fig. 17*. It is equipped with two controls for setting initial latitude and longitude before the flight.

The *mode control unit* serves for the switching of several functions. It contains switches for "Stand-by", "Navigation" and "Off", and controls to compensate for the Earth's rotation and gyro drift.

Preparation of the equipment for use starts with switching on the heating of the gyros to bring the fluid in which the floating gimbals are immersed to the right density for operation; the heating time is about 20 or 30 minutes. In the meantime the RAMPs are roughly aligned to enable the platform to find the local vertical direction. For this purpose the electronic tuning is switched to a pendulum period of 3 minutes and afterwards to a period of 8.4 minutes. Through heavy damping which is switched on during this period the platform now finds the vertical. When the amplitude of the oscillations is sufficiently small, the Schuler period of 84.4 minutes is switched on.

By rotating the platform around its vertical axis from the rough initial alignment (where the *y*-axis is pointing east) to the opposite direction (*y*-axis pointing west), the gyro drift of the *Y*-RAMP and true East are established. The gyro drift of the *X*-RAMP is measured by rotating the platform through 90°. The *Z*-gyro drift is measured when the platform is accurately aligned.

The *power supply* converts the aircraft mains voltage, 115 V, 400 Hz, three-phase, to stabilized d.c. voltages and accurate a.c. voltages. The gyro power is supplied by a 26 V, 400 Hz, three-phase generator with a fre-

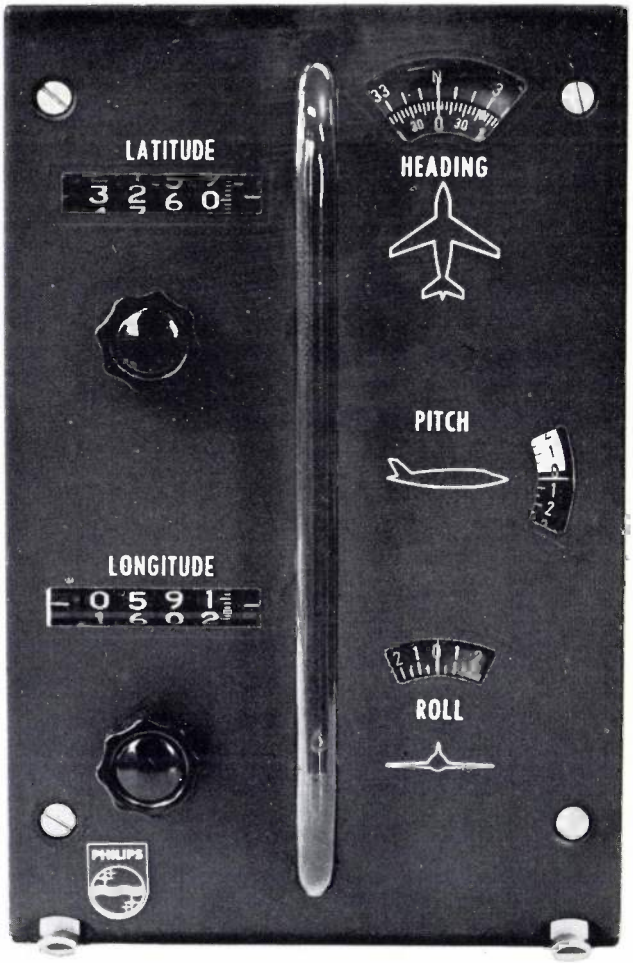


Fig. 17. The display unit, which indicates latitude, longitude, heading, roll and pitch of the aircraft.

quency precision of 10^{-5} and a total distortion and amplitude error of about 10^{-3}.

Analogue computer investigations

A mathematical analysis of the complete RAMP system is given in the Appendix. One of the principal results of this analysis is that the introduction of correction terms effects a coupling between the *X*-RAMP and the *Y*-RAMP. This coupling induces two kinds of oscillations of the system; one with a period equal to the Earth's rotation time and one with the Schuler period (84.4 min). The latter oscillations are modulated; the modulation frequency corresponds to the speed of rotation of a Foucault pendulum at the corresponding latitude.

In order to check these characteristics of the system a model of the information flow, which is shown in fig. 27 in the Appendix, was made in an analogue computer. In this case the main object was to deduce the relation between the gyro drift and the errors in position and velocity. This was done by applying a step function for the gyro drift of 0.01°/h to each gyro individually. The corresponding position and attitude deviations are shown in *fig. 18*. As will be seen from these curves, the dominant errors have a 24-hour period while the errors having the Schuler period can hardly be seen. To give some idea of the corresponding position errors we note that 1 minute of arc longitude at a latitude of 60° corresponds to 0.5 nautical mile.

Laboratory tests

In order to investigate the behaviour of the inertial system over longer periods of operation it was run for periods of 12 and 24 hours in the laboratory. During these times normal drift tests of gyros and alignment were made as in real flight tests. The absence of shocks and accelerations did of course result in smoother curves and also in considerably smaller total errors.

Fig. 19 shows the positional errors, latitude and lon-

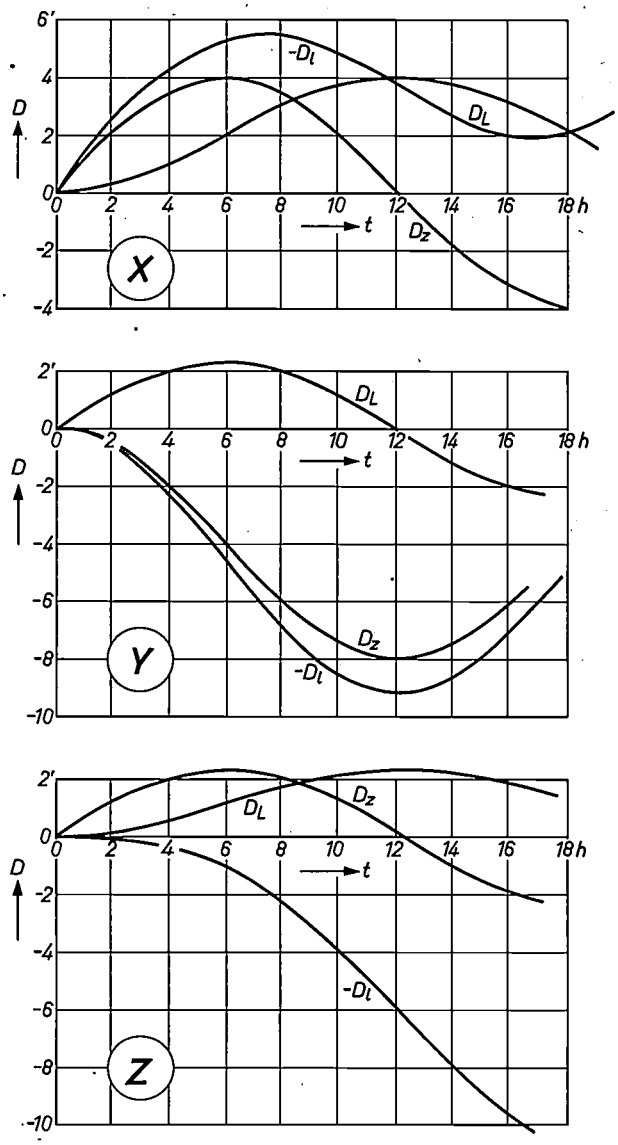


Fig. 18. Recordings of analogue computer investigations on the inertial navigation system. The recordings show the longitude error D_l, the latitude error D_L, and the direction error D_z, all resulting from the application of a step-function gyro drift of 0.01 °/h successively to the *X*-gyro, the *Y*-gyro and the *Z*-gyro, as a function of the time *t* (in hours). The roll and pitch errors were very small.

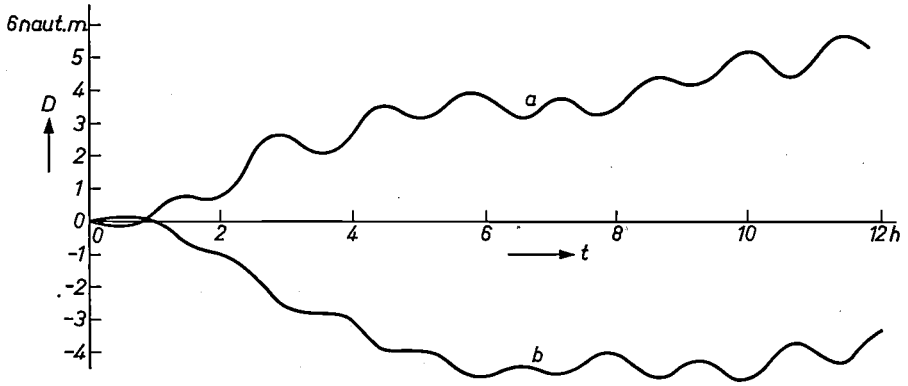


Fig. 19. Errors *D* in longitude (*a*) and latitude (*b*) during a 12-hour laboratory test. The curves show a half period of a 24-hour oscillation and Schuler oscillations (period 84.4 min). These oscillations are superimposed on a linear gyro drift, which amounts to about 0.5 nautical miles per hour in longitude and 0.3 nautical miles per hour in latitude.

gitude, obtained in a typical 12-hour test. The curves show Schuler oscillations and a 24-hour oscillation, superimposed on a linear gyro drift of about 0.5 naut. m.p.h. for the longitude and of about 0.3 naut.m.p.h. for the latitude.

Fig. 20 shows a recording of the velocity signals from the two RAMPs when an oscillation of about 1′ amplitude was initiated. It was found from the recording that very smooth Schuler oscillations remain, with a harmonic distortion as low as about 5%, which corresponds to second-order disturbances smaller than 0.01 °/hour. Furthermore the Foucault effect can be recognized as a modulation of the Schuler oscillations.

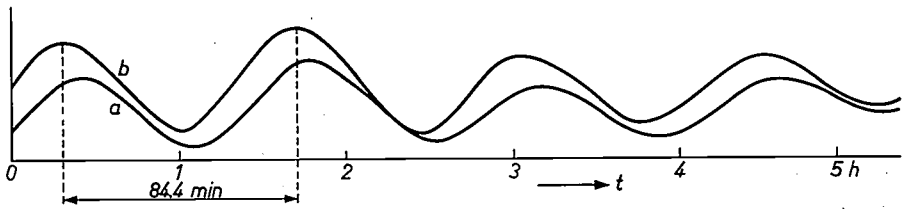


Fig. 20. Part of a recording of the velocity signals from the *X*-RAMP (*a*) and from the *Y*-RAMP (*b*) at an initiated oscillation of about 1′ amplitude during a 12-hour laboratory test.

Flight tests

Two flight test programmes were carried out. The first one was made in one of the twin-engine aircraft owned by Philips, a De Havilland Dove. The second programme was carried out in a DC8 aircraft, belonging to Scandinavian Airline Systems, during flights over the European continent and over the North Atlantic.

Test flights with the De Havilland Dove

The installation of the equipment is shown in *fig. 21*. Three different tracks were used: two East-West tracks and one North-South track. The pilot flew over the check points at a low altitude (50 to 100 metres), thus ensuring an accurate position indication. When he gave the signal, the position — latitude and longitude — indicated on the display was photographed or recorded manually together with the time.

Altogether 36 flights were made during this period. In the beginning the results varied due to teething troubles in the electronic and mechanical equipment. At the end of the flight test period, however, the results became fairly reproducible. The results of the flights number 32, 33 and 35 are shown in *fig. 22*. The average uncertainty amounted to 2-3 nautical miles per hour. This corresponds to the estimated accuracy, based on measured gyro drift.

Flight test in the DC8

The same equipment that was tested in the De Havilland Dove was mounted in a special rack, adapted to one of the seats in the lounge of the DC8. Before the flights (normal passenger flights) the equipment was warmed up and aligned in a service bus on the ground, after which, operating on its own battery, it was installed in the aircraft and connected to its mains supply. During the first flights, which were carried out over the European continent (Copenhagen-Teheran, 3000 nautical miles, flight time about 9 hours

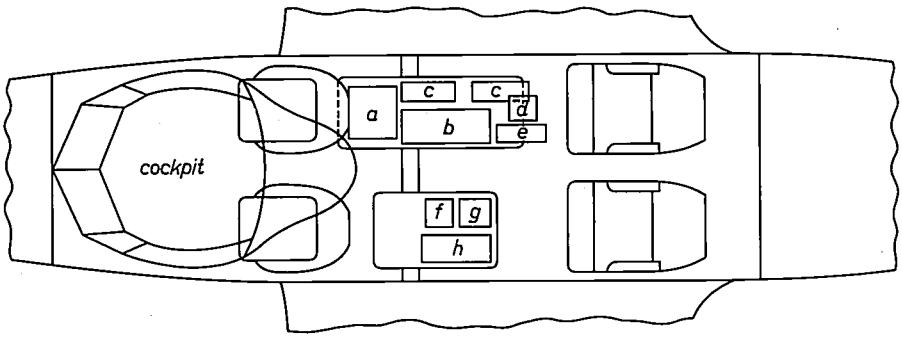


Fig. 21. Installation of the test equipment in the De Havilland Dove. *a* platform. *b* electronic units. *c* power converters. *d* mode control units. *e* display unit. *f* auxiliary equipment. *g* recorder. *h* oscilloscope.

with stops), the position was checked with the ordinary navigation system on board. Accurate check points were reported, particularly when passing over the radio beacons. Over the North Atlantic (Copenhagen-New York, 3000 nautical miles, flight time about 8.5 hours) the Loran C network was used which, at the beginning and at the end of the flights, offered a fairly good check. The radio beacons along the Scandinavian and American Atlantic coast were of course also used.

The results of these long-flight tests showed that the recordings from the flights over the Continent were smoother and more accurate. A typical recording is shown in *fig. 23*, where the maximum error corresponds to 3.5 naut.m.p.h. and the average to 2.2 naut.m.p.h. The recordings of the transatlantic flights are more disturbed, due probably to the checking method. The distribution of the longitude and latitude errors after six hours' flight time with the DC8 is shown in *fig. 24*. The equipment is shown fitted into the DC8 in *fig. 25*.

ments for inertial navigation equipment. Thus, the gyros were fire control components of high accuracy, improved so as to come close to inertial navigation accuracy. Although gyros were specified with a drift of less than 0.01-0.02 °/h, gyro drifts of several tenths of a degree per hour were often measured. For practical reasons an analogue computer constructed from electro-mechanical components and electronic amplifiers and networks was chosen. Important sources of inaccuracy were the lack of linearity and the uncontrolled drift of the integrators, both of which amounted to about 0.1 %. Even more serious errors come from the co-ordinate transforming process. Here it was rather difficult to maintain an accuracy of 0.1 %. Finally the mechanical unit of the pendulum system in the RAMP had some not unexpected teething troubles like thermal drift. The total error was expected to be of the order of a few nautical miles per hour and it was encouraging to note that the flight test results were very close to this expected accuracy. Introduction of better gyros, a digital computer and an improved mechanical design of the pendulum system would very probably lead to an accuracy of the order of 1 nautical mile per hour.

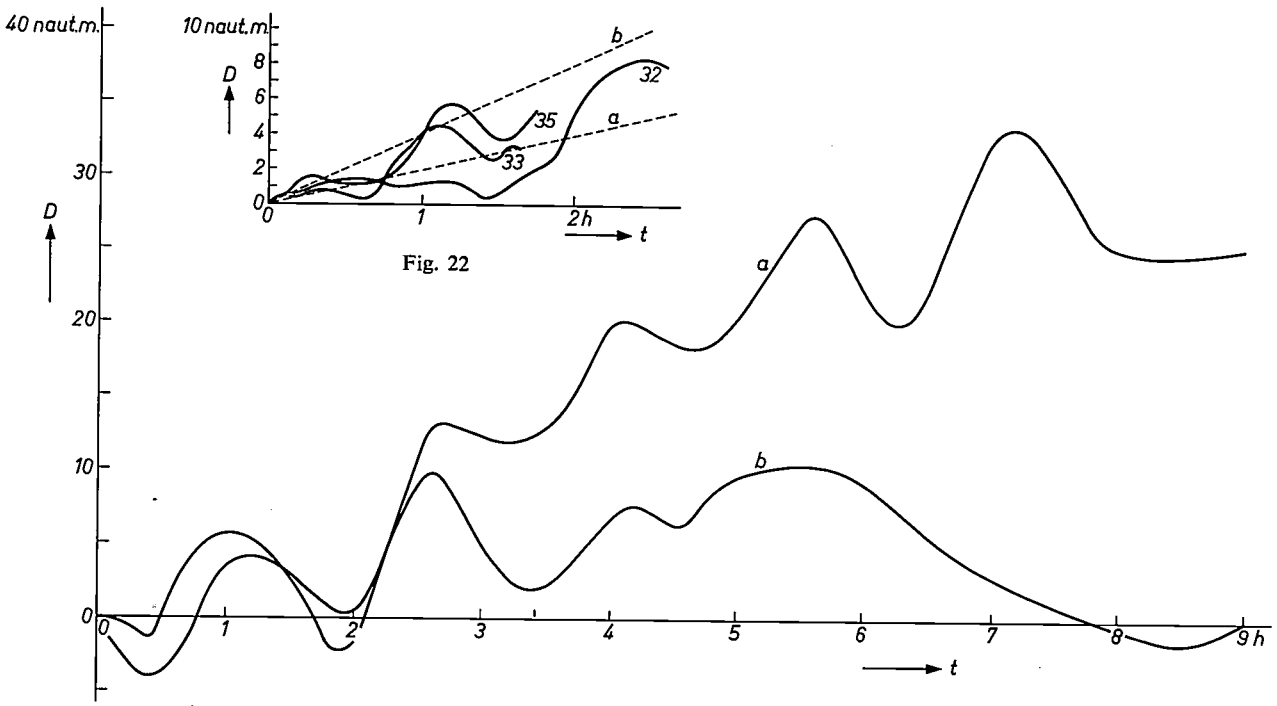


Fig. 22

Fig. 23

Fig. 22. Results obtained during three of the later test flights with the De Havilland Dove (flight numbers 32, 33 and 35). The solid lines show the position error D, in nautical miles, as a function of time. The dotted lines a and b correspond to errors of 2 and 4 naut.m.p.h.

Fig. 23. Recording of the longitude error (a) and the latitude error (b) in nautical miles as a function of time during a flight over the European continent.

Comments on the test results

When discussing the test results it should be pointed out that in order to lower the cost some components were used which did not fulfil the advanced require-

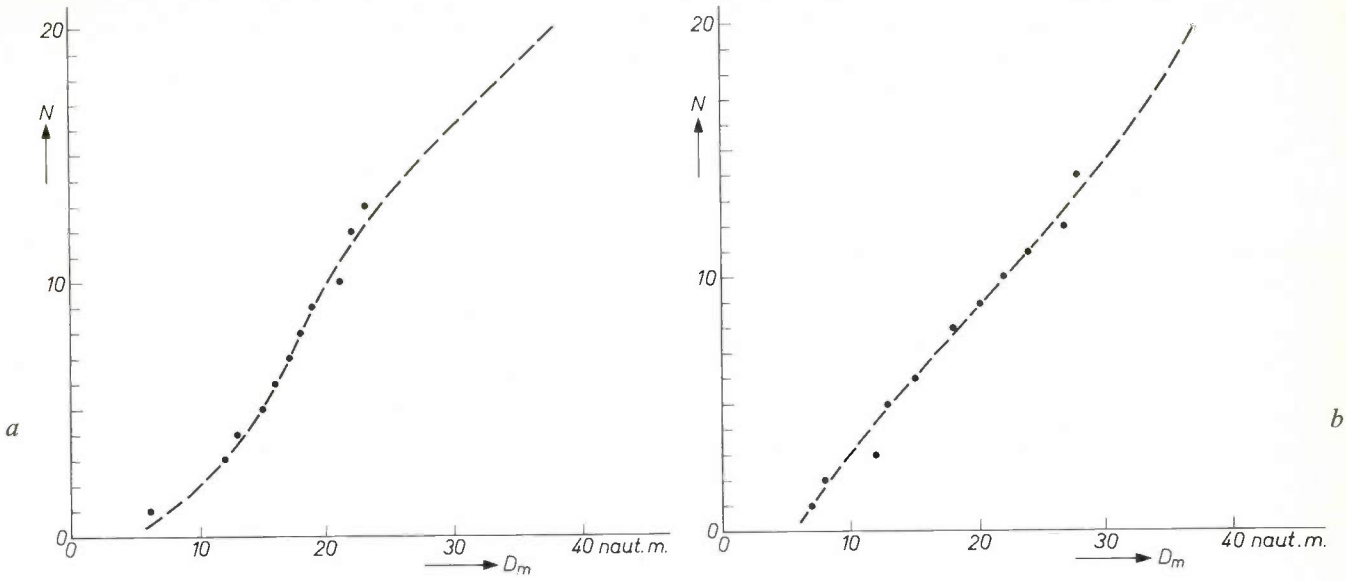


Fig. 24. Distribution of the errors in longitude (*a*) and in latitude (*b*) indicated by the RAMP equipment during six-hour flights over the European continent and over the Atlantic Ocean. The dots show the number of flights N with a maximum error less than D_m (in naut. miles). Errors larger than about 30 naut. miles occurred only in the beginning.

Appendix: Analysis of the complete RAMP navigation system

For an analysis of the complete navigation system we must know 1) the location on the Earth's surface of the vehicle in which the system is installed; 2) the forces, motions and their mathematical relations. Referring to *fig. 26* we establish a three-axis co-ordinate system, which is geographically orientated: the x-axis points north, the y-axis east and the z-axis downwards along the vertical. The co-ordinate system thus rotates with the Earth. The angular velocities imparted to the platform by this rotation and by the movement of the vehicle follow the equations:

$$\omega_{1x} = (\omega_{IE} + \dot{l}) \cos L, \quad (13)$$

$$\omega_{1y} = -\dot{L}, \quad (14)$$

$$\omega_{1z} = -(\omega_{IE} + \dot{l}) \sin L, \quad (15)$$

$$\text{(hence } \omega_{1z} = -\omega_{1x} \tan L), \quad (16)$$

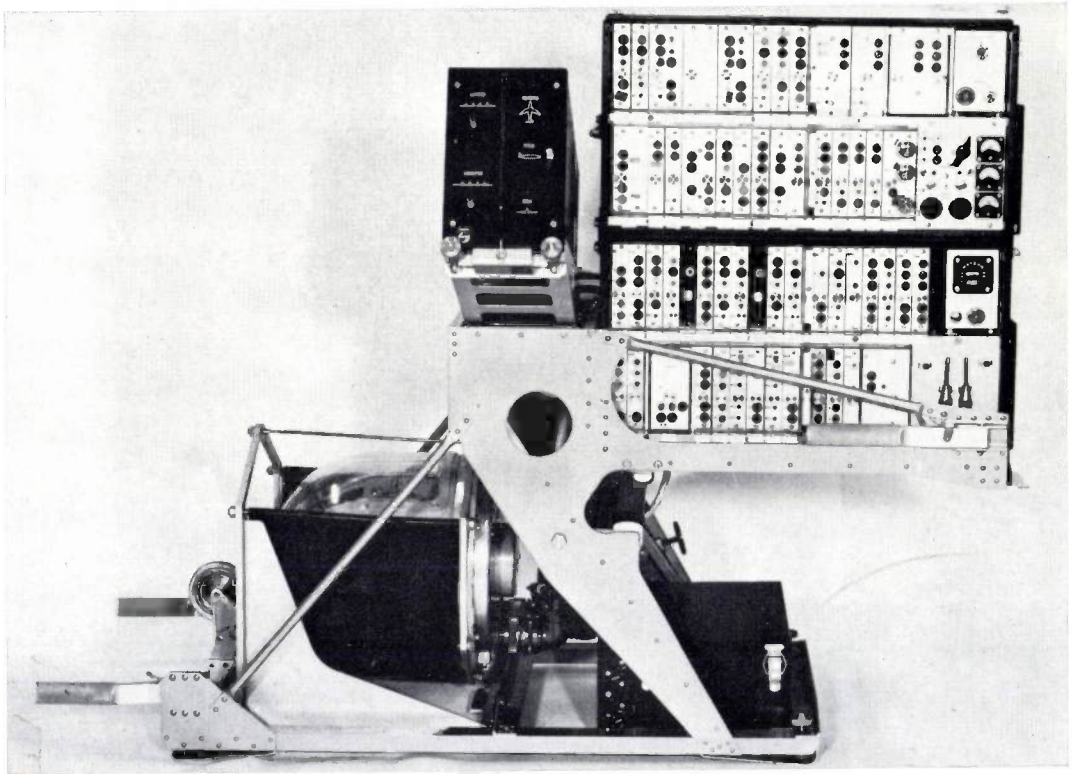


Fig. 25. Installation of the test equipment in the DC8.

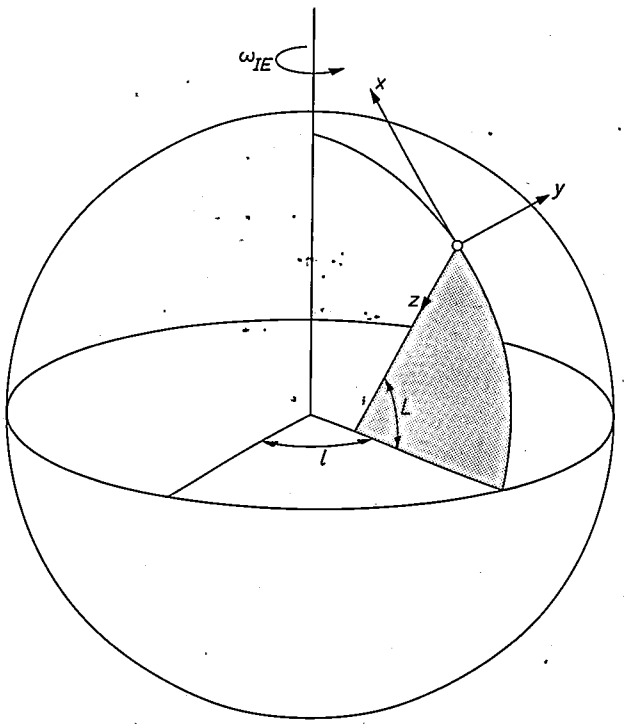


Fig. 26. Geographically orientated co-ordinate system x, y, z, pointing north, east and downwards. l longitude. L latitude. ω_{IE} rotation speed of the Earth.

where ω_{IE} is the frequency of the Earth's rotation, l is the longitude and L the latitude.

The mathematical derivation of the accelerations acting on the platform along the three axes gives an expression whose principal terms are:

$$a_{px} = -R(1 - 2e\cos^2 L)\dot{\omega}_{Iy} - 2\dot{R}\omega_{Iy} - \\ -R\omega_{Ix}\omega_{Iz} - e\ddot{R}\sin 2L + \ldots,$$

$$a_{py} = R\dot{\omega}_{Ix} + 2\dot{R}\omega_{Ix} - R\omega^2_{Ix} - R\omega_{Iy}\omega_{Iz} + \ldots,$$

$$a_{pz} = -\ddot{R} + R(\omega^2_{Ix} + \omega^2_{Iy}) + \ldots,$$

where R is the Earth's radius and e is the eccentricity of the reference ellipsoid of the Earth. These accelerations will generate torques about the pendulum axes. The torque equations for two single-axis pendulums each contain a dominant term:

$$M_{sx} = -mhR\dot{\omega}_{Ix},$$

$$M_{sy} = -mhR\dot{\omega}_{Iy}(1 - 2e\cos^2 L).$$

They also contain terms which are small compared with the main terms and therefore may be treated as correction terms; these are:

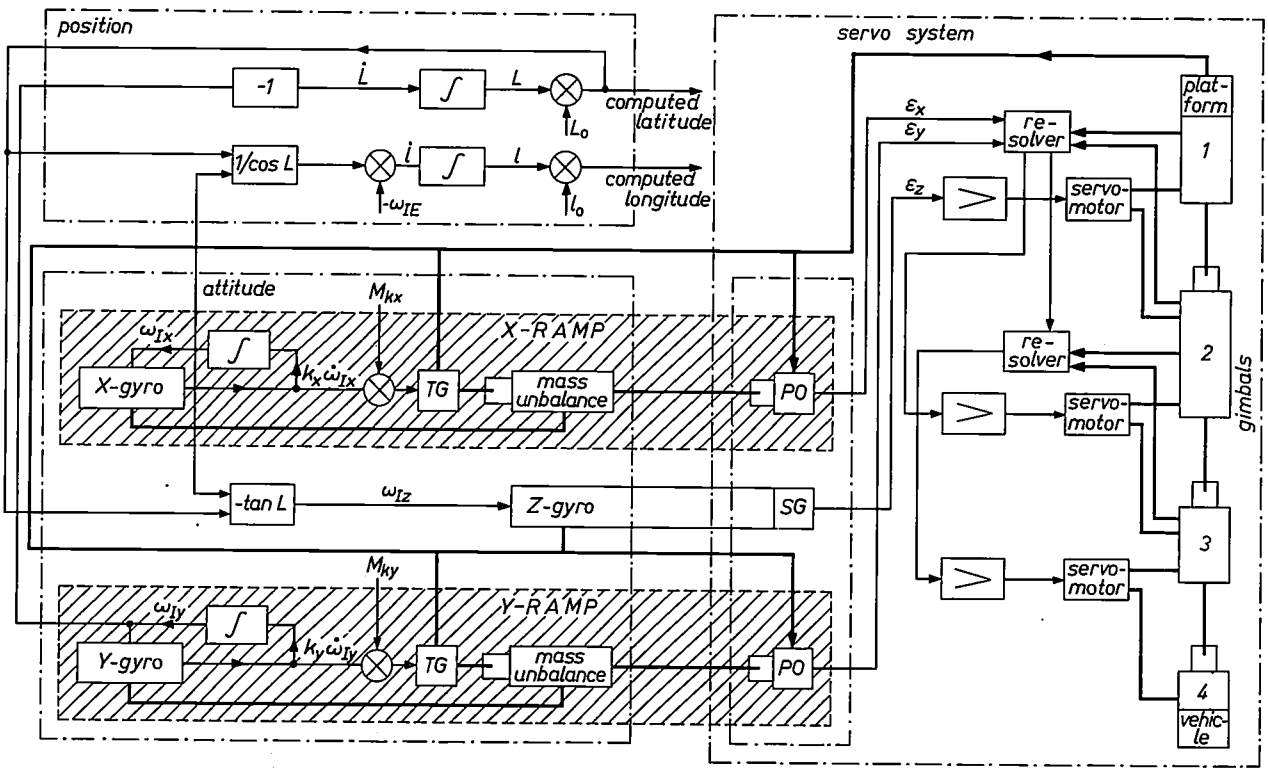


Fig. 27. Signal flow diagram of the complete RAMP navigation system. The bold lines correspond to mechanical connections, the thin lines to electrical connections.

The hatched blocks represent the X-RAMP and the Y-RAMP. In the block marked "position" the output signals of the RAMPs undergo the mathematical operations represented by the equations (13) and (14). After correction for the Earth's rotation, integration and introduction of the geographical latitude and longitude at the starting point of the flight, L_0 and l_0, the latitude and longitude of the actual location of the airplane are found.

In the block "$-\tan L$" the computation of equation (16) is carried out. This yields a signal proportional to the z-component of the Earth's angular velocity, which is used for correcting the Z-gyro for Earth's rotation.

The blocks to the right represent the servo system for keeping the platform in a horizontal position. The servomotors are controlled by the amplified error signals ε_x, ε_y and ε_z of the RAMPs and the Z-gyro. The signal of the Z-gyro, which has to control only one servomotor, is fed directly to this motor. The error signals of the RAMPs however have to be resolved into two components, which must be fed to the other two motors in a certain ratio (depending on the position of the gimbals). This is done in the blocks marked "resolver".

$$M_{kx} = -mh(2\dot{R}\omega_{Ix} - R\omega_{Iy}\omega_{Iz}),$$
$$M_{ky} = -mh(R\omega_{Ix}\omega_{Iz} + R\omega^2_{IE}\sin L\cos L + 2\dot{R}\omega_{Iy} + e\ddot{R}\sin 2L).$$

A block diagram of the signal flow in a complete RAMP system is shown in *fig. 27*. The correction torques are calculated in a computer and added to the total torques. As the equations indicate, there is a coupling between the two RAMPs, which will lead to induced oscillations.

A three-axis system analysis (which will not be presented here) leads to the following characteristic equation:

$$(p^2 + \omega^2_{IE})\{p^4 + 2(g/R + 2\omega^2_{IE}\sin^2 L)p^2 + (g/R)^2\} = 0,$$

where p is the linear time derivative operator d/dt. We see that this equation is of the 6th order. The solution consists of terms containing the following factors:

$$\sin\ \omega_{IE}t,$$
$$\cos\ \omega_{IE}t,$$
$$\sin\ (\omega_s + \omega_{IE}\sin L)t,$$
$$\cos\ (\omega_s + \omega_{IE}\sin L)t,$$
$$\sin\ (\omega_s - \omega_{IE}\sin L)t,$$
$$\cos\ (\omega_s - \omega_{IE}\sin L)t,$$

where ω_s corresponds to the frequency of a Schuler pendulum. As the last four terms reveal, these "Schuler oscillations" are modulated; the modulation frequency, $\omega_{IE}\sin L$, corresponds to Foucault oscillations. Consequently the RAMP system will be subject to oscillations at the "Earth's frequency" (period 24 hours) and to modulated Schuler oscillations (period 84.4 min).

Further literature:

C. S. Draper, W. Wrigley and L. R. Grohe, The floating integrating gyro and its application to geometrical stabilization problems on moving bases, Institute of Aeronautical Sciences, S. M. F. Fund Paper No. FF-13, New York, 1955.
W. Wrigley, F. E. Houston and H. R. Whitman, Indication of the vertical from moving bases, Institute of Navigation Paper, Massachusetts Institute of Technology, Cambridge, Mass., 1956.
F. Hector and K. J. Åström, Swedish Pat. 189 939, application made 24 September 1959, application granted 16th January 1964.
K. J. Åström and F. Hector, Vertical indication with a physical pendulum based on electromechanical synthesis of a high moment of inertia, Report 590802 of the TTN-Group, August 1959, reprinted in the Journal of the Swedish Association of Electrical Engineers, Eleknik **8**, 1965, No. 4.
W. R. Markey and J. Hovorka, The mechanics of inertial position and heading indication, Methuen, London 1961.

Summary. Description of an inertial navigation system, developed by the Swedish Philips Co., Ltd, primarily for aircraft navigation. The system is based on indication of the change of the vertical with a Schuler pendulum, whose oscillation time is 84.4 min. This device, which contains a gyroscope with integrating feedback, has been named "RAMP" (Rate and Acceleration Measuring Pendulum). The complete system comprises two RAMP units, mounted on an inertial platform, which is "slaved" to the pendulums. For indication of heading, a directional gyro is mounted on the same platform. The equipment was tested both in the laboratory and during a number of test flights, including flights over the European continent and transoceanic flights. The average uncertainty varied between 2 and 5 nautical miles per hour.

Oversized rectangular waveguide components for millimetre waves

H. J. Butterweck and F. C. de Ronde

Since microwaves first came into practical use, the hollow pipe, usually of rectangular cross-section, has been the standard means for transmission of the waves. In the conventional use of this waveguide the waves can travel in a single precisely defined pattern — a very valuable property of the guide for many applications. If, however, the same techniques are used in the millimetre and submillimetre wavebands, power loss in the guide becomes a very severe problem and the components become prohibitively small. These have been two of the chief reasons for the development of other types of transmission lines.

In the "oversized waveguide" the losses are much lower than in the standard guide, while the advantage of propagation in a well-defined pattern can be maintained by proper "mode control". The size of the guide can be chosen in a convenient range. Another important advantage is that very broad-band components can be made.

Introduction

At microwave frequencies electromagnetic waves can propagate along several types of transmission line. These can be divided into two classes. In the first class the electromagnetic field extends to infinity in the direction transverse to the axis of propagation; the corresponding structures are usually said to be "open". In the second class the electromagnetic field has only a finite transverse extension. Since the boundaries are imposed by metallic screening, this class is said to be "shielded" or "closed".

Well-known transmission lines such as the dielectric line [1], the image line [2], the H-guide [3] and the microwave beam [4], belong to the first class. The shielded structures of the second class include hollow pipes and pipes containing one or more inner conductors. The rectangular waveguide and the coaxial line are typical representatives of this class.

In the millimetre-wave region, to which we shall confine ourselves, both classes of structures have their specific fields of application and both have their special advantages [5]. In this article we shall be concerned with the closed structures. Let us begin by noting some of their advantages. First, the shielding prevents radiation and makes the device insensitive to parasitic electromagnetic fields; there is no coupling even between closely spaced guides. Secondly, the shielding part of the structure gives sufficient rigidity without requiring additional support, and alignment is quite easy. Finally, many almost ideal components (bends, directional couplers, etc.) can be produced: this is not so with the open structures, where diffraction phenomena may easily give rise to unpredictable deviations from the ideal behaviour.

A traditional closed microwave structure is designed to operate in a single mode: waves of a single definite pattern only can be propagated. For this to be true the structure must be of "standard size", i.e. the cross-sectional dimensions must not exceed certain critical values which are of the order of magnitude of the wavelength used. For millimetre waves in particular, if a cross-section is to be of standard size it must be smaller than a few millimetres across. Coaxial transmission lines of this size are extremely lossy and therefore not very suitable for millimetre-wave applications. On the other hand, standard-size rectangular waveguide components for wavelengths as short as 2 mm have been developed and are in current use [6].

If, at a given frequency, the cross-sectional dimensions of a guide do exceed the critical value referred to above, transmission is possible in several modes. As we shall explain more fully below, this will in general, lead to an unstable electromagnetic field pattern, that is to say, the pattern may vary erratically along the guide. If, however, it is possible to prevent the excitation of the higher-order modes, such "oversized" transmission lines have several advantages over standard-size lines. To mention a few: the attenuation can be considerably lower, broadband components can be

Dr. H. J. Butterweck, formerly with Philips Research Laboratories, Eindhoven, is now Professor of Electrical Engineering at the Technical University of Eindhoven, Netherlands; Ir. F. C. de Ronde is with Philips Research Laboratories, Eindhoven.

designed, and, particularly in the millimetre and submillimetre range, manufacture can be easier. The specific problems associated with oversized transmission lines will be the subject of the present article.

Transmission lines of three main types of cross-section are usually considered. These are: the rectangular waveguide, the circular waveguide, and the coaxial line. The latter is of no interest here, as the design of components not exciting higher-order modes proves to be very difficult. The oversized circular waveguide has become very promising, particularly for long-distance transmission [7]. For laboratory applications, however, the rectangular form offers certain advantages.

The present article is mainly about the rectangular type of oversized waveguide, which has already attracted some interest [8]; a brief comparison between the circular type and the rectangular type will be given at the end. After a general discussion of the mode spectrum of the rectangular waveguide, we pay special attention to the measurement and control of higher-order modes. Several components which have been developed in our laboratory will be described in detail. In particular, we discuss mode filters, tapers, bends, directional couplers, and variable attenuators, phase shifters, and impedances. Finally various applications are enumerated.

Most of our components have been designed for and tested in the 4-mm wave region, and have been built using standard 3-cm waveguide as a starting point.

Mode spectrum of the rectangular waveguide

To give an insight into the problems associated with higher-order modes let us consider some fundamental properties of hollow waveguides [9]. Assuming that we have a waveguide of constant cross-section, let us raise the frequency (instead of taking increased dimensions at a given frequency, as above). To begin with no propagation of electromagnetic waves is possible at all below a critical frequency, the *cut-off frequency* f_{c1}. This cut-off frequency is a specific property of the cross-section under consideration. For frequencies f slightly above f_{c1} the electromagnetic field propagates in a characteristic pattern, the *fundamental mode* of the waveguide. If the frequency is raised further, a second cut-off frequency f_{c2} is obtained, above which another mode with a different field pattern can be propagated. (For f_{c2} the waveguide has the critical dimensions referred to in the introduction.) Every waveguide has an infinite spectrum of such cut-off frequencies $f_{c1}, f_{c2}, f_{c3}, \ldots$, so that for sufficiently high frequencies the energy can be transmitted in a large number of modes simultaneously. In most cases it turns out that the cut-off wavelength λ_{c1} corresponding to the first cut-off frequency f_{c1} has the same order of magnitude as the dimensions of the cross-section (this wavelength is given by $\lambda_{c1} = c/f_{c1}$, where c is the free-space velocity of light). Thus any wave that can be propagated has a free-space wavelength λ comparable to these dimensions or smaller.

A specific mode is propagated along the waveguide without change of the transverse field distribution. In a lossless waveguide, the amplitude is constant. The phase velocity v_i of the mode i, i.e. the velocity of propagation of the field pattern, is given by:

$$v_i = c/\sqrt{1-(\lambda/\lambda_{ci})^2}, \quad \ldots \quad (1)$$

where λ_{ci} is the cut-off wavelength of the mode. Thus, at a given frequency, the phase velocities are different for different modes. Therefore, if several modes are excited, the transverse electromagnetic field distribution varies along the guide owing to the continuously varying phase differences between the individual modes. This unstable situation may occur if $f > f_{c2}$, and if the frequency satisfies this condition the waveguide is said to be *overmoded* or *oversized*. If the frequency lies between f_{c1} and f_{c2}, the case which we shall refer to as *standard-size operation*, only the fundamental mode is present and the propagation is stable. If stable operation is required in oversized waveguide, special care has to be taken to excite only one mode, which need not necessarily be the fundamental one. Such "mode control" will be discussed in detail in the next section.

We now apply these general considerations to the special case of a rectangular cross-section. Let a and

[1] G. Schulten, Archiv elektr. Übertr. **14**, 163-166, 1960; see also: H. Severin and G. Schulten, Surface-wave transmission lines for microwave frequencies, Philips tech. Rev. **26**, 342-356, 1965.

[2] S. P. Schlesinger and D. D. King, IRE Trans. **MTT-6**, 291-299, 1958. The image line is also discussed in the second article in ref. [1].

[3] J. W. E. Griemsmann and L. Birenbaum, Proc. Symp. Millimeter Waves, New York 1959, publ. Polytech. Inst. Brooklyn 1960, p. 543-562.

[4] T. Nakahara and N. Kurauchi, IEEE Trans. **MTT-15**, 66-71, 1967 (No. 2); A. G. van Nie, Philips Res. Repts. **19**, 378-394, 1964 and Nachr.techn. Z., edn. Comm. J. **6**, 264-269, 1965.

[5] An extensive survey of several types of transmission line studied in view of their application at millimetre and submillimetre wavelengths and paying special attention to losses has been made by D. J. Kroon and J. M. van Nieuwland of this laboratory. This survey is published in: D. H. Martin, Spectroscopic techniques for far infra-red, submillimetre and millimetre waves, North-Holland Publ. Co., Amsterdam 1967, Chap. 7: Techniques of propagation at millimetre and submillimetre wavelengths, p. 308-380.

[6] C. W. van Es, M. Gevers and F. C. de Ronde, Waveguide equipment for 2 mm microwaves, Philips tech. Rev. **22**, 131-125 and 181-189, 1960/61.

[7] S. E. Miller, Bell Syst. tech. J. **33**, 1209-1265, 1954.

[8] J. J. Taub, H. J. Hindin, O. F. Hinckelmann and M. L. Wright, IEEE Trans. **MTT-11**, 338-345, 1963.

[9] For extensive theoretical treatments of this subject the reader is referred to standard textbooks on waveguide theory. An introduction is given in: W. Opechowski, Electromagnetic waves in waveguides, Philips tech. Rev. **10**, 13-25 and 46-54, 1948/49.

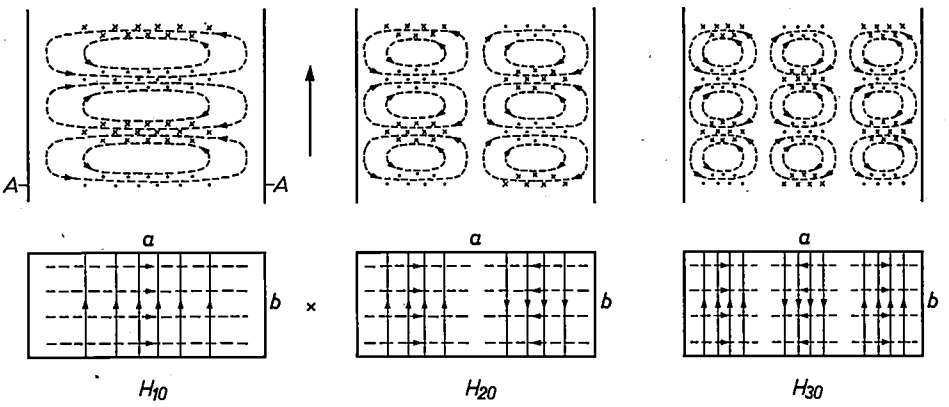


Fig. 2. Field pattern of the fundamental mode and some higher-order modes in rectangular waveguide. The longitudinal cross-section is shown above, and the transverse cross-section below. The electric field lines are shown solid and the magnetic field lines dashed. It should be noted that many of the lines drawn in the cross-sectional planes are not the actual field lines, but projections of the field lines on these planes. In the longitudinal cross-sections the directions indicated (arrows) are those of the lines near the top of the guide, and in the transverse cross-section the directions are those of the lines near the section A-A (as indicated in the first drawing). These directions correspond to the direction of propagation indicated between the first two drawings. The various modes are drawn in such a way that they have the same guide wavelength $\lambda_g = v_t/f$ (the length of the longitudinal section drawn is $\frac{3}{2}\lambda_g$); this implies that, if the λ_{ct}'s differ, the frequencies will be different (cf. eq. 1).

b be the width and the height of the rectangle, respectively, where $a > b$. The first cut-off wavelength is then given by:

$$\lambda_{c1} = 2a. \quad \ldots\ldots\ldots \quad (2)$$

The field pattern of the fundamental mode, often referred to as the H_{10} mode, is shown in *fig. 1*. The electric field lines E are parallel to the shorter side of the waveguide while the magnetic field lines H lie in planes perpendicular to the electric field. The cut-off wavelengths λ_{cmn} of the higher-order modes are given by:

$$(2/\lambda_{cmn})^2 = (m/a)^2 + (n/b)^2. \quad \ldots \quad (3)$$

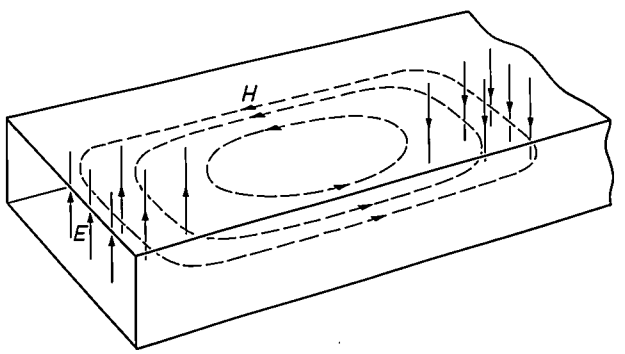


Fig. 1. Field pattern of the H_{10} mode in rectangular waveguide.

This expression yields the complete mode spectrum: an infinite set of modes, each characterized by two mode numbers m and n, one of which, but not both, may be zero.

In *fig. 2* the field patterns of the fundamental mode and some higher-order modes of the rectangular waveguide are shown. If $m = 0$ or $n = 0$, then only the magnetic field has an axial component H_z; the axial component of the electric field E_z vanishes. Modes of this type are generally called H modes. Thus we have an H_{20} mode, an H_{30} mode, an H_{01} mode, etc. If neither m nor n is zero, two different field patterns are possible which have the same cut-off wavelength λ_{cmn}. The first, known as H_{mn}, has again a vanishing electric field component E_z, but a non-vanishing H_z; the second, known as E_{mn}, has a vanishing magnetic field component H_z, but a non-vanishing E_z. The indices m and n are equal to the number of half periods of the wave pattern along the width a and the height b respectively.

Fig. 3 is an illustration of how the field pattern of the fundamental mode in an oversized waveguide is elongated in the transverse direction and compressed in the longitudinal direction, compared with the pattern in standard-size guide. When a guide is highly oversized the pattern is close to that of a transverse electromagnetic wave.

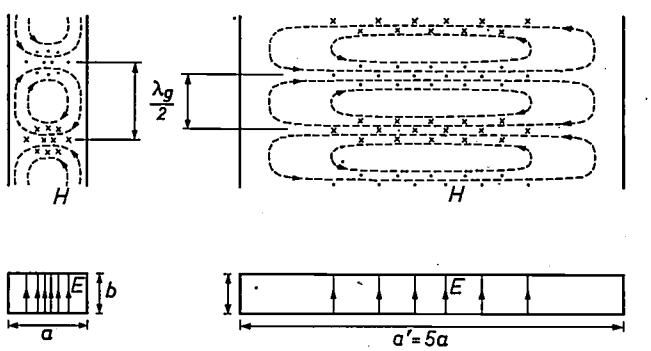
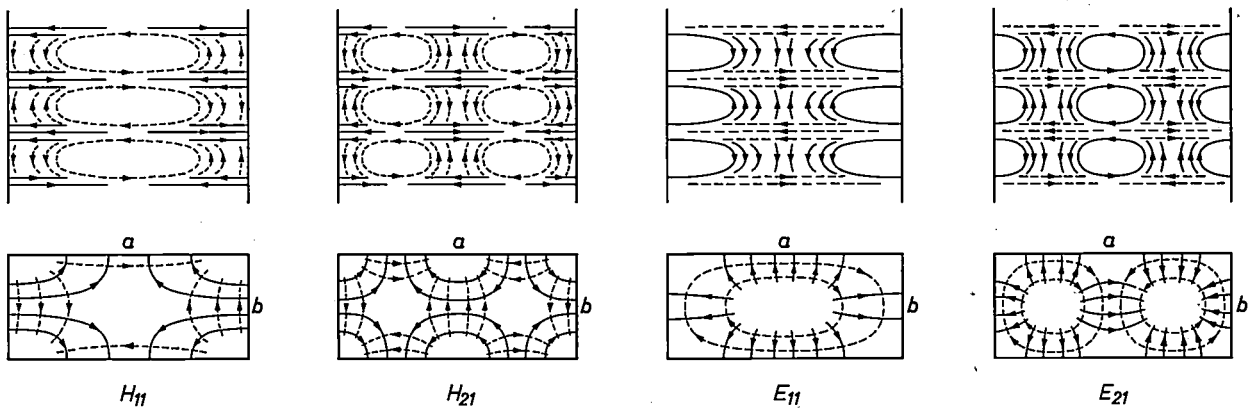


Fig. 3. The field pattern of the fundamental mode H_{10} at a given free-space wavelength λ in a standard-size guide (the width a is smaller than λ) and in a guide of width $a' = 5a$ which is oversized if $5a > \lambda$ (cf. eq. 3). Oversizing has been obtained here by increasing the width; sufficient increase of the height alone would also give oversizing: in practical oversized waveguide *both* dimensions are usually considerably larger than in the standard-size guide.



Eq. (3) yields the cut-off frequencies $f_{cmn} = c/\lambda_{cmn}$ for any combination of indices m and n. The mode spectrum $f_{c1}, f_{c2}, f_{c3} \ldots$, arranged in order of increasing cut-off frequencies, depends upon a and b. The ratio a/b determines the higher cut-off frequencies relative to f_{c1}. Mode spectra of this type are shown in *fig. 4* for the two values $a/b = 22.86/10.16 = 2.25$ and $a/b = 22.86/1.55 = 14.8$. The first case corresponds to standard 3-cm waveguide (inner dimensions $0.9'' \times 0.4''$, i.e. 22.86×10.16 mm), the second to a flat waveguide, with the same width as 3-cm waveguide but height equal to that of 4-mm waveguide. For both configurations the ratio f_{c2}/f_{c1}, which determines the frequency band of stable H_{10} mode transmission, is equal to 2. It is easy to show that this is the case whenever $a/b \geqq 2$. If this condition is met the rectangular waveguide, for standard-size operation, can be used theoretically over a frequency range of an octave. For practical purposes, however, only part of this band is useful: in the vicinity of f_{c1} the losses of the H_{10} mode become very high, and in the vicinity of f_{c2} the next higher-order mode H_{20} is almost propagating, which means that this mode, if excited by an obstacle, dies away very slowly. The frequency ratio f_{max}/f_{min} of the

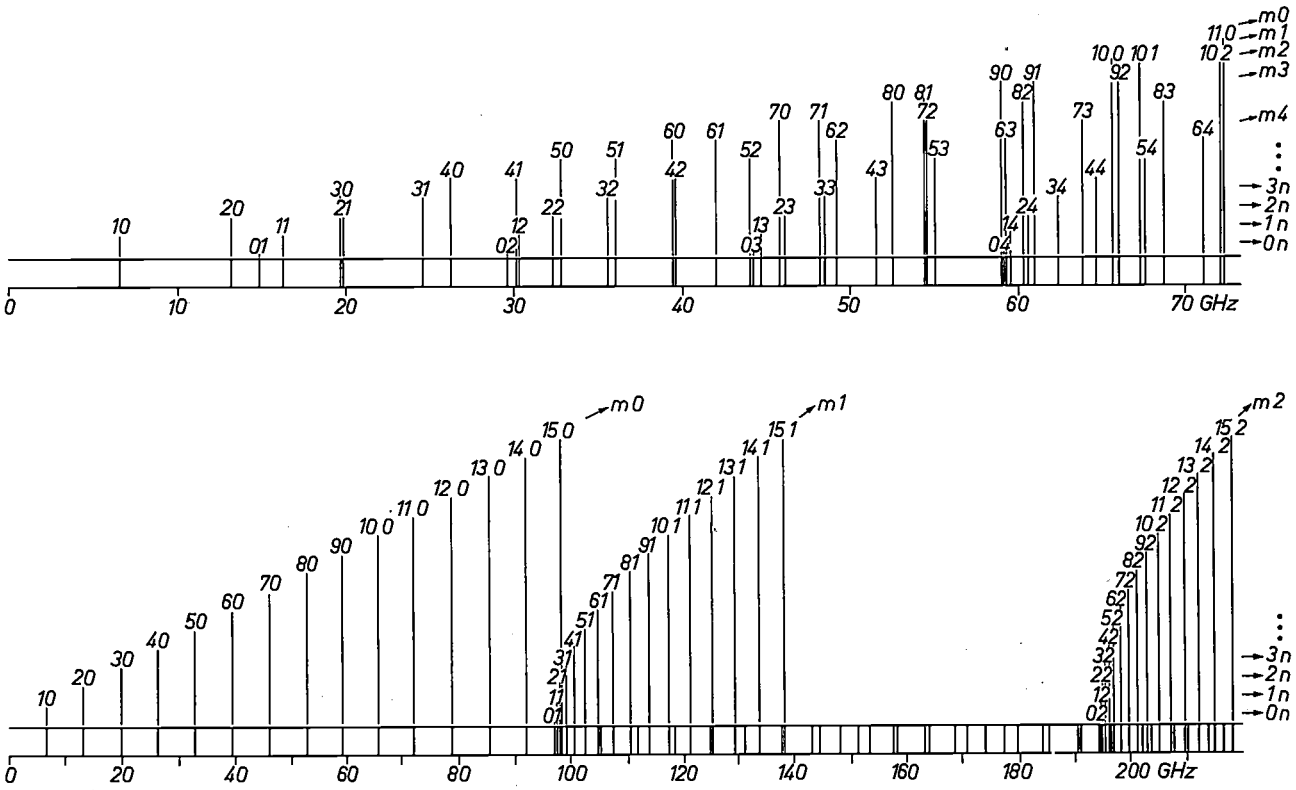


Fig. 4. Mode spectrum of standard 3-cm waveguide (upper diagram; $a/b = 2.25$) and spectrum of a flat waveguide with the same width as 3-cm waveguide but height equal to that of 4-mm waveguide (lower diagram; $a/b = 14.8$). The spectra show the cut-off frequencies f_{cmn} above which the mode (m, n) can propagate. The numbers m and n are indicated in the figure. For a specific pair m,n, both $\neq 0$, two modes exist: the E_{mn} mode and the H_{mn} mode, with the same cut-off frequency. For m or $n = 0$ only the modes H_{0n} and H_{m0} are possible; E_{0n}, E_{m0} and H_{00} do not exist.

band recommended for standard-size operation of rectangular waveguides is only 1.6.

Mode control

As mentioned above, special care has to be taken if, in an oversized waveguide, a single mode is wanted. This is because, even when a single mode is launched at the input, many modes will easily arise in a circuit: in each microwave component and even at each irregularity in a waveguide there will be *mode conversion*, i.e. if a wave of a certain mode is incident on the component or at the irregularity, other modes are excited.

One may ask just why multimode operation is so undesirable. The answer is that in multimode operation a simple description of the electromagnetic waves in a circuit is no longer possible. Even in a simple piece of oversized waveguide the transverse field distribution is not well defined and varies along the guide. More generally, if a component is considered as a black box with a number of waveguide arms, the reflection and transmission coefficients of the wanted modes as well as the conversion coefficients must be known for a complete description of such a component. If there are several components in the circuit, the mathematical description may soon become so unwieldy that in practice the field distribution is unpredictable; on the other hand, measurements in such a circuit will not lead to definite conclusions about the characteristics of a certain component.

If the conversion from the wanted mode to unwanted modes is negligible, we can refer to single-mode or stable operation of a specific oversized waveguide circuit. However, it should be emphasized that, even if the coefficients describing the conversion into the separate components are very small, higher-order modes with substantial amplitudes may arise. This may occur for example if the higher-order modes, excited at some irregularity, are trapped between two tapers, as shown in *fig. 5*. A taper (to be treated in the following section) gives a smooth transition between an oversized waveguide and a standard-size waveguide. Since only the fundamental mode can propagate in the standard guide, a higher-order mode incident from the oversized section is totally reflected. Multiple reflections of the higher-order modes may give rise to resonance of these modes. If the frequency of the source coincides with one of the resonance frequencies, absorption takes place causing a considerable loss in the total transmitted power. This, when plotted as a function of frequency (fig. 5), may exhibit a great number of such resonance absorptions. The density of these resonances increases with increasing degree of oversizing and increasing length of the trapped-mode resonator. Furthermore, the larger the conversion coefficient of the irregularity and the lower the attenuation per unit length of the resonating mode, the stronger the absorption. In the limiting case of vanishing attenuation the transmission may even be zero [10], which corresponds to a total reflection of the incident wave.

There are some special applications of oversized waveguides where it would be very difficult to achieve single-mode transmission, but multimode operation is however acceptable. For instance, if submillimetre waves are transmitted through a standard centimetre-waveguide, this guide is oversized to a very high degree, and in such a case it is extremely difficult to make an almost ideal taper and launch only a single mode [11]. Nevertheless, such a guide may be useful, if only for transferring power over normal laboratory distances with low attenuation. In this case the existence of higher-order modes does not essentially affect the overall transmission, for if the waveguide is long enough the trapped-mode resonances are well damped out.

Such special cases apart, oversized waveguides will usually be used in single-mode operation. This is almost invariably so in measurement work. If, for instance, a bridge circuit has to be designed, the components, such as power dividers and directional couplers should be substantially free from mode conversion, since the excitation of higher-order modes may easily lead to unpredictable errors.

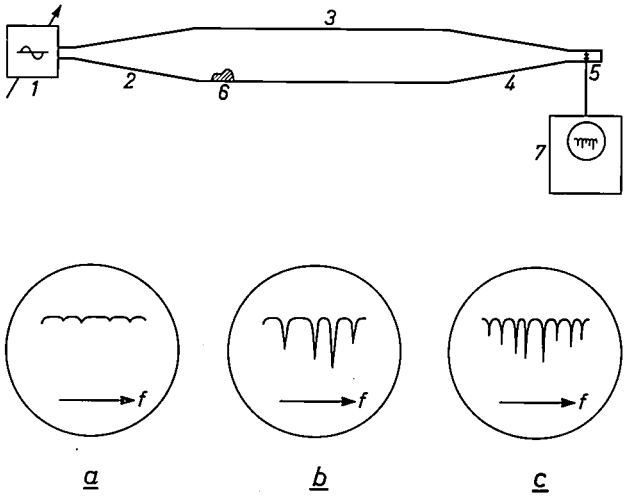


Fig. 5. Trapped-mode resonances. Microwave power from a swept-frequency generator *1* with standard-size output is fed through a taper *2* into oversized waveguide *3* and by a second taper *4* into a broadband detector *5* in standard-size waveguide. The tapers are assumed to be free from mode conversion. Mode conversion will take place at any irregularity *6*. A higher-order mode will be reflected in the tapers and be trapped, giving resonance losses at certain frequencies, which can be determined from the power-frequency plot on the oscilloscope *7*. *a*) Low mode conversion at *6*; *b*) high mode conversion; *c*) the density of resonances increases if the length of the oversized waveguide *3* or its degree of oversizing is increased.

Some time ago, we began an investigation in this laboratory to find out how to design components with low mode conversion. For this work, we needed equipment for measuring the conversion coefficients. An essential element of such equipment is the *mode transducer*, several types of which are described below. If the components to be developed do not satisfy certain requirements for mode purity, this can be improved by *mode filters*, which will be described later. The measurements to test the components that had been developed were all made with the aid of a swept-frequency technique in which we used a backward-wave oscillator that could be swept between 72 and 77 GHz.

Mode transducers

A typical arrangement for the measurement of mode conversion using a mode transducer is shown in *fig. 6*. The waveguide component under test is excited by a wave in the wanted H_{10} mode incident from the left. An H_{10} mode and several higher-order modes produced

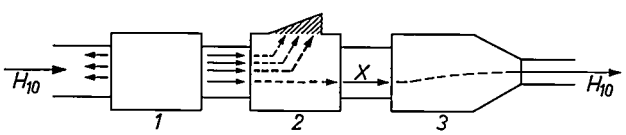


Fig. 6. Measurement of the conversion of the wanted mode H_{10} into a specific higher-order mode X in a component (*1*) to be tested. The test is carried out with the aid of a mode transducer *3* combined with a mode filter *2*. The fundamental mode is incident at the input (standard or oversized) of *1*. The fundamental mode and several higher-order modes may leave by the output as well as by the input (reflection). In *2* all the modes from the output of *1*, including H_{10}, are absorbed, except the mode X, which is converted in *3* into the H_{10} mode in the standard-size output of the transducer. In an ideal filter-transducer combination none of the modes are reflected at the input of the mode filter.

in the component by mode conversion leave by the right-hand arm. All these modes except the one to be measured (X) are completely absorbed by a mode filter before the transducer. (The mode filter is often incorporated in the transducer.) The remaining mode X is converted into the H_{10} mode in the outgoing standard-size waveguide of the transducer. The way in which this conversion is actually performed is different for each type of mode transducer, and will be discussed below. If the transmission characteristics of the mode transducer are known the overall $H_{10} \rightarrow H_{10}$-transmission of the arrangement is a measure of the mode conversion $H_{10} \rightarrow X$ in the component under test. Care must be taken that the transducer does not reflect the mode X to be measured, since a reflected wave might cause further mode conversion in the unknown device. The input waveguide of the component under test may be oversized or of standard size.

A different approach for testing components is to measure the distribution of the modes in the oversized waveguide by a method proposed by Peaudecerf [12] and by Levinson and Rubinstein [13]. This method is based on probing the electromagnetic field with the aid of a single movable probe or a number of identical fixed probes. The amplitudes of the individual modes can be evaluated from the field distribution measured in this way. The probe method has the advantage that the mode distribution can be measured in a system while it is operating, while the method using mode transducers only permits the evaluation of the mode conversion in a single component. Small distortions of the wanted mode pattern, however, cannot be accurately determined by the probe method, whereas with the transducers to be described below, spurious mode levels of more than 40 dB below the fundamental mode level can easily be measured. Another disadvantage of the probe method is that the probes may disturb the original field distribution.

Before describing a few types of mode transducer in detail, let us consider the possible higher-order modes in a rectangular waveguide. These modes (cf. fig. 2) can be divided into two classes according to whether the field does or does not vary in the y-direction (the direction perpendicular to the broad face of the guide). Class 1, in which the field does vary with y, consists of the E_{mn} and H_{mn} modes with $n \neq 0$. Class 2, in which the field does not depend on y, contains the H_{m0} modes with $m \neq 1$. It will be shown that mode filters can easily be designed for the modes belonging to class 1. Thus no difficulties arise from conversion into these modes provided the power loss of the H_{10} mode is not excessive. For this reason, mode transducers have been designed for the second class only, and in this class only for the most interfering types H_{20} and H_{30}.

An $H_{10} \rightleftarrows H_{20}$ mode transducer first proposed by Montgomery, Dicke and Purcell [14], is shown in *fig. 7*. A standard-size waveguide is split into two parts by means of a metal sheet perpendicular to the electric field. The two guides then twist through 90° in opposite senses. This brings the two field patterns side by side, giving the wanted H_{20} mode. Finally there is a smooth linear taper to an oversized waveguide. As this device is reciprocal it can be used in both directions, i.e. as an H_{20} detector or as an H_{20} exciter. This transducer is subject to some mode conversion in the taper, but this effect can be made arbitrarily small by using a suffi-

[10] H. J. Butterweck, A theorem about lossless reciprocal three-ports, IEEE Trans. CT-15, No. 1, March 1968.
[11] H. J. Butterweck, unpublished; see also: H.-G. Unger, Bell Syst. tech. J. 37, 899-912, 1958.
[12] M. Peaudecerf, Thesis, University of Toulouse, 1966.
[13] D. S. Levinson and I. Rubinstein, IEEE Trans. MTT-14, 310-322, 1966 and 15, 133-134, 1967 (No. 2).
[14] C. G. Montgomery, R. H. Dicke and E. M. Purcell, Principles of microwave circuits, Radiation Laboratory Series No. 8, McGraw-Hill, New York 1948, p. 339.

ciently smooth taper. Furthermore, a small asymmetry in the twisted section may give rise to a residual H_{10} mode in the oversized waveguide; this is absorbed by means of a resistive sheet placed at the centre of the guide. An additional function of this sheet is the absorption of any H_{10} mode energy that might arrive from extremely thin or not positioned exactly, there is some mode conversion $H_{30} \rightarrow H_{10}$, diminishing the effectiveness of the mode filter. (In the H_{20} mode filter described above the absorbing sheet need not be extremely thin, as there is no mode conversion $H_{20} \rightarrow H_{10}$ even in a thick sheet.)

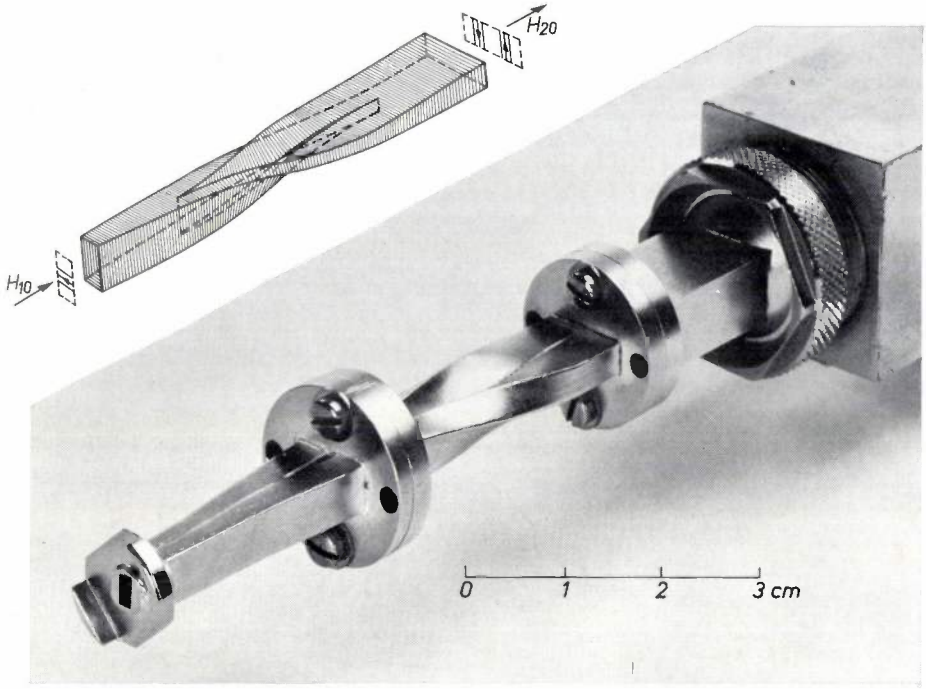


Fig. 7. $H_{10} \rightleftarrows H_{20}$ mode transducer.

the oversized section. Finally, care must be taken to avoid trapped-mode resonances associated with modes which cannot propagate through the taper. It is therefore advisable to use this transducer in connection with one of the separate mode filters described in the following section.

We have developed a mode transducer for the H_{30} mode, making use of the principle of mode-selective directional coupling (*fig. 8*). Two waveguides whose widths differ by a factor of *3* are coupled by means of a slot. Since the H_{10} mode in the narrow guide and the H_{30} mode in the broad guide have equal phase velocities they interact strongly. In a way analogous to the action of two coupled pendulums of equal resonance frequency, the energy of the H_{10} mode in the narrow guide is transferred into the H_{30} mode in the broad guide, the distance for complete transfer depending on the width of the slot. A smooth taper guides the H_{30} mode into the oversized waveguide.

Since the directional coupler and the taper are not ideal the residual power propagating in the H_{10} mode has to be dissipated in a suitable mode filter. We have used two thin resistive sheets located at the nodes of the electric field of the H_{30} mode. If the sheets are not

Measurements on a complete mode-transducer assembly, using a 12-μm "Mylar" [*] sheet with a resistive metallic film for the mode filter, have indicated an H_{30} mode purity of better than 50 dB. The surface impedance of the resistive sheets of the mode filter can be optimized by applying well-known principles described elsewhere [15]. For standard *X*-band (3-cm) waveguide

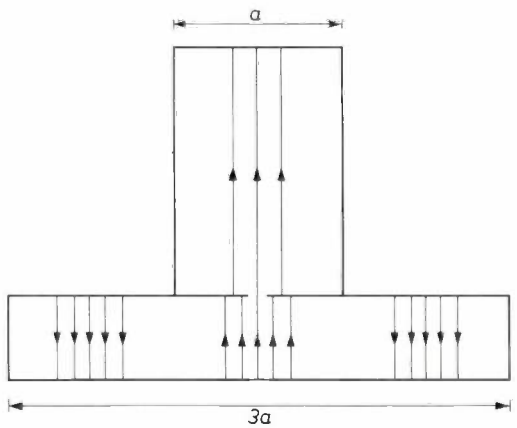


Fig. 8. Transverse cross-section of $H_{10} \rightleftarrows H_{30}$ mode transducer. Owing to the equal phase velocities of the H_{10} mode in the upper guide and the H_{30} mode in the lower guide, power in one of these modes will be completely transferred into the other in a distance along the guides which depends on the width of the coupling slot.

and a frequency $f = 75$ GHz (4 mm) one obtains an optimum surface resistance of $R_\square = 560$ ohms giving an attenuation of $\alpha = 2.4$ dB/cm for the H_{10} mode. As in the H_{20} mode filter the resistive sheets have the additional function of absorbers for any H_{10} mode energy incident from the oversized waveguide section. These sheets also suppress potential trapped-mode resonances.

Mode filters

In the mode transducers described above suitable mode filters were incorporated where necessary. As, however, mode conversion is an effect which can easily occur in oversized components, separate mode filters are desirable for mode control. A mode filter is essentially a two-port which strongly attenuates unwanted modes, while transmitting the wanted mode (usually the H_{10} mode) almost unperturbed. Furthermore, all modes should be matched; in other words, the mode-selective transmission should not be achieved by means of reflecting obstacles as these might give rise to trapped-mode resonances.

A simple design for a mode filter for the higher-order modes of class 1 (cf. p. 91), that is to say, the E_{mn} and H_{mn} modes with $n \neq 0$, consists of one or more resistive sheets perpendicular to the electric field lines of the wanted H_{10} mode (*fig. 9*). Such sheets do not perturb the propagation of the H_{10} mode, but attenuate E_{mn} and H_{mn} modes with $n \neq 0$ since these modes have an electric field component parallel to the sheets. Unperturbed transmission can however occur if the planes of the sheets happen to coincide with nodal planes of one of these modes.

It should be noted that an H_{mn} mode ($n \neq 0$) incident at one port of this general type of mode filter gives not only an (attenuated) H_{mn} mode at the other port, but an E_{mn} mode as well (with the same subscripts m, n). Likewise, an incident E_{mn} mode sets up an H_{mn} mode at the other port. This mode conversion is related to what is called the "degeneracy" of the two modes E_{mn} and H_{mn}, which means that their phase velocities are equal. If the modes E_{mn} and H_{mn} are superimposed in a rectangular waveguide, their phase difference does not vary along the guide, i.e. the field pattern is stable. Thus every superposition of these modes also gives a mode.

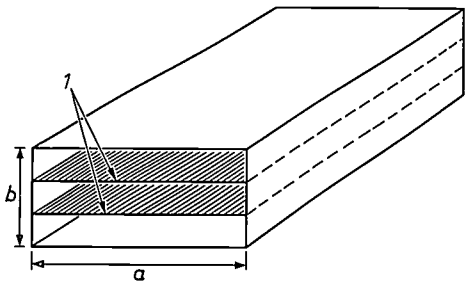


Fig. 9. Filter for the modes E_{mn} and H_{mn} with $n \neq 0$. These modes are attenuated by the resistive sheets *1*, unless they happen to have no tangential E-component at the sheets. The H_{10} mode is unperturbed since its E-field is perpendicular to the sheets.

It has been shown [15] that there are two such combinations of the E_{mn} and H_{mn} mode which undergo no mode conversion in this type of mode filter. These combinations are characterized by the properties $E_y = 0$ and $H_y = 0$, respectively. This implies that either the electric or the magnetic field lines lie in planes parallel to the sheet. These two modified mn-modes are often called "longitudinal-section modes" (LS-modes). An E_{mn} or H_{mn} wave incident from the empty guide sets up both types of LS-mode. Thus for an efficient filtering of E_{mn} and H_{mn} modes both LS-modes have to be attenuated sufficiently in the mode filter.

The electric field patterns of the LS-modes inside the mode filter are shown in *fig. 10* for the case $m = n = 1$ and with one symmetrically located sheet. The

[15] H. J. Butterweck, Mode filters for oversized rectangular waveguide, to be published in IEEE Trans. MTT.

[*] Registered trade mark of Du Pont de Nemours.

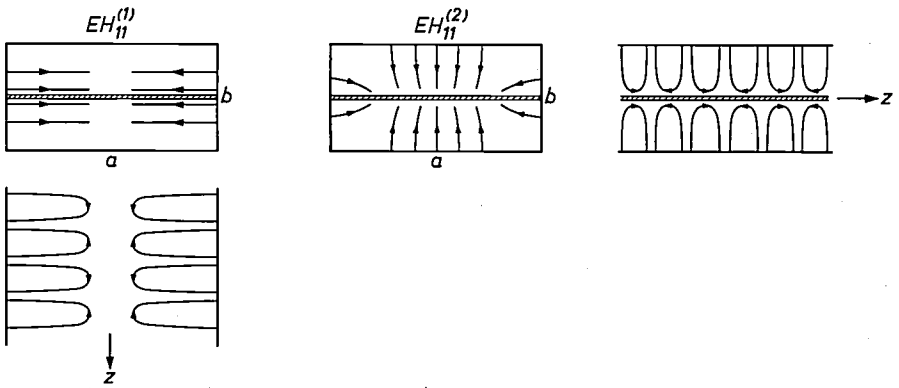


Fig. 10. The electric field pattern of the longitudinal-section (LS) modes with $m = n = 1$, $EH_{11}^{(1)}$ and $EH_{11}^{(2)}$, in a mode filter with one symmetrically located resistive sheet. The LS modes do not undergo mode conversion in a filter of the type of fig. 9.

first mode, called $EH_{11}^{(1)}$, has two electric field components, E_x and E_z, parallel to the sheet, E_z being negligible for highly oversized waveguide. This mode is considerably attenuated by the sheet. For the second mode, $EH_{11}^{(2)}$, however, the only important component parallel to the sheet is E_z and this axial component is very small anyway, as might be expected since waves in highly oversized waveguide are approximately transverse electromagnetic in character (cf. fig. 3). Thus the surface resistance of the sheet has to be chosen carefully to obtain a reasonable value for the attenuation of the $EH_{11}^{(2)}$ mode. As is usual in absorption problems, there is an optimum value of the surface resistance $R_\square$ (neither a perfect conductor nor a perfect insulator absorb electromagnetic waves), and for the case of a *single* symmetrically located sheet this value has been shown to be [15]:

$$R_\square = 140\lambda/b \text{ ohms}. \quad . \quad . \quad . \quad . \quad . \quad (4)$$

The corresponding maximum attenuation of the $EH_{11}^{(2)}$ mode (in dB per unit length) is:

$$\alpha_{max} = 3.8\ \lambda/b^2. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

For our example of standard 3-cm waveguide ($b = 10.16$ mm) and $\lambda = 4$ mm we obtain $R_\square = 56$ ohms and $\alpha_{max} = 1.5$ dB/cm.

It should be mentioned that the comparatively low value of $R_\square$ given by eq. (4) is quite different from the corresponding optimum value for the $EH_{11}^{(1)}$ mode. For a high degree of oversizing, the attenuation of the $EH_{11}^{(1)}$ mode can be even lower than that of the $EH_{11}^{(2)}$ mode, if the optimum value for the latter is chosen from eq. (4). In this case a compromise for the value of the surface resistance $R_\square$ is recommended [15].

For the higher-order modes of class 2 (p. 91), that is to say the H_{m0} modes with $m \neq 1$, we were unable to design a mode filter with such an almost ideal performance as that of the filter described above. However, we give details of two designs which do at least meet certain requirements. The first type is inherently broadband and attenuates all the higher modes simultaneously; but the wanted H_{10} mode is also attenuated to a certain extent. The second type of mode filter transmits the H_{10} mode without losses, but is narrow-band and absorbs only a specific higher-order mode.

The first design consists of a waveguide whose narrow walls are made of lossy material (see *fig. 11*). If the broad walls are assumed to be perfectly conducting, the attenuation (in dB per unit length) of the H_{m0} modes is given by

$$\alpha_m = 8.67\ \frac{m^2\lambda^2}{2a^3}\ \frac{v_m}{c}\ \frac{R_w}{Z_0}. \quad . \quad . \quad . \quad (6)$$

Here v_m is the phase velocity of the mode under consideration as given by eq. (1), R_w is the surface resistance of the lossy walls, and $Z_0 = 377$ ohms is the free-space characteristic impedance. If the waveguide is highly oversized and if m is not too high, we have $v_m \approx c$ so that α_m may be assumed to increase in proportion to m^2. Thus the H_{30} mode is attenuated nine times as much as the H_{10} mode.

This dependence of the attenuation upon the mode number m can be illustrated with the aid of the well-known representation of an H_{m0} mode by the zigzag reflections of a plane wave [16]. The angle Θ (see fig. 11) between its direction of propagation and the reflecting wall follows from the relation:

$$\sin\Theta = \lambda/\lambda_{cm} = m\lambda/2a. \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

Thus Θ increases with increasing mode number m. The distance L which the wave travels along the guide between two reflections is given by:

$$L = a/\tan\Theta. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

The relative power loss per reflection equals $1 - |r|^2$, where r is the complex amplitude reflection coefficient of the wall. The power loss in dB per unit length of the guide, which equals twice the attenuation constant α_m, is therefore:

$$2\alpha_m = 8.67(1 - |r|^2)/L = 8.67(\tan\Theta)\,(1 - |r|^2)/a. \quad . \quad (9)$$

The power reflection coefficient $|r|^2$ of a plane wave obliquely incident on a dissipative medium with surface resistance R_w can be found in standard textbooks and is given by:

$$|r|^2 = 1 - 4(R_w/Z_0)\sin\Theta. \quad . \quad . \quad . \quad . \quad . \quad (10)$$

If eqs. (7) and (10) are substituted in (9) and use is made of eq. (1), we obtain eq. (6) for the attenuation of the H_{m0} modes. It is clear that the m^2 law is due to the fact that a) the number of reflections per unit length, and b) the power loss per reflection increase with increasing mode number m.

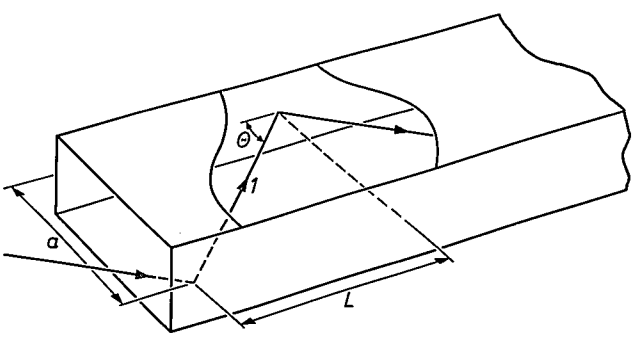


Fig. 11. Filter for the modes H_{m0} with $m \geqq 2$. The side walls are made of lossy material. This filter is not ideal in principle as the H_{10} mode undergoes some attenuation, but the attenuation increases with the mode number m (cf. eq. 6). To illustrate the derivation of eq. (6) an H_{m0} mode is represented by the plane wave *1* which is reflected in a zigzag path by the narrow walls.

The second type of mode filter, designed for the suppression of a specific H_{m0} mode in a highly oversized waveguide, is shown in *fig. 12*. The filtering action is based on the principle of mode-selective directional coupling. Two subsidiary waveguides are coupled to

the main guide by means of dielectric sheets. Let the mode to be absorbed by the filter be of the H_{30} type. In this case the width of the subsidiary guides is made about a third of that of the main guide. The phase velocity of the H_{10} mode in these subsidiary waveguides then approximately equals the phase velocity of the H_{30} mode in the main guide, resulting in a strong interaction between these two modes. As in the $H_{10} \rightleftarrows H_{30}$ transducer described in the preceding section, all the power of the H_{30} mode in the main guide can be transferred to the subsidiary guides and dissipated there, provided all the parameters (physical dimensions, dielectric constant of the sheet) are related in a certain manner which we shall not go into here. A detailed theoretical treatment is given elsewhere [15].

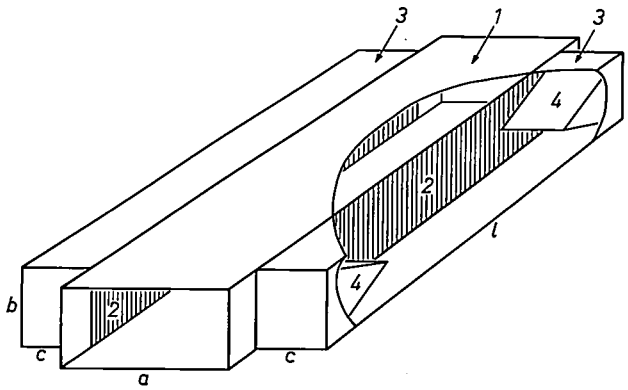


Fig. 12. Filter for the H_{30} mode. Power in the H_{30} mode in the main waveguide *1* is transferred by means of dielectric sheets *2* into the H_{10} mode in the subsidiary guides *3*, which is absorbed in the loads *4*. For the filter to be effective the phase velocities of H_{30} in *1* and H_{10} in *3* must be equal, i.e. c must be equal to $a/3$.

Both types of H_{m0} mode filter described above suffer from the general disadvantage of mode conversion from the H_{10} mode to higher-order modes H_{m0} with m odd. This effect is also described in reference [15].

Mode filters of the three designs described have been made in standard 3-cm waveguide and their performance in the 4-mm wave region has been measured. In the first type (cf. fig. 9) we used three equidistant resistive sheets with a surface resistance of 70 ohms and a length of 10 cm. We are not able to quote quantitative results since we had no suitable mode transducer, but the complete suppression of strong trapped-mode resonances did give a rough indication of an effective mode filtering. The second type (fig. 11) was constructed with a lossy wall having a surface resistance of about 250 ohms. With a length of 15 cm we obtained the following attenuations: 0.6 dB, 2.5 dB, and 5.4 dB for the H_{10}, H_{20}, and H_{30} modes, respectively. A filter of the third type (fig. 12), with a length of 15 cm, gave attenuations of 0.3 dB and 30 dB for the H_{10} and H_{30} modes respectively. For both of the last two types of filter the mode conversion $H_{10} \rightarrow H_{30}$ was about −30 dB.

Components

We shall now describe several components for oversized rectangular waveguide, some of them new, paying regard to mode conversion. Special attention will be paid to their broadband characteristics. In most cases we were able to design components for a 10 to 1 frequency band with the lowest frequency corresponding to standard-size operation.

Tapers

The waveguide output of coherent microwave sources is usually designed for standard-size operation. This means that some sort of transition is needed from standard-size to oversized waveguide. In order to launch an almost pure H_{10} mode in the oversized waveguide, this transition has to be very smooth. Such a "taper" is also needed for the normal types of detector mounted in standard-size waveguide. In principle, a detector can also be designed for oversized waveguide, e.g. in the form of a homogeneous detecting sheet mounted transversely in the waveguide. However, such a detector is too insensitive for most applications. For good sensitivity some sort of concentration of the energy is required. This can be achieved with a taper much more efficiently than with other devices such as concave mirrors and lenses, which moreover will excite higher modes in the oversized guide.

In a taper the field pattern of the H_{10} mode is elongated in the transverse direction and compressed in the longitudinal direction. This must not involve a substantial perturbation of the characteristic field pattern of the mode; in other words, no serious mode conversion is allowed to occur. Furthermore, the reflection coefficient of a taper must be as low as possible. Fortunately in most cases, tapers optimized for low mode conversion also automatically exhibit negligible reflection coefficients. Measurements have indicated that the reflection coefficient is of the order of 1 %.

The tapers considered here have a smoothly-varying cross-section with a) every cross-section rectangular, b) the centres of all the rectangles on a straight line, the "axis" of the taper, c) no twisting of the rectangles with respect to each other. It has been shown [11] that in a smooth rectangular waveguide taper the excitation of the H_{30} mode, which is found to be the most troublesome one, is dependent only upon the axial variation of the width a of the taper. Roughly speaking, the mode

[16] S. Ramo and J. R. Whinnery, Fields and waves in modern radio, Wiley, New York 1944, p. 348. See also the article by Opechowski [9].

conversion decreases with increasing length of the taper, but for a given length the form of the taper can be optimized in order to yield minimum mode conversion.

In the theoretical design providing a minimum $H_{10} \rightarrow H_{30}$ mode conversion the function $a(z)$ has a bottle-like shape as shown in *fig. 13*; b can be a linear function of z. A taper of this form, 140 mm long, between standard 4-mm waveguide (1.55×3.10 mm inner dimensions) and standard 3-cm waveguide, gave a mode conversion $H_{10} \rightarrow H_{30}$ of less than −40 dB (i.e. the excited H_{30} mode is at least 40 dB below the H_{10} mode). On the other hand, a taper with a linear function $a(z)$ and the same length gave a much larger $H_{10} \rightarrow H_{30}$ conversion (a value of about −18 dB was measured).

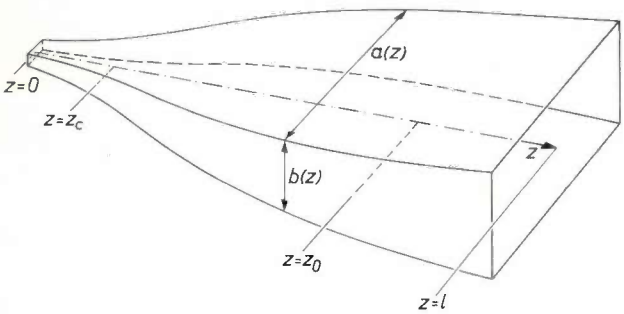
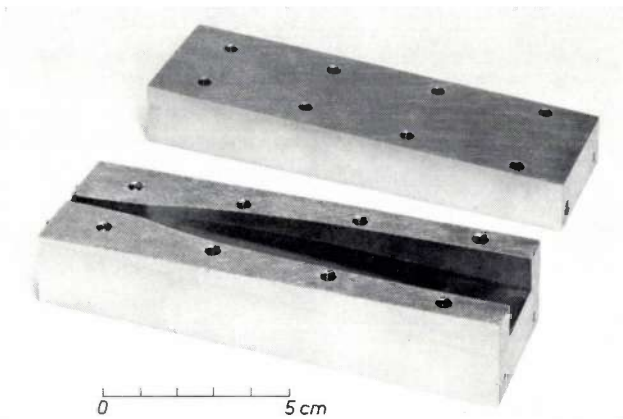


Fig. 13. The taper for transferring the H_{10} mode from standard-size to oversized waveguide. The design is based on a minimum $H_{10} \rightarrow H_{30}$ mode conversion for a given length of the taper. The width a is a function of z with a bottle-like shape, the height b a linear function of z.



We can understand the superiority of the "bottle" form shown in fig. 13 as follows: owing to the finite inclination of the walls an infinitesimal element of the taper at $z = z_0$ gives rise to two elementary waves of the unwanted H_{30} mode; one travels towards the oversized section (to the right in fig. 13), the other in the opposite direction. The wave travelling to the left is totally reflected in the cross-section at $z = z_c$, for which the H_{30} mode becomes cut-off, then it too travels, as a secondary wave, toward the oversized section. Now, for the primary waves, let us consider the phase difference $\Delta\varphi$ between the H_{10} wave incident at the origin ($z = 0$) and the H_{30} wave leaving at the end ($z = l$). $\Delta\varphi$ is determined by the phase velocity of the H_{10} wave from $z = 0$ up to $z = z_0$ and that of the H_{30} wave from $z = z_0$ up to $z = l$. It is therefore dependent upon the position z_0 of the infinitesimal mode-converting element under consideration. Obviously the total H_{30} amplitude found by integration of the elementary contributions is far less than the sum of the absolute values because of "destructive interference" (i.e. contributions of opposite phase cancel each other, partially or totally, depending on their magnitudes). Since the phase constants (phase shifts per unit length) of the H_{10} and H_{30} mode both approach the free-space phase constant with increasing width a of the taper, their difference tends to zero as a tends to infinity. The destructive interference therefore becomes less effective with increasing width, which means that the contributions from the end of the taper, corresponding to the largest values of the width a, have a relatively large weighting factor. This part therefore has to be designed with a very small slope to ensure small elementary contributions to the mode conversion, whereas the narrower part may have a greater inclination. This approach leads to the "bottle" shape for the taper. With the secondary waves reflected at the cut-off plane, the *sum* of their two phase constants determines the destructive interference; this leads, after integration, to total amplitudes which are always negligible when compared with the effect of the primary waves.

An important concept in connection with the analysis of tapers and other oversized waveguide components is the *beat wavelength* λ_b of two distinct modes, i.e. the length along which the phase difference of the two modes changes by 2π. In other words λ_b equals 2π divided by the difference of the two phase constants. To obtain sufficient destructive interference in a component, it should be long with respect to the beat wavelength λ_b. For two specific modes λ_b is dependent on the frequency and the dimensions a and b. Thus, for the modes H_{10} and H_{30}, $\lambda_b = 130$ mm for $f = 75$ GHz and standard 3-cm waveguide.

Because of the symmetrical structure of this kind of taper H_{mn} modes with m even or n odd cannot be excited, and therefore only the types H_{30}, H_{50}, . . . , H_{12}, H_{32}, . . . , H_{14}, H_{34}, . . . occur. Of these, the H_{30} mode usually has the largest amplitude, since the corresponding beat wavelength λ_b is by far the largest. The tendency towards destructive interference is therefore particularly small with this type of mode, and on this account

our taper has only been optimized for low $H_{10} \rightarrow H_{30}$ mode conversion.

The only part of the taper which is significant in mode conversion is the part where the unwanted higher mode can propagate. The rest of the taper can be designed more or less arbitrarily: we chose a linear function for $a(z)$. We made the variation of b with z linear throughout, since variation in height excites only the types $H_{12}, H_{32}, \ldots, H_{14}, H_{34}, \ldots$ which always have small amplitudes on account of their small beat wavelengths.

Bends

There are two different approaches to the design of the majority of the components for oversized waveguide. The first, used by most workers in this field [8] [17], is the "optical approach", in which the propagation of electromagnetic energy is assumed to take place essentially along "rays", which undergo reflection and refraction, as in geometrical optics. This approach is useful especially for a high degree of oversizing, since in this case the fundamental H_{10} wave behaves very much like a plane wave. However, there are some components such as tapers and mode filters which cannot be designed by using optical principles. It has been found very difficult to make these indispensible elements with low enough mode conversion if the waveguide is highly oversized. In the work described here we therefore used a different approach, suitable for a medium degree of oversizing. This is the "waveguide approach" in which the propagated wave is separated into the waveguide modes, between which conversions can take place.

H- and *E*-plane bends designed by the optical approach are shown in *fig. 14*. The incident wave, which may be represented by a "ray", falls upon a plane mirror at a certain angle and is reflected at the same angle towards the output port. However, some diffraction occurs, since in both waveguide arms a part of the wall has to be left out to permit the desired deflection of the ray. As the wall currents in the broad walls of a highly oversized waveguide are much higher than in the narrow walls, this diffraction is obviously more marked in the *E*-plane bend (on the left in fig. 14). In waveguide terms diffraction means excitation of higher-order modes; these belong to the E_{1n} and H_{1n} types for the *E*-plane bend and to the H_{m0} types for the *H*-plane bend. Although the E_{1n} and H_{1n} modes can be absorbed by using the mode filters described above, the overall losses become so excessive (several decibels) that the *E*-plane bend designed by the optical method is not very attractive. On the other hand, the *H*-plane bend gives satisfactory results provided the waveguide is sufficiently oversized. An extensive theoretical analysis has been given by Dr. C. J. Bouwkamp of this laboratory [18]. For instance, the power loss due to mode conversion into the H_{m0} modes has been shown to be less than 0.4 dB if the frequency exceeds eight times the cut-off frequency, in other words, if the waveguide is more than about five times oversized. However, it should be emphasized that the power loss in a complete system may be substantially higher because of trapped-mode resonances.

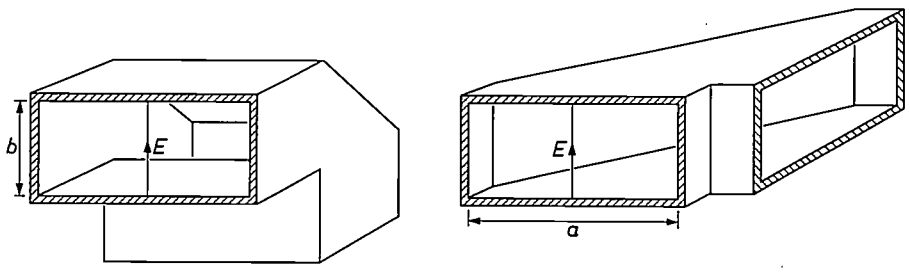
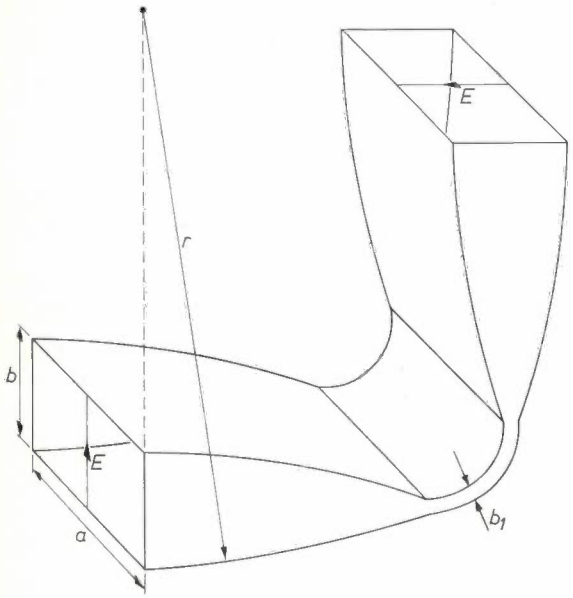


Fig. 14. The 90° *E*-plane bend (on the left) and the 90° *H*-plane bend (on the right), designed by the "optical approach", in which the wave propagation is assumed to take place along "rays".

Since the optical approach has proved to be unsatisfactory in the case of an *E*-plane bend, a solution based on the waveguide principle has been sought. In the arrangement shown in *fig. 15* the actual bend is made of flat waveguide, which is connected with the full-height oversized waveguides by means of two tapers. In the bend section there is some mode conversion into the E_{1n} and H_{1n} modes; but if the height of the flat waveguide does not exceed half the wavelength these modes are evanescent and do not appear at the ports. A bend of this type is usable from frequencies giving standard-size operation to those at which there is a high degree of overmoding, because the quantity determining the cut-off frequency of the desired mode, i.e. the width, does not change. The power loss, which is only of the order of a few tenths of a decibel, is mainly due to the wall losses in the flat guide.

[17] B. Z. Katsenelenbaum, Soviet Phys. Uspekhi **7**, 385-400, 1964/65; Radio Engng, Electronic Phys. **8**, 1098-1106, 1963.

[18] C. J. Bouwkamp, unpublished. See also: I. P. Kotik and A. N. Sivov, Radio Engng. Electronic Phys. **10**, 140-142, 1965.



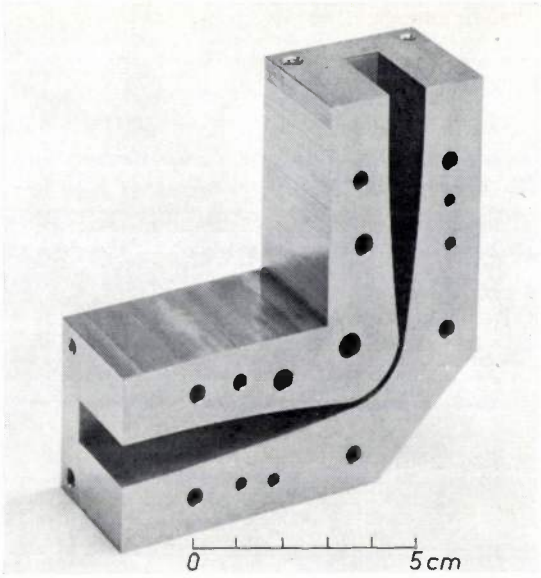


Fig. 15. The 90° *E*-plane bend designed from a "waveguide approach" in which the wave is split up into waveguide modes. Conversion into E_{1n} and H_{1n} can occur, but these modes do not propagate in the flat guide provided $b_1 < \lambda/2$. In a design for use at $\lambda \geqq 4$ mm the following dimensions were chosen: $a \times b = 22.86 \times 10.13$ mm, $r = 420$ mm; $b_1 = 2$ mm.

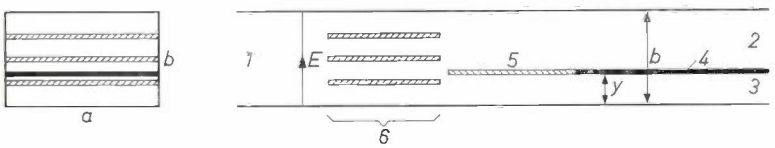


Fig. 17. Principle of a three-port directional coupler. At the left, transverse cross-section; on the right, longitudinal cross-section. Power from port *1* is divided in the ratio $(b - y) : y$ between the ports *2* and *3*. Ports *2* and *3* are decoupled. *4* metallic sheet, *5* semi-transparent resistive sheet, *6* higher-order mode filter.

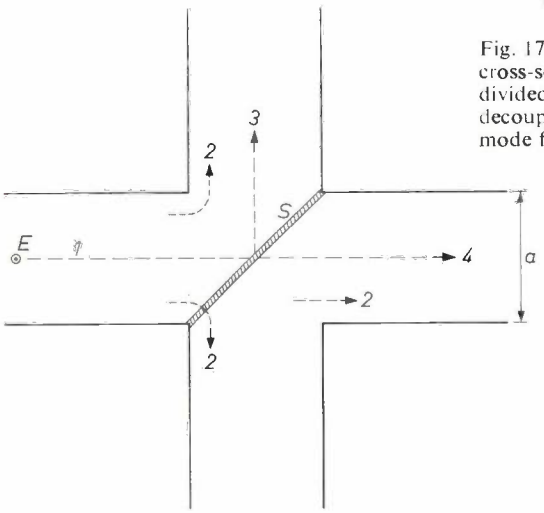


Fig. 16. Directional coupler designed from optical principles. *H* is in the plane of drawing. At the dielectric sheet *S* the incident wave *1* is partly reflected (*3*), partly transmitted (*4*). There is some unwanted coupling $1 \rightarrow 2$ owing to diffraction, even without any sheet.

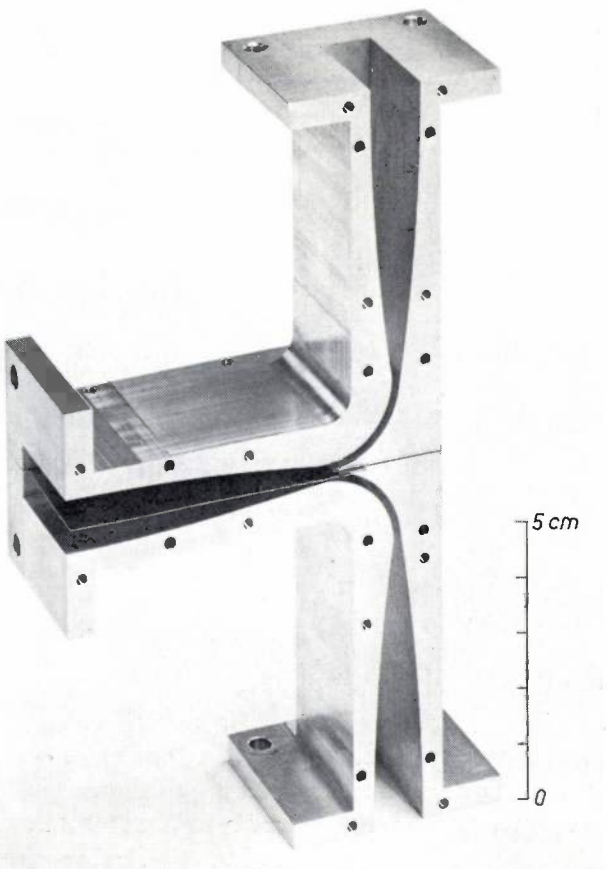


Fig. 18. Practical design of a 3-dB three-port directional coupler, from the principle of fig. 17.

Directional couplers

A well-known optical design for a directional coupler is shown in *fig. 16*. It consists of an *H*-plane cross-junction of two waveguides, with a dielectric sheet placed in the diagonal plane of the cross. This sheet acts as a semi-transparent mirror and provides a directional coupling, as in optical interferometers. However, even without any sheet, a fundamental mode incident in

one arm causes some radiation into the side arms owing to diffraction. The distribution of this deflected power as well as of the reflected and transmitted power among the different modes has been calculated by Katsenelenbaum [17] and by Bouwkamp [18]. These diffraction phenomena give rise to certain errors in the performance of the directional coupler which decrease with increasing degree of overmoding. Special care has to be taken if the directional coupler is used close to the cut-off frequency of a higher-order H_{m0} mode which may give resonances in a system.

If there is no special reason for using a four-port directional coupler, and a three-port type of directional coupler with built-in load is sufficient, the following arrangement based on the waveguide principle is to be preferred [19]. A waveguide is divided into two parts by a septum parallel to the broad walls (*fig. 17*). This septum consists of two parts: a metal sheet and a resistive film which is semi-transparent to microwaves. Power incident from the left (port *1*) is divided between waveguides *2* and *3* in the ratio of their heights. This ratio determines the coupling factor of the directional coupler. Port *2* does not couple to port *3*, because in two parallel guides separated by a homogeneous sheet, a wave propagating in one of the guides does not excite a wave propagating in the opposite direction in the other guide. The higher-order modes excited by the resistive sheet are attenuated by the mode filter at port *1*. The resistive sheets of this filter and the resistive part of the septum are equivalent to the built-in load used in conventional standard-size directional couplers. A more practical form of oversized waveguide directional coupler is given in *fig. 18*. With this construction a directivity of about 50 dB has been obtained for a coupling of 3 dB ($\lambda = 4$ mm, standard 3-cm waveguide). A great advantage of the present device is the frequency independence of the coupling, which is solely determined by the physical dimensions. Like the *E*-plane bend shown in fig. 15 the directional coupler is inherently broadband.

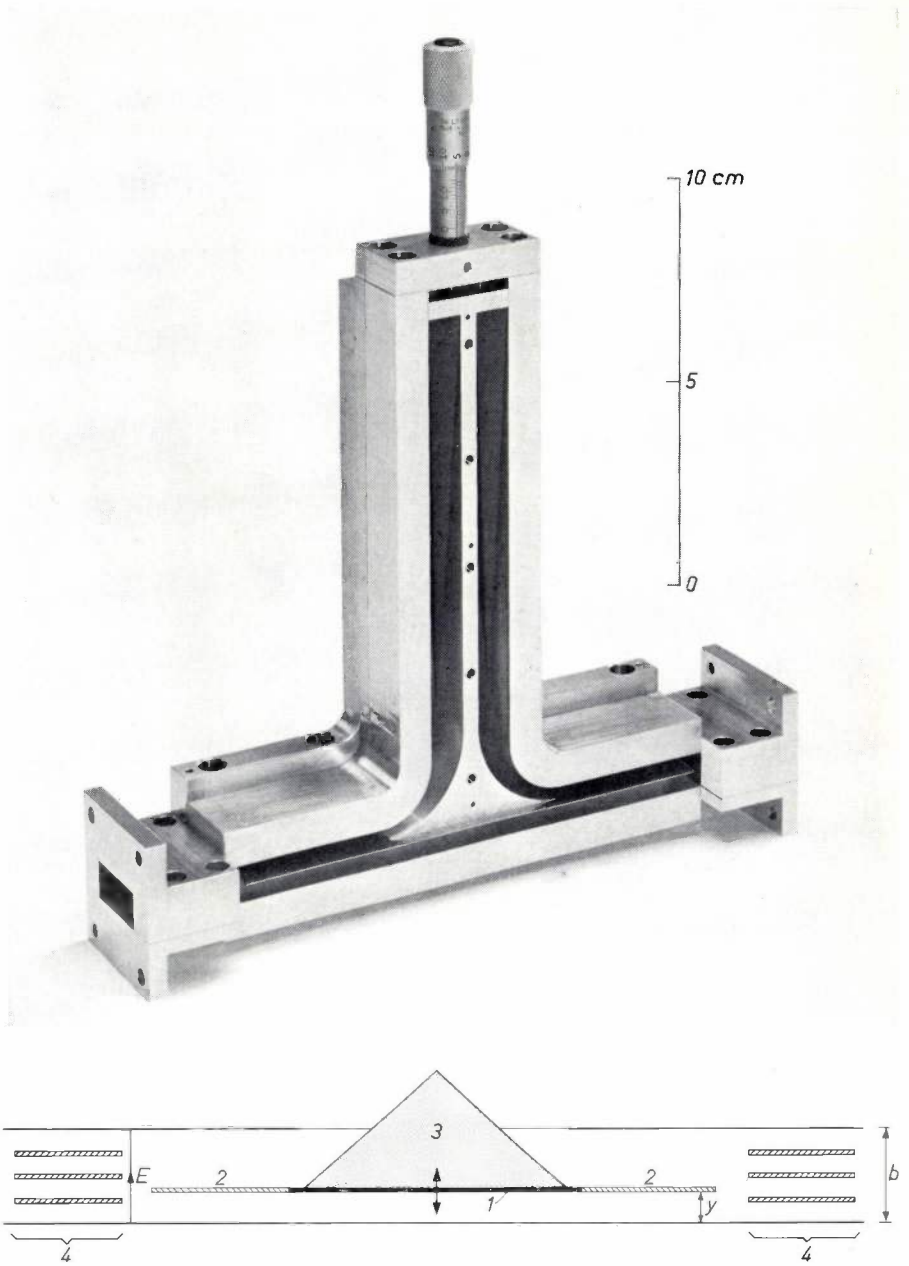


Fig. 19. Variable attenuator derived from the directional coupler of fig. 17. The guide is again divided by a metallic sheet *1* and resistive sheets *2*. Power in the upper guide is absorbed in the matched load *3*. The attenuation is varied by vertical displacement of the unit consisting of the sheets *1* and *2* and the load *3*. Higher-order modes are filtered out, as before, by the mode filters *4*.

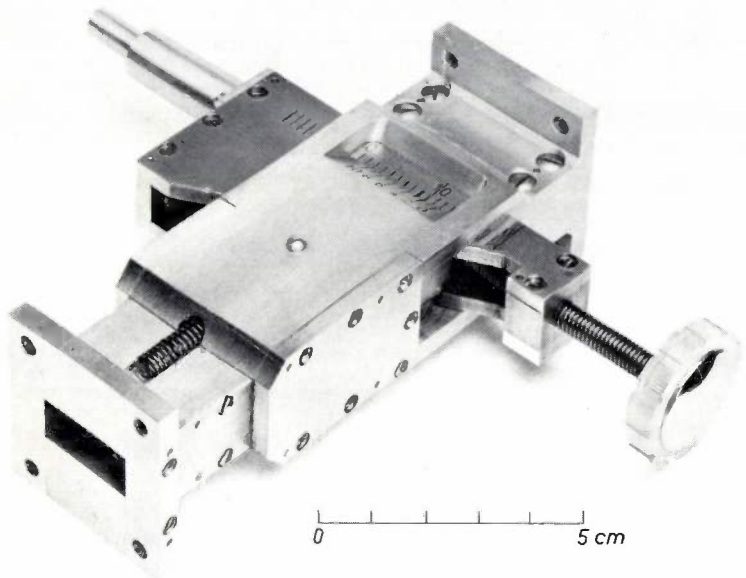
Some variable components

In principle the directional coupler described above

[19] S. B. Cohn, Microwaves **5**, No. 6, 62, 1966; G. W. Epprecht, International Microwave Symposium Digest, Boston, May 1967, p. 10; W. K. Kahn, *ibid.* p. 54-57.

can also be used as a frequency-independent *attenuator*, if one arm is terminated in a matched load. In *fig. 19* two such directional couplers are connected symmetrically in cascade, so that the input and output waveguides are full-height again. The device can easily be made into a *variable attenuator* by making the septum movable. As each of the directional couplers with one arm matched is a reciprocal two-port the total attenuation (in dB) is theoretically equal to twice the coupling (in dB) of each directional coupler; it is given by (cf. fig. 19):

$$\text{attenuation} = 20 \log (b/y). \quad . \quad . \quad . \quad (11)$$



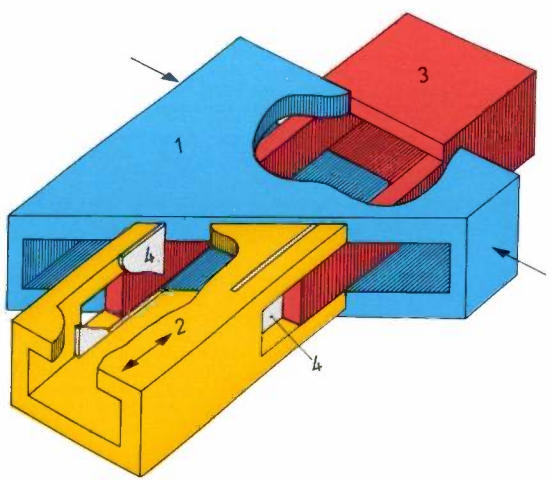


Fig. 20. Variable phase shifter (line stretcher). The phase shift between the two ends of a waveguide is varied by varying the length of the line. By turning the knob the part *1* is made to slide in the transverse direction, thereby causing part *2* of the waveguide to slide in the longitudinal direction over the fixed part *3*. The top and bottom of the waveguide are partly formed by *1*. *4* is a thin metal sheet covering the hole of varying size between *2* and *3*.

Measurements ($\lambda = 4$ mm, standard 3-cm waveguide) have indicated a minimum attenuation of 0.5 dB, and a maximum attenuation of 50 dB.

Fig. 20 shows a simple construction of a *variable phase-shifter*; some of the details are explained in the caption. This "line stretcher" can be used in all cases where the variation of its total length causes no difficulty.

A variable phase-shifter and a variable attenuator connected in cascade and terminated in a short-circuit constitute a *variable impedance*. Since in this case the output port need not be accessible, the arrangement can be simplified by using half of the attenuator shown in fig. 19. This gives the configuration of *fig. 21* where the metallic sheet of the directional coupler has clearly become superfluous. The magnitude of the reflection coefficient can be varied by vertical displacement of the short-circuiting part together with the resistive sheet and the load: its value is simply given by y/b. The phase is varied by the line-stretcher described above, which is placed in front of the present device.

Conclusions

Now that we have described the particular features of several special components at some length, let us summarize some principal aspects of the design of single-mode oversized waveguide components. We have seen that the components designed from optical principles work satisfactorily only if the waveguide is sufficiently oversized. If, on the other hand, the waveguide is more than about ten times oversized, the tapers and mode filters — indispensable components for most applications — have to be inconveniently long. The quasi-optical components will therefore be usable only in a limited frequency interval, roughly characterized in practice by the limits "five times oversized" and "ten times oversized", corresponding to an octave. On the other hand, some of the components described in this paper and designed from "waveguide" principles, work over a frequency decade, that is to say, between the frequencies corresponding to "not oversized" and "ten times oversized".

Almost all components give rise to some mode conversion into unwanted modes. This phenomenon is annoying not only because of the loss in the power transmitted in the fundamental mode, but also because

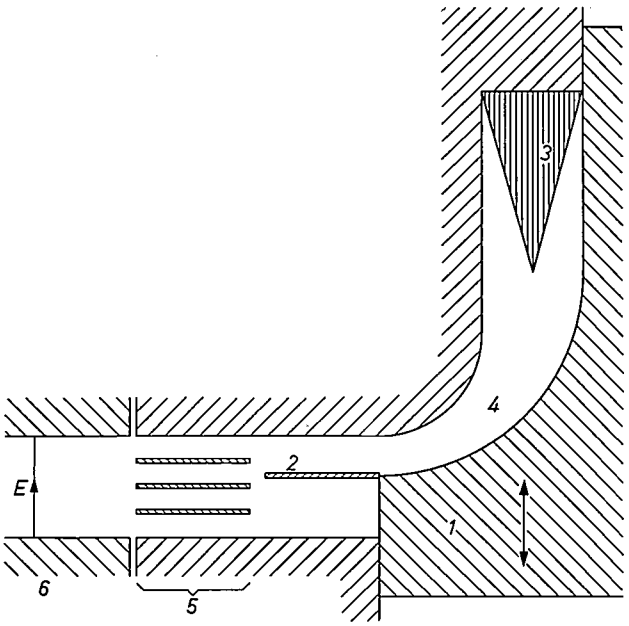


Fig. 21. Termination giving a reflection of variable magnitude and phase. This device, together with the variable phase shifter of fig. 20 (*6* in this figure) constitutes a variable impedance. The magnitude of the reflection is varied by vertical displacement of the part *1* together with the resistive sheet *2*. *3* is a load, *4* a 90° *E*-plane bend, both reflectionless. *5* mode filter. The magnitude of the reflection coefficient is equal to y/b (cf. fig. 19).

of possible interactions between the higher-order modes excited by different components and, in particular, the possible occurrence of trapped-mode resonances. On account of these effects mode filters have to be used which suppress or at least reduce these interactions and resonances. Since for the higher-order modes of class 1 (E_{mn} and H_{mn} modes with $n \neq 0$) almost ideal filters can be constructed in the form of resistive sheets, the preferred components should be those which only give conversion into these modes. This is the case with the *E*-plane bend of fig. 15 and the directional coupler of fig. 18. These components have the common characteristic that all longitudinal cross-sections parallel to the narrow wall are identical. Furthermore, they are inherently broadband because the only limitation to the frequency of the dominant mode to be transmitted is a lower limit: the cut-off frequency determined by the width of the waveguide.

Tolerances in the form and dimensions of components have not been discussed in this article. Due attention has to be paid to precision manufacturing of the components and careful alignment in a system, as any deviation from ideal dimensions, any discontinuity at the flanges or inhomogeneity along the guide, can in principle give rise to higher-order modes.

Finally, let us enumerate some of the applications of oversized rectangular waveguide equipment:

1) Short-distance transmission, e.g. in radar aerial feeds.
2) Measurements of physical properties of materials such as dielectric constant, conductivity, surface impedance.
3) Simultaneous transmission of microwave signals of several frequencies covering a wide band (multiplex).
4) High-power transmission. This application, however, may be limited by breakdown or excessive rise of temperature in some of the components described in this paper.
5) It may also be possible to apply the oversized rectangular waveguide method to the submillimetre region, by scaling down the components.

In applications such as these the rectangular waveguide has proved to be superior to the circular waveguide operated in the H_{01} mode. A serious drawback of the H_{01} circular mode is that it is very difficult to design mode-launchers and filters, which means that its use is justified only if full advantage can be taken of the extremely low loss in this mode. Mode control in the circular guide is rather difficult, particularly since the H_{01} mode in this guide is degenerate with the E_{11} mode, so that mode conversion easily takes place.

Summary. The use of oversized waveguide for the transmission of millimetre waves has advantages over standard-size waveguide: the larger guide is easier to manufacture and to handle, the losses are smaller, and very broadband components can be made. For measuring purposes in particular, single-mode operation of the oversized waveguide is essential, as with multimode operation the field pattern is in practice unpredictable. If single-mode operation is to be achieved, mode control is necessary, i.e. measures have to be taken to prevent the conversion of the fundamental mode into higher-order modes, or to eliminate the effects of such conversion. An essential element for examining the mode conversion of components is the mode transducer, which converts a specific higher-order mode into the fundamental mode. Another indispensable element for mode control is the mode filter, which ideally absorbs all the higher modes without attenuating the fundamental mode. Several mode transducers and filters are described. Some of the oversized waveguide components discussed: tapers, *E*-plane bends, directional couplers, variable attenuators, phase shifters and impedances can be used over a 10 to 1 frequency band.

Stereo drawings made with a digital computer

I. Making stereo drawings with the COBRA-controlled drawing machine

II. Stereographic presentation of the "Duphaston" molecule

III. A solution of the Van der Pol equation, represented in three dimensions

When a computer is used to calculate a three-dimensional figure, the results usually have to be studied in the form of a number of projections drawn by an x-y plotter. A much more complete picture is obtained, however, if the figure can be seen in three dimensions, preferably from different sides. This is illustrated in the following three short articles. The first article describes how a numerically-controlled drawing machine, starting with the co-ordinates of the figure (determined, for example, by the computer), can make a pair of stereo drawings by first computing the required projections. The technique is shown applied to the design of a television picture tube. The second article gives another example of stereographic presentation: a large number of drawings are made of a molecule whose structure has been determined by a computer from X-ray diffraction data; these drawings, put together in a film, give the impression that the molecule is rotating, and combined in pairs they also provide a three-dimensional picture. Here the projections are calculated by the computer itself, and the drawings are then made by an x-y plotter. As the last example the third article discusses the stereographic representation of a solution of Van der Pol's non-linear equation. All the examples are illustrated by anaglyphs.

I. Making stereo drawings with the COBRA-controlled drawing machine

G. C. M. Schoenaker

Introduction

Drawing and engraving machines are now being extensively used for making workpieces of widely different types, such as photographic masks for integrated circuits and printed wiring, and also drawings for checking the design of a product or of a particular stage in its manufacture. An example of the latter is the checking of punched tapes for numerically-controlled machine tools such as milling machines. The drawing machine in this case draws the projections of the cutter path on the three co-ordinate planes; any errors in the programme can then be seen. The control system for such a drawing machine is usually of the same type as used for the control of machine tools.

We have recently described in this journal a small digital computer, the COBRA, which was designed specially for use as a universal control system for machine tools, and which can also readily be used for controlling a drawing machine [1]. For making check drawings of the type mentioned above, the COBRA has two important advantages over conventional control systems. Firstly the COBRA-controlled drawing machine can in principle check the punched tapes for all the existing control systems, whereas most conventional systems can check only those punched tapes that have been made to be processed by the same type of control

Ir. G. C. M. Schoenaker is with Philips Research Laboratories, Eindhoven.

system. Secondly, without requiring the help of a computer, this combination can not only produce drawings of the orthogonal projections on the co-ordinate planes, it can also produce drawings of orthogonal, oblique and perspective projections on any arbitrary plane, resulting in a much clearer picture of the cutter path. Moreover, a stereographic presentation can then be obtained from two perspective projections, which is often a substantial improvement. We shall now deal briefly with the operation of the COBRA before discussing the making of check drawings with the COBRA-controlled drawing machine.

Drawing with the COBRA

The COBRA is a small computer of conventional construction, comprising a store (1024 addresses of 24 bits), an arithmetic unit and an internal control unit. To enable it to control a given machine tool, the COBRA is supplied with the necessary instructions in the form of a *control programme* which is placed in its store (stored logic). The data for the workpiece to be produced, forming the *workpiece programme*, are fed to the COBRA by means of punched tape during the machining operation. The COBRA thus constitutes a very flexible system: it is possible to switch over to the control of another machine tool very quickly by putting a different control programme, which is stored on punched paper tape into the store: this programme need be set up only once for any given machine tool.

To make a workpiece programme the required tool path is calculated from the given shape of the workpiece; the data that have to be fed to the COBRA in the workpiece programme are derived from these calculations. In milling a cam, for example, one needs to know various points of the tool path and certain information about the cutter and the cutter speed. The COBRA then computes the path to be followed by the cutter by interpolation between the given points, the computation being done during the machining. For making a workpiece programme it is often necessary to make use of a large computer, and in this case the programming language used is generally the APT language [2].

The COBRA works in the same way when used as a control unit for a drawing machine. During the several years that the COBRA has been used in combination with a drawing machine (of the Aristomat type) at the Philips laboratories, the making of drawings for checking punched tapes for numerically-controlled machine tools has been an important part of its work. We shall take as an example the various stages in the design and manufacture of a new type of television picture tube. The starting point is set by various requirements, some of them technical and others concerned with industrial design. From these requirements and a number of fixed data, a computer determines the shape of a tube that meets these requirements. The next step is to judge from drawings whether the processing by the computer has resulted in any unexpected effects, e.g. dents or flattenings in the shape of the tube.

In the next stage the workpiece programme is made for the production of two hard-metal moulds for pressing the tube. Before this operation is carried out it is first necessary to check the punched tape containing the workpiece programme for errors. For this check the punched tape is fed to the COBRA-controlled drawing machine, which then produces a projection of the paths to be followed by the centre of the cutter while the programme is being run (*fig. 1*). Any errors

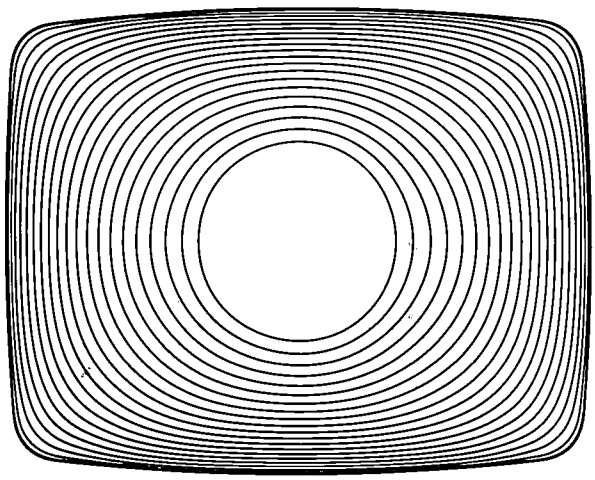


Fig. 1. Check drawing of the workpiece programme used for milling one of the moulds for pressing a television picture tube. The drawing represents an orthogonal projection on the x,y-plane of various paths which the cutter centre will follow during the milling operation.

in the programme now become apparent. The control programme stored in the COBRA has to enable this combination of computer and drawing machine to give a complete simulation of the numerically-controlled milling machine that will be used for production. The drawing of these paths calls for great accuracy; the drawing and engraving machine used in combination with the COBRA has an absolute accuracy of ± 25 μm over a working area of 85×85 cm and a reproducibility of 10 μm, which is more than adequate.

For a three-dimensional workpiece the check drawings consist as a rule of the orthogonal projections on the three co-ordinate planes. In some cases, for example that of fig. 1, this provides sufficient information

[1] R. Ch. van Ommering and G. C. M. Schoenaker, The COBRA, a small digital computer for numerical control of machine tools, Philips tech. Rev. **27**, 285-297, 1966.

[2] J. Vlietstra, The APT programming language for the numerical control of machine tools, Philips tech. Rev. **28**, 329-335, 1967 (No. 11).

about the correctness of the data on the punched tape. Sometimes, however, it is very difficult to pick out the cutter path in these projections. This is particularly so in the case of a workpiece of complex shape made on a three-axis milling machine, where only one slide can move at a time during the milling operation, so that the cutter is displaced either in the x-, the y- or the z-direction. The projection drawings then give a large number of straight lines in the x-, y- and z-directions, from which it is difficult, if not impossible, to determine the cutter path (*fig. 2*).

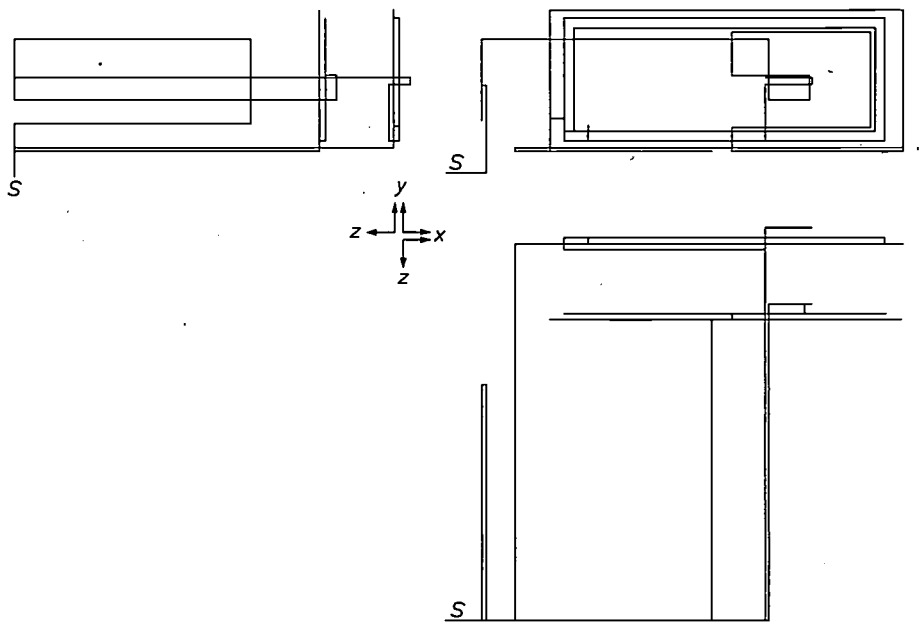


Fig. 2. Check drawing of a workpiece programme for milling a workpiece on a three-axis milling machine, where the displacement of the cutter during the operation always takes place along one axis only. The drawing represents the orthogonal projections on the three co-ordinate planes; S is the starting point of the cutter path. From drawings of this kind it is very difficult to determine the cutter path. (This workpiece was a frame for an instrument.)

We noted earlier that a clearer picture of the cutter paths is obtained if in addition to the orthogonal projections on the co-ordinate planes, check drawings can also be produced as orthogonal, oblique or perspective projections on an arbitrary plane. The additional calculations needed for drawing these projections when x-, y- and z-co-ordinates are given of particular points of a workpiece, or of a tool path, can also be carried out by the COBRA. With the control programme set up for this purpose the COBRA-controlled drawing machine can produce, for example, a perspective drawing from the data of the normal workpiece programme which will later be used for manufacturing the workpiece. We shall now first consider the calculations required for making a perspective projection.

The perspective projection

To make a perspective projection, in which an object is seen from one particular point, the object has to be projected from that point, called the centre of projection, on to a plane which is perpendicular to the "viewing direction". If the object is to be seen from the resultant drawing in the correct perspective, it is advisable to make this plane lie within the object or close to it.

A situation of this kind is shown in *fig. 3*. The position of the centre of projection C in the co-ordinate system is determined by the angles α and β and the distance $OC = r$. The projection plane is the plane through the origin O of the co-ordinate system x,y,z perpendicular to the line CO. We now define a second co-ordinate system u,v,w, whose origin is also at O, and whose w-axis coincides with the line CO as the u-axis lies in the x,y-plane. The chosen projection plane thus coincides with the u,v-plane. We consider point P of the object, which is projected from C to a point P' on the u,v-plane. We shall now determine the relation between the u- and the v-co-ordinate of P' and the given co-ordinates of P and C.

The calculation goes as follows. A perpendicular is drawn from P to the u,v-plane. The foot P'' of this perpendicular then lies on OP', and the length PP''

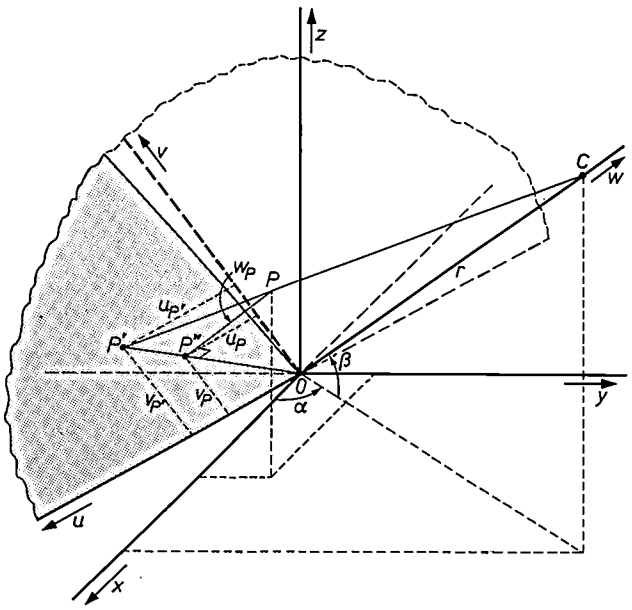


Fig. 3. Diagram for computing the perspective projection P' of a point P seen from the centre of projection C. The projection plane is the plane through O perpendicular to the line CO. The w-axis of a co-ordinate system u,v,w coincides with CO, the u,v-plane with the projection plane. The u,v-co-ordinates of P' can be expressed in terms of α, β, r and the x,y,z-co-ordinates of P.

is equal to the w-co-ordinate of the point P. In the right-angled triangle COP' we now have:

$$\frac{OP'}{OP''} = \frac{r}{r - w_P}. \qquad \ldots\ldots (1)$$

With the help of (1) we can now express the u- and the v-co-ordinates of P' in terms of the u,v,w-co-ordinates of P:

$$\frac{u_{P'}}{u_P} = \frac{v_{P'}}{v_P} = \frac{r}{r - w_P}. \qquad \ldots\ldots (2)$$

By means of a simple linear transformation the u,v,w-co-ordinates of P can be calculated from the given x,y,z-co-ordinates, and the co-ordinates of P' then follow from (2). The transformation can be rather elegantly written as a multiplication of the vector of the point P by a matrix M whose elements are determined by the angles α and β. If is $\mathbf{p}$ the row vector (x_P,y_P,z_P) and $\mathbf{p}'$ is the row vector $(u_{P'},v_{P'})$, we can write:

$$\mathbf{p}' = \frac{r}{r - w_P}\,\mathbf{p}M, \qquad \ldots\ldots (3)$$

where

$$w_P = \mathbf{pm}. \qquad \ldots\ldots (4)$$

Here $\mathbf{m}$ is a column vector whose elements are likewise determined by α and β.

For completeness we mention here the form of the matrix and the column vector:

$$M = \begin{pmatrix} -\sin\alpha & -\cos\alpha\sin\beta \\ \cos\alpha & -\sin\alpha\sin\beta \\ 0 & \cos\beta \end{pmatrix},$$

$$\mathbf{m} = \begin{pmatrix} \cos\alpha\cos\beta \\ \sin\alpha\cos\beta \\ \sin\beta \end{pmatrix}.$$

The calculations using equations (3) and (4) include a number of multiplications and divisions and can thus be carried out entirely by the COBRA. This does need a relatively long computing time (about 200 milliseconds), since the COBRA operates as a normal, though simple, computer and carries out the operations as a series of instructions, such as "add", "shift" etc.

In drawing a perspective projection it would in principle be necessary to carry out the calculation for every point of the figure to be projected. However, as we saw above, equation (3) represents a linear transformation, which means that for linear motions of the cutter only one calculation is needed for every line element. Since in programming the cutter paths (e.g. with the APT programming language) the curved lines are approximated by a large number of straight-line elements, the number of calculations is limited. This is important in view of the relatively long computing time needed per point; an unduly large number of calculations would require the drawing machine to be slowed down, since the COBRA would not have the commands or the drawing machinefavailable in time.

To control a drawing machine the COBRA has to be provided with a control programme containing subroutines for carrying out linear interpolation, for the automatic acceleration and deceleration of the slides, for working with various scale factors in the x- and y-directions, and so on. For drawing the projections a programme is drawn up which in addition to these subroutines contains the subroutines for carrying out the transformation calculations in accordance with equations (3) and (4). The situation of the centre of projection C can still be selected at will. When the control programme has been put into the COBRA store, the data for the centre of projection must therefore be fed in before drawing can begin. The punched tape containing the data for the drawing (the workpiece programme) is then fed in once the process is under way. These extra data — the distance r and the elements of the matrix M and the vector $\mathbf{m}$ — can be put on to a punched tape within a few minutes by means of a conventional tape punch. If a number of these punched tapes (each about 9 cm long) are held in reserve an object can thus quickly be drawn from various viewpoints.

As an example, based on the punched tape used for the drawing in fig. 1, a perspective drawing was made of the cutter paths for one of the moulds for a television picture tube (*fig. 4*). The tube is of the 22 inch type, that is to say the diagonal of the face of the tube is 22 inch long. We took as the centre of projection a point for which $\alpha = \beta = 45°$ and $r = 64$ cm. The actual drawing was made at full scale, but is shown here on a reduced scale; the tube will appear in the right perspective if fig. 4 is viewed from a distance of about 20 cm. As a second example *fig. 5* shows a perspective drawing made with the workpiece programme of fig. 2. Here again, the centre-of-projection data are $\alpha = \beta = 45°$ and $r = 64$ cm. The cutter path can be picked out much more easily from this figure than from fig. 2.

In some cases the distortion occurring in a perspective projection can be a disadvantage. It may make it impossible, for instance, to measure certain important dimensions in the check drawing. In such cases a compromise can be obtained between the different requirements by drawing an *oblique parallel projection*, i.e. a parallel projection on one of the co-ordinate planes with the projection lines making an angle with that

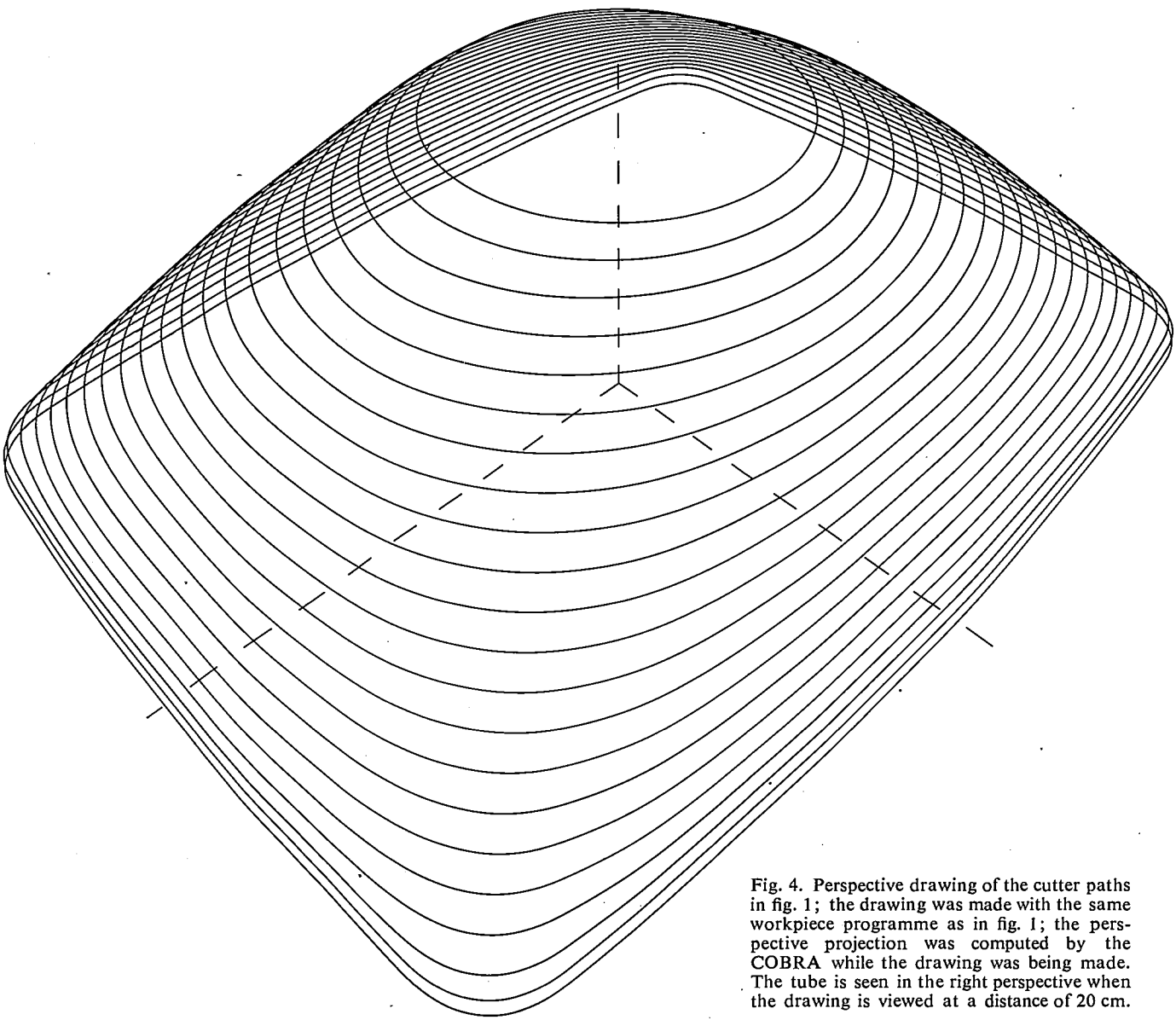


Fig. 4. Perspective drawing of the cutter paths in fig. 1; the drawing was made with the same workpiece programme as in fig. 1; the perspective projection was computed by the COBRA while the drawing was being made. The tube is seen in the right perspective when the drawing is viewed at a distance of 20 cm.

plane. If we take the projection angle as 45°, not only do the dimensions in the relevant co-ordinate plane remain unchanged; those in the third (obliquely drawn) co-ordinate direction also remain unchanged. All dimensions in the three co-ordinate directions can now be checked in the drawing while at the same time the cutter path can still be picked out fairly clearly from this projection. *Fig. 6* shows such a projection for the cutter path of fig. 5.

The oblique parallel projection, like the orthogonal projection, can be made with the same control programme as the perspective projection; all that is needed for this purpose is to supply the COBRA with a different matrix M and vector $\mathbf{m}$. In fact one should also put $r = \infty$. This is of course impossible, but in equations (3) and (4) the same effect can be achieved by substituting the zero vector for $\mathbf{m}$.

Stereographic presentation

If we arrange things so that the drawing machine draws *two* perspective projections of a three-dimensional object, seen from two different centres of projection, these drawings form a stereoscopic pair where the centres are taken as positions of the eyes. To obtain a natural three-dimensional impression the centres of projection should of course be situated at places which really could correspond to the positions of the viewer's eyes. If we look at these drawings in a stereoscope we receive the same impression of depth as the observer would receive from the real thing. The improvement which stereoscopic viewing offers when compared with the viewing of a single perspective projection is of special interest in forming an opinion of the design of an object. Here it is essential to have the clearest possible idea of the final result at an early stage. By varying certain parameters one can then modify the shape of the object to meet specific requirements. This method is in fact used for such complicated objects as car bodies, where of course the visual impression is of great importance.

As an illustration, *fig. 7* shows a stereo pair in the form of an anaglyph. The pair consists of two projections of the 22 inch television picture tube of fig. 1 and fig. 4. The two perspective projections were drawn with the aid of the same tape containing the workpiece data, but the tube is seen from a different viewpoint from that of fig. 4. The dimensions of the tube were halved by the COBRA before the perspective transformations were computed (so that in fact an 11 inch tube is represented). The centres of projection used are two points at a distance equal to the distance between the eyes and situated on both sides of a point M for which $r = 64$ cm, $\alpha = 60°$ and $\beta = 30°$. The projections which were drawn are shown in fig. 7 at the true scale. If we look at this figure from a distance of about 64 cm we therefore see an 11 inch tube in natural perspective. We should have the same impression if we viewed the 22 inch tube from a distance of 128 cm, but with twice the distance between the eyes.

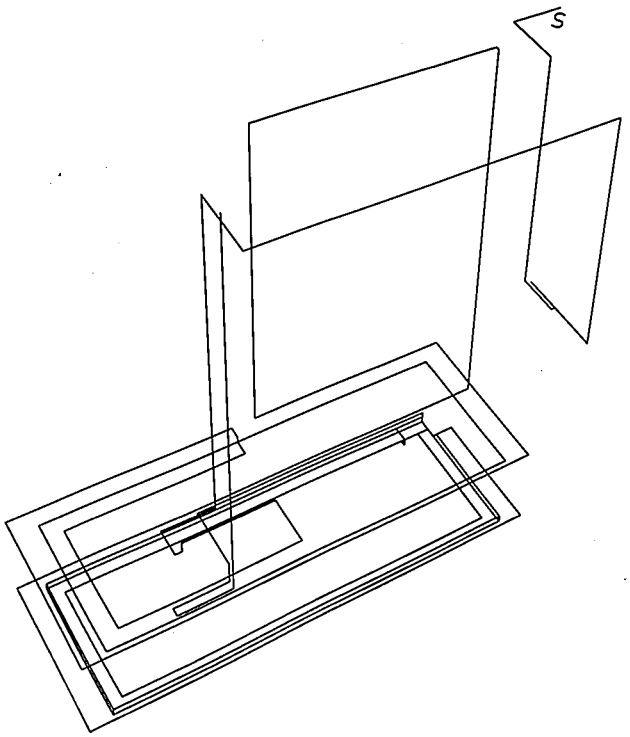


Fig. 5. Perspective drawing of the cutter path in fig. 2, made with the same workpiece programme. The cutter path here can be followed clearly; S is the starting point of the path. The dimensions in this figure are of course considerably distorted. The path is seen in the right perspective when the drawing is viewed from a distance of about 20 cm.

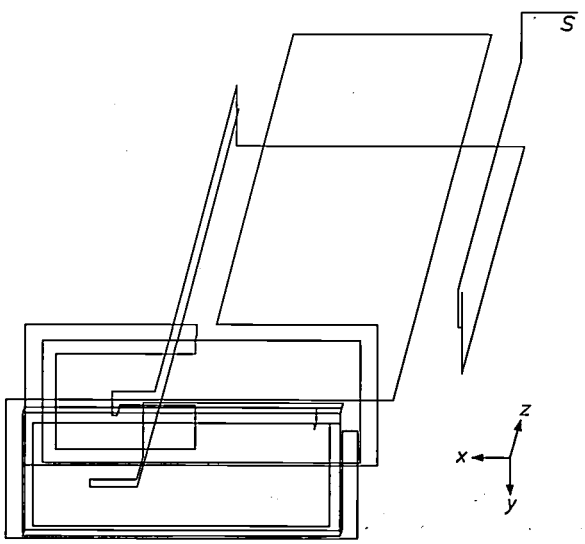


Fig. 6. Oblique projection of the cutter path of fig. 5. S starting point. This drawing has the advantage that the dimensions correspond to the actual dimensions not only in the x,y-plane, but also in the z-direction. The cutter path is not so easy to follow, however, as in fig. 5.

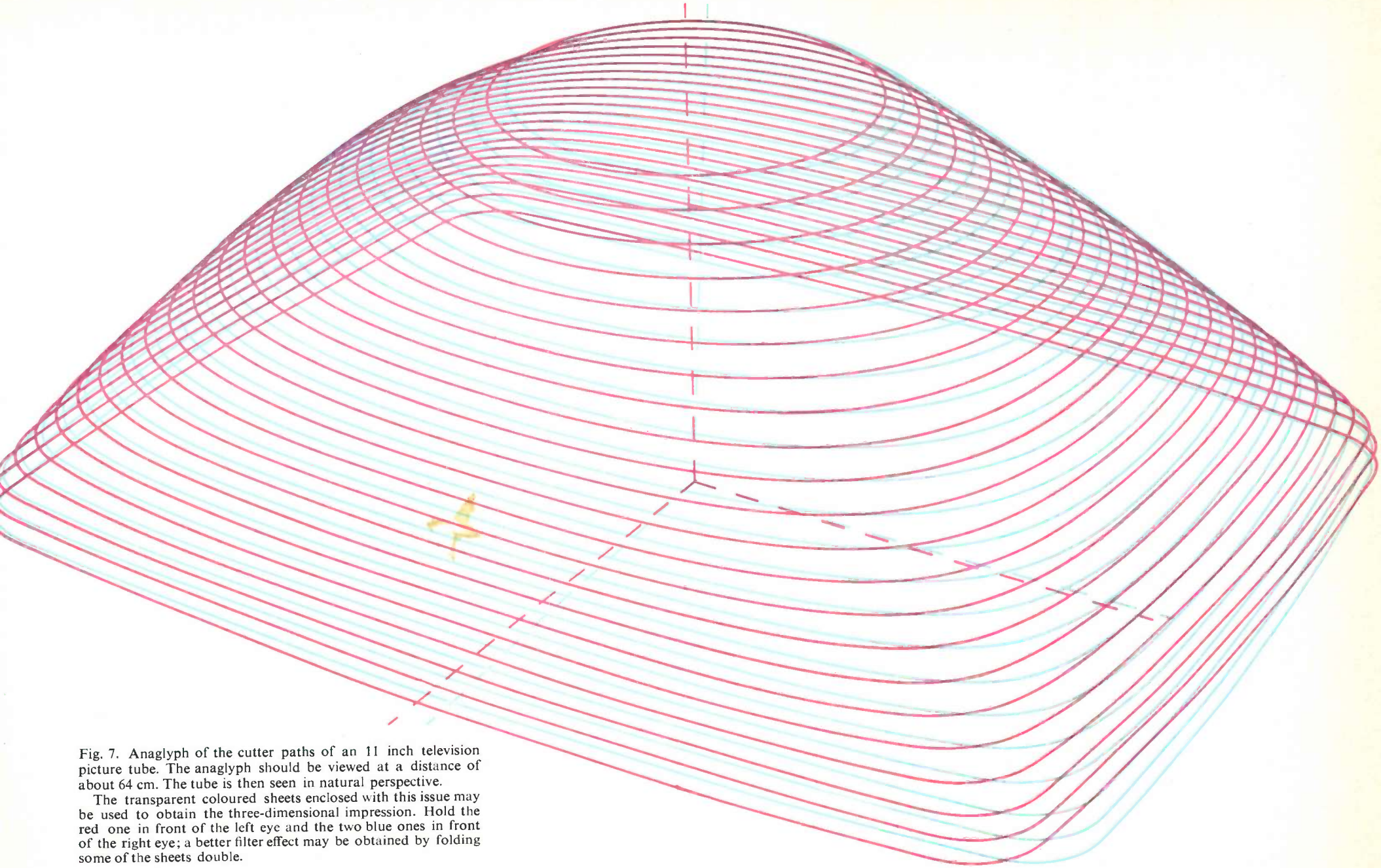


Fig. 7. Anaglyph of the cutter paths of an 11 inch television picture tube. The anaglyph should be viewed at a distance of about 64 cm. The tube is then seen in natural perspective.

The transparent coloured sheets enclosed with this issue may be used to obtain the three-dimensional impression. Hold the red one in front of the left eye and the two blue ones in front of the right eye; a better filter effect may be obtained by folding some of the sheets double.

If it is intended to give a stereo pair in the form of an anaglyph, the calculation of the perspective projections must be slightly modified. The projection plane in the method used in fig. 3 is perpendicular to CO. If we calculate the projections for a stereo pair in this way, then on viewing each drawing must be located so that the line connecting O to the appropriate eye is perpendicular to the plane of the drawing. If the drawings are viewed in a stereoscope, this requirement is fulfilled (*fig. 8a*). In this arrangement the optical axes should converge to a point (e.g. O' in fig. 8*a*) in the region where the object is observed, thus enabling the eye to accommodate to this region without difficulty. This accommodation and the convergence of the optical axes are both of great importance if an optimum three-dimensional impression is to be obtained.

When viewing an anaglyph the situation is somewhat different, both projections then being in the same plane, i.e. the plane of the drawing, which is at right angles to OM, the perpendicular bisector of the connecting line between the eyes (fig. 8*b*). In calculating the two projections the corresponding plane must therefore be taken as the projection plane, that is to say the plane through O which is at right angles to the line connecting O with the centre of the line connecting the two centres of projection. Provided the anaglyph is viewed from the right distance, the above-mentioned requirements for convergence and accommodation are then satisfied.

In this case equation (4) for the perspective transformation takes the form:

$$\mathbf{p}' = (\pm a, 0) + \frac{r}{r - w_P}\left[\mathbf{p}M - (\pm a, 0)\right], \quad \ldots \quad (6)$$

where $2a$ is the distance between the centres of projection (equal to the distance between the eyes); the plus sign relates to the projection intended for the right eye, and the minus sign to the other projection.

Obviously, the calculations needed for drawing the various projections can also be carried out quite easily by a large computer in a computer centre. This is in fact done in many places [3]. The advantage of the method described above as compared with the use of a normal computer is that, provided the various punched tapes for the required centre-of-projection transformations are available, the different drawings can be made with the use of only one punched tape containing the x,y,z-co-ordinates of the object. If the drawings are for checking a workpiece programme, a further great advantage of this method is that all it needs is the punched tape which will be used in a later stage for manufacturing the workpiece.

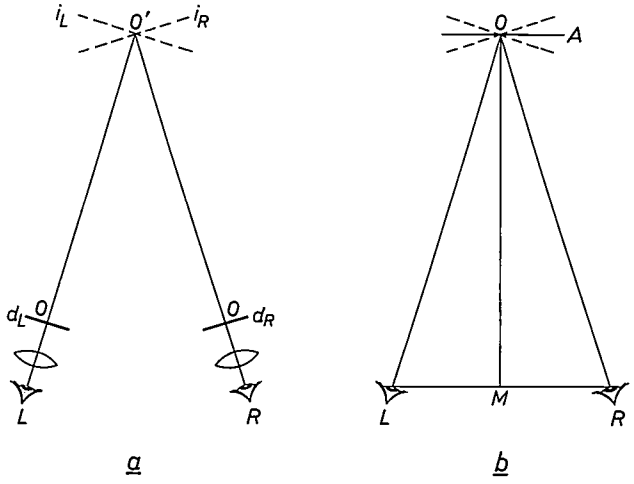


Fig. 8. *a*) Schematic diagram of a stereoscope. L and R are the viewer's eyes. The stereo drawings d_L and d_R are observed with the aid of lenses; the eye sees images of these projections at i_L and i_R. *b*) Arrangement of the eyes L and R when viewing the anaglyph A.

[3] See R. Hartwig, Graphische Datenverarbeitung mit Präzisions-Zeichenanlagen, Elektron. Rechenanl. **9**, 185-193, 1967 (No. 4);
A. M. Noll, Stereographic projections by digital computer, Computers and Automation **14**, No. 5, 32-34, 1965;
G. Horn, Zeichnungen — maschinell hergestellt, Elektrotechn. Z. **B 19**, 65-69, 1967 (No. 3).

Summary. The COBRA, a small digital computer specially designed for the numerical control of machine tools, can also be used as the control unit for an automatic drawing machine (for instance of the Aristomat type). In this application the co-ordinates of a number of points of the object to be drawn are fed to the COBRA on punched tape. Three-dimensional objects can be produced by drawing the orthogonal projections on the three co-ordinate planes, but the COBRA can also be programmed to produce a perspective drawing for any arbitrary centre of projection. Starting from the punched tape with the given x,y,z-co-ordinates, the COBRA then computes the co-ordinates of the desired projection while the figure is being drawn. Oblique and orthogonal projections on an arbitrary plane can also be drawn in this way. Two perspective drawings with suitably chosen centres of projection can be combined to form a stereoscopic pair. The article describes the calculations which the COBRA has to carry out and discusses as an application the making of drawings for checking punched tapes to be used for numerically-controlled production. It is shown how the path traversed by the cutter when a complex workpiece is being made can be followed clearly from a perspective drawing: this is not always possible from three orthogonal projections. The usefulness of stereographic presentation is illustrated by an anaglyph of a stereo pair of drawings of a mould for pressing a television picture tube.

II. Stereographic presentation of the "Duphaston" molecule

J. I. Leenhouts

In the investigation of a chemical compound it is often necessary to determine not only the structural formula but also the three-dimensional structure of the molecule, since this has an important influence on the properties of the substance, in particular its physiological action. The three-dimensional structure of the molecule can be determined with the aid of X-ray diffraction data [1], and for this purpose the electronic computers nowadays available are an invaluable help. The procedure used at the Philips laboratories in Eindhoven will shortly be described elsewhere [2]; here we shall deal only with the potential offered by the computer for clear and rapid visual presentation of the results.

As an example we take the retro-steroid 6-dehydro-retro-progesterone, or dydrogesterone [3], a pharmaceutical marketed under the name of "Duphaston" [*]. *Fig. 1* shows the structural formula of this substance.

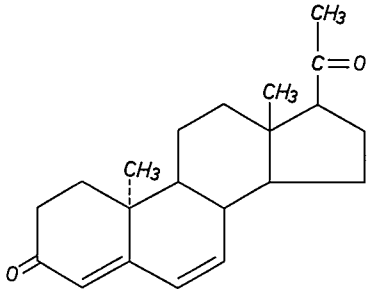


Fig. 1. The structural formula of "Duphaston".

We determined the three-dimensional arrangement of the atoms in the molecule from the X-ray diffraction data with the aid of the CDC 3600 computer, and in the course of the calculations the computer also determined from the interatomic distances which of the atoms were linked by chemical bonds We then used the same computer for drawing a number of orthogonal parallel projections, displaying the molecule from many sides. After each projection the direction of projection was rotated through a small angle (4°) about an axis lying roughly along the longitudinal axis of the molecule. (The computer can easily effect this rotation *without* the direction of projection being changed, by applying a rotational transformation to the computed co-ordinates of the centres of gravity of the atoms.)

J. I. Leenhouts is with Philips Research Laboratories, Eindhoven.

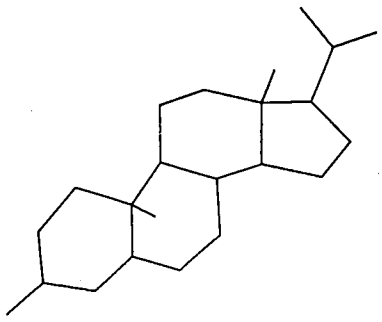


Fig. 2. An orthogonal parallel projection of the "Duphaston" molecule. The centres of gravity of atoms that are linked by chemical bonds are interconnected here by straight lines; the direction of projection has been chosen so as to obtain close resemblance to the structural formula.

Fig. 2 shows a projection of this kind, chosen so as to give a picture that corresponds broadly to the chemical structural formula. The centres of gravity of atoms linked by chemical bonds are shown here connected by straight lines. For clarity, the hydrogen atoms have been omitted in all the figures.

The projections could again have been drawn by means of the COBRA-controlled drawing machine. In this case, however, the accuracy required was much less critical than in the examples described in the first

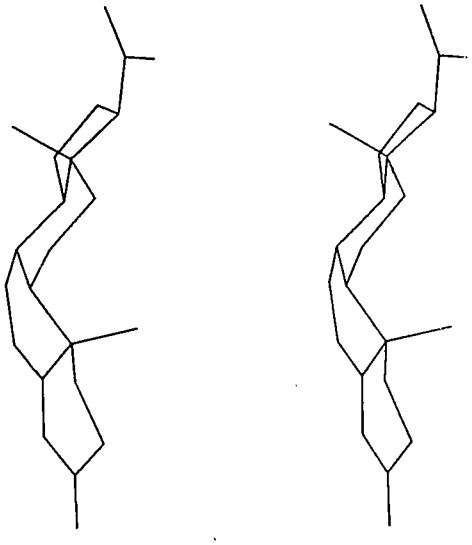


Fig. 3. Stereo pair of drawings of the "Duphaston" molecule, seen more from the side than in fig. 2. The pictures should be viewed from a distance of about 40 cm, the left eye looking at the left-hand figure and the right eye at the right-hand one. Better results can often be obtained if a sheet of paper is held upright between the two figures.

article of this series. It was therefore sufficient to use an x-y plotter of the type normally used in conjunction with a computer. The figures in this article were made by a plotter which gave an accuracy of $\frac{1}{4}$ mm over an area of 30×30 cm.

An x-y plotter of this type can carry out the following instructions:

1) Draw a straight line between points x_1,y_1 and x_2,y_2.
2) Move the pen, without writing, from x_3,y_3 to x_4,y_4.

With a suitable combination of these instructions, the plotter can draw each figure once this has first been approximated by the computer with straight-line sections. Since the x-y plotter obviously draws much more slowly than the computer calculates, the instructions are first recorded on magnetic tape, which is then fed to the x-y plotter (off-line operation). The preparation of the magnetic tape (including the calculation of the projections) takes the computer in our example only 60 seconds for all 90 projections together.

All the successive drawings can be taken together to form a ciné film, which shows the molecule rotating. This gives a remarkably clear view of the three-dimensional structure of the molecule. Two projections of this series made from slightly different angles may also be combined to form a stereoscopic pair. *Fig. 3* shows an example, in which the angle between the projection planes is 4°. The best impression of depth is obtained when the angle between the axis of the eyes when viewing is also 4°; this corresponds to a viewing distance of about 40 cm.

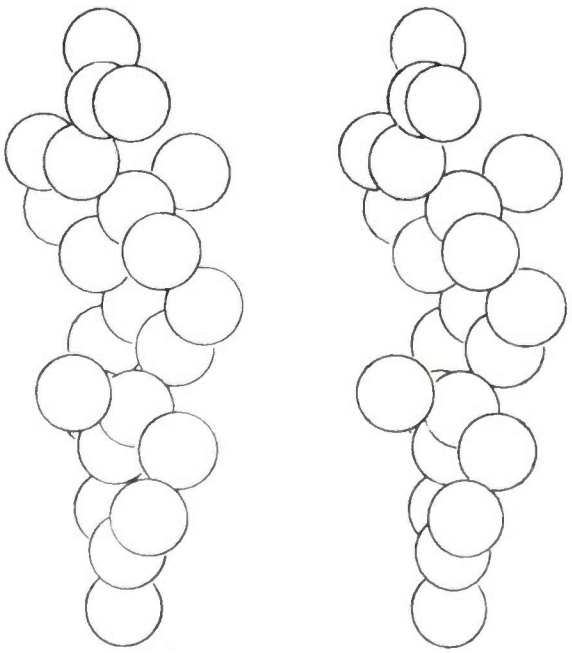


Fig. 4. Stereo pair of drawings of the "Duphaston" molecule, with the atoms represented by circles around the computed centres of gravity. Optimum viewing distance about 20 cm.

A more "natural" picture of the molecule can be obtained by drawing circles around the centres of gravity in the projections, to represent the atoms. This operation was carried out for all the projections with the help of the computer, which at the same time omitted the parts of the circles which were hidden by the circles of atoms in front of them. A film can also be made of these drawings as well; two of them are shown combined here to form a stereoscopic pair (*fig. 4*) and are also given in the form of an anaglyph in *fig. 5*. The angle between the projection planes here was 8°.

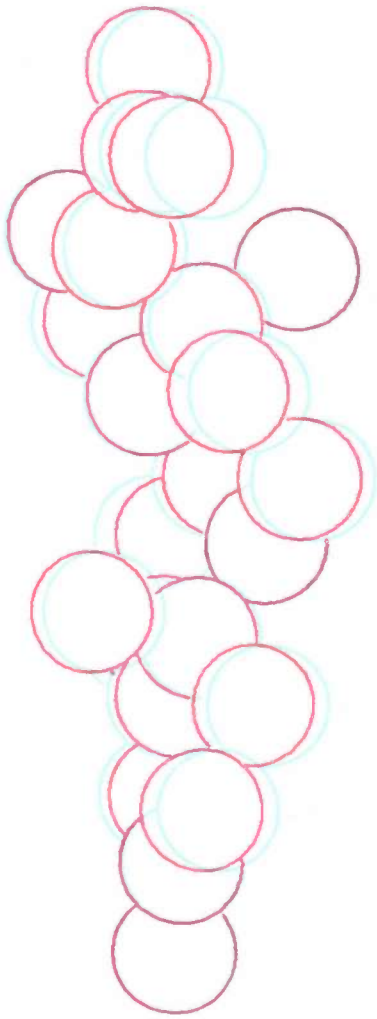


Fig. 5. Anaglyph of the stereo pair from fig. 4. Optimum viewing distance about 50 cm. The figure should be viewed using the coloured sheets included in this issue; see the caption to fig. 7 in article I.

The three-dimensional impression given by this figure

[1] P. B. Braun and A. J. van Bommel, X-ray determination of crystal structures, Philips tech. Rev. **22**, 126-138, 1960/61.
[2] A paper by P. B. Braun, J. Hornstra and J. I. Leenhouts on this subject will appear in Philips Res. Repts.
[3] O. A. de Bruin, H. F. L. Schöler and J. N. Walop, Retrosteroids, a new class of compounds with sex-hormone action, Philips tech. Rev. **28**, 70-79, 1967 (No. 3/4).
[*] Registered trade mark of N.V. Philips-Duphar.

is in fact relatively poor compared with the examples given in article I. This is so because we have not used perspective projections, and the projections were not both computed for the same projection plane, as they really should be for an anaglyph. Moreover, owing to the absence of surface details we do not see the "atoms" as spheres — which would be desirable as the simplest image — but as flat discs. Nevertheless, for practical purposes these figures do give a good idea of the three-dimensional form of the molecules.

III. A solution of the Van der Pol equation, represented in three dimensions

J. A. Zonneveld

The differential equation

$$y'' - \varepsilon(1 - y^2)y' + y = 0, \quad \ldots \quad (1)$$

introduced in about 1926 by Van der Pol in the investigation of relaxation oscillations [1], has become the prototype of a non-linear differential equation. For every value of the dimensionless parameter ε this equation has a solution $y(x)$ which, irrespective of the initial values $y(x_0)$ and $y'(x_0)$, approximates to a periodic function at large values of x. At large values of ε this stationary solution has the form of a "square sine". As a rule only the period and the amplitude of the stationary solution are of importance; for large values of ε asymptotic series expansions for period and amplitude are known [2]. In order to find the complete solution $y(x)$ for a given value of ε, it is necessary to integrate equation (1) using methods of approximation. This used to be done by a combined graphic-numerical method (the method of isoclines) [1] which was not very accurate; nowadays, however, the whole integration is done numerically with the aid of a computer, using for example a Runge-Kutta method [3]. In view of the highly abnormal behaviour of the functions at large values of ε, the integration has to be carried out with strongly varying discrete steps.

We have computed the stationary solution in this way for $\varepsilon = 10$ making use of the Runge-Kutta formula RK4n [3]. This method gives a result in the form of sets of values of x, y and y', called triples (x,y,y'), which constitute a curve in the x,y,y'-space; this curve gives the stationary solution of (1); see *fig. 1*. The data for fig. 1, which was drawn by hand, were calculated by the computer.

The drawing can be carried out fully automatically, however, using the COBRA-controlled drawing machine as described in article I of this series. The computer (we used an Electrologica X8) need then only cal-

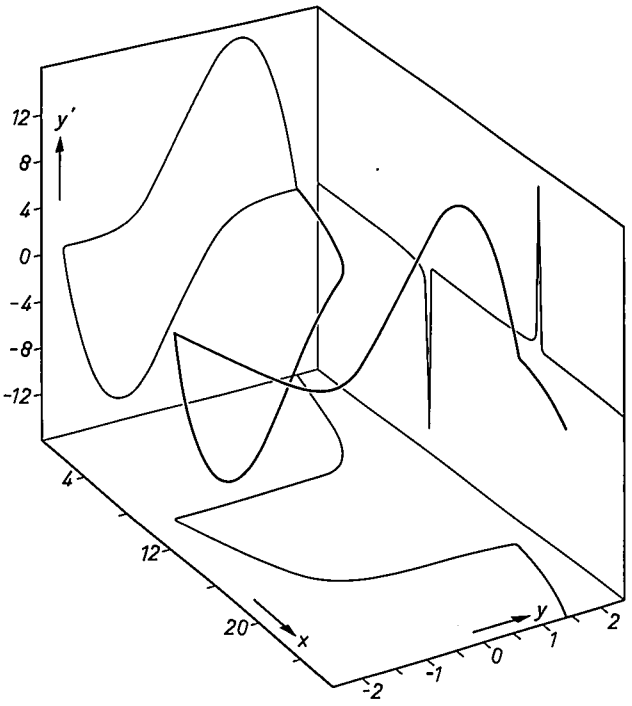


Fig. 1. Stationary solution (independent of the initial values) of the Van der Pol equation for $\varepsilon = 10$. The three-dimensional curve and its orthogonal projections on the three co-ordinate planes were calculated by a computer; so also was the complete projection on the plane of the drawing.

Dr. J. A. Zonneveld is with Philips Research Laboratories, Eindhoven.

[1] Balth. van der Pol, Phil. Mag. **2**, 978-992, 1926; see also Philips tech. Rev. **1**, 39-45, 1936; H. Bremmer, The scientific work of Balthasar van der Pol, Philips tech. Rev. **22**, 36-52, 1960/61.

[2] A. A. Dorodnicyn, Prikl. Mat. Mekh. **11**, 313-328, 1947, English translation in Amer. Math. Soc. Transl. No. **88**, 1953. J. A. Zonneveld, Periodic solutions of the Van der Pol equation, Indag. math. **28**, 620-622, 1966.

[3] J. A. Zonneveld, Automatic numerical integration, Math. Centre Tracts No. 8, Amsterdam 1964.

is in fact relatively poor compared with the examples given in article I. This is so because we have not used perspective projections, and the projections were not both computed for the same projection plane, as they really should be for an anaglyph. Moreover, owing to the absence of surface details we do not see the "atoms" as spheres — which would be desirable as the simplest image — but as flat discs. Nevertheless, for practical purposes these figures do give a good idea of the three-dimensional form of the molecules.

III. A solution of the Van der Pol equation, represented in three dimensions

J. A. Zonneveld

The differential equation

$$y'' - \varepsilon(1 - y^2)y' + y = 0, \quad \ldots \quad (1)$$

introduced in about 1926 by Van der Pol in the investigation of relaxation oscillations [1], has become the prototype of a non-linear differential equation. For every value of the dimensionless parameter ε this equation has a solution $y(x)$ which, irrespective of the initial values $y(x_0)$ and $y'(x_0)$, approximates to a periodic function at large values of x. At large values of ε this stationary solution has the form of a "square sine". As a rule only the period and the amplitude of the stationary solution are of importance; for large values of ε asymptotic series expansions for period and amplitude are known [2]. In order to find the complete solution $y(x)$ for a given value of ε, it is necessary to integrate equation (1) using methods of approximation. This used to be done by a combined graphic-numerical method (the method of isoclines) [1] which was not very accurate; nowadays, however, the whole integration is done numerically with the aid of a computer, using for example a Runge-Kutta method [3]. In view of the highly abnormal behaviour of the functions at large values of ε, the integration has to be carried out with strongly varying discrete steps.

We have computed the stationary solution in this way for $\varepsilon = 10$ making use of the Runge-Kutta formula RK4n [3]. This method gives a result in the form of sets of values of x, y and y', called triples (x,y,y'), which constitute a curve in the x,y,y'-space; this curve gives the stationary solution of (1); see *fig. 1*. The data for fig. 1, which was drawn by hand, were calculated by the computer.

The drawing can be carried out fully automatically, however, using the COBRA-controlled drawing machine as described in article I of this series. The computer (we used an Electrologica X8) need then only cal-

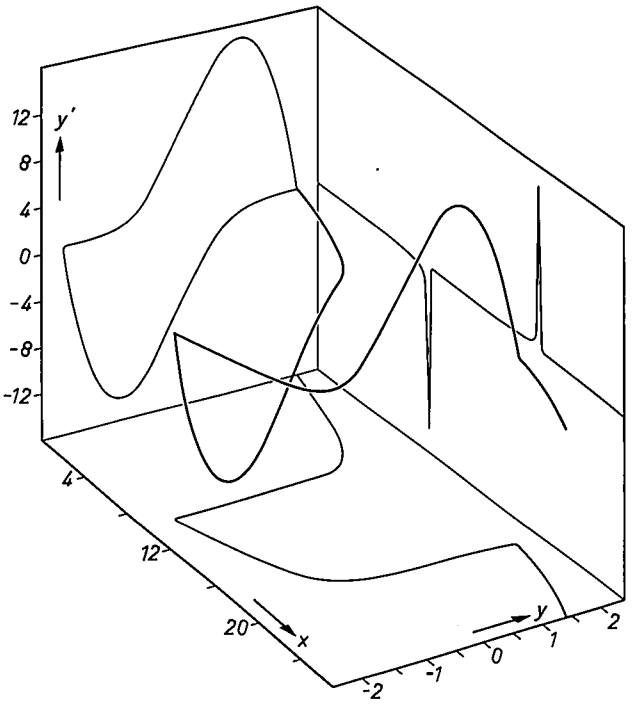


Fig. 1. Stationary solution (independent of the initial values) of the Van der Pol equation for $\varepsilon = 10$. The three-dimensional curve and its orthogonal projections on the three co-ordinate planes were calculated by a computer; so also was the complete projection on the plane of the drawing.

Dr. J. A. Zonneveld is with Philips Research Laboratories, Eindhoven.

[1] Balth. van der Pol, Phil. Mag. **2**, 978-992, 1926; see also Philips tech. Rev. **1**, 39-45, 1936; H. Bremmer, The scientific work of Balthasar van der Pol, Philips tech. Rev. **22**, 36-52, 1960/61.

[2] A. A. Dorodnicyn, Prikl. Mat. Mekh. **11**, 313-328, 1947, English translation in Amer. Math. Soc. Transl. No. **88**, 1953. J. A. Zonneveld, Periodic solutions of the Van der Pol equation, Indag. math. **28**, 620-622, 1966.

[3] J. A. Zonneveld, Automatic numerical integration, Math. Centre Tracts No. 8, Amsterdam 1964.

culate as many triples as are needed for linear interpolation between these points to yield a curve which approximates to the solution with sufficient accuracy. A punched paper tape containing the x-, y-, and y'-coordinates of these points is now fed to the COBRA, which then causes an orthogonal, oblique or perspective projection to be drawn, whichever may be required. As an illustration, two perspective projections, forming a stereo pair, have been drawn for the solution for $\varepsilon = 10$; in *fig. 2* these are given by an anaglyph.

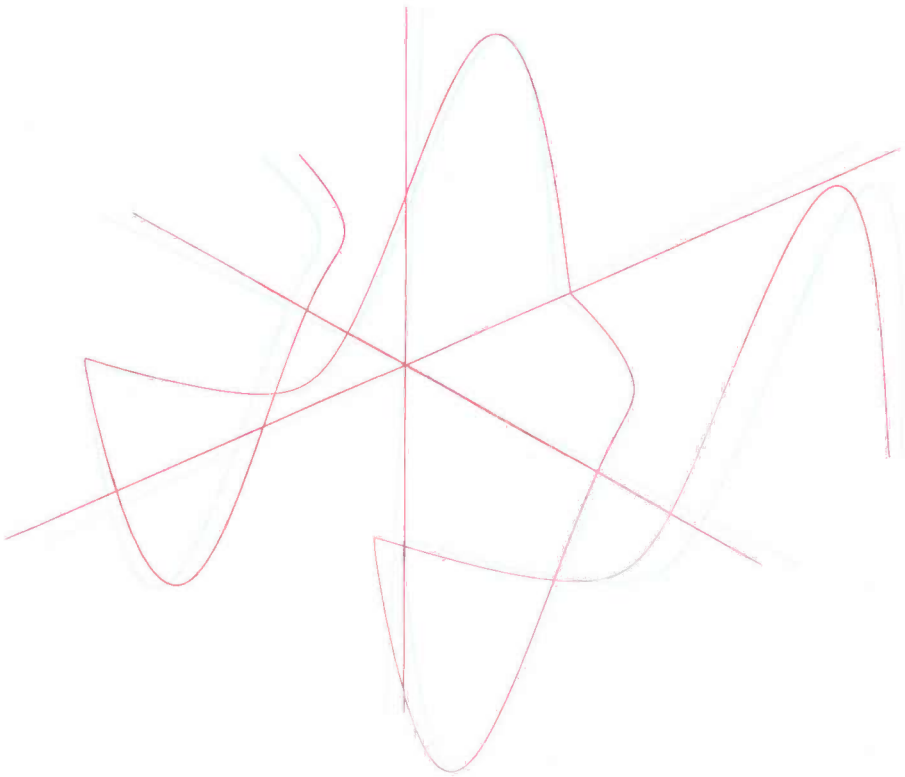


Fig. 2. Anaglyph of the solution represented in fig. 1. The two perspective projections needed for these were computed and automatically drawn by the COBRA-controlled drawing machine. The figure should be viewed using the coloured sheets included in this issue; see the caption to fig. 7 in article I.

Anomalous transmission of X-rays in highly perfect crystals

B. Okkerse and P. Penning

X-ray diffraction is well known as a means for the investigation of crystal structures. It is not so well known that the "anomalous transmission" of X-rays can be utilized to investigate lattice disturbances in highly perfect crystals with striking results: dislocations can be made visible and bending of lattice planes with a radius of curvature as great as 500 km can be detected.

In recent years techniques for producing highly perfect single crystals have been developed to such an extent that "dislocation-free" crystals are no longer the rarity they used to be. This is particularly true in the case of the semiconductor materials silicon and germanium and to a lesser extent with gallium arsenide and indium antimonide.

One consequence of this has been to awaken renewed interest in X-ray diffraction theory, particularly in the "dynamical" theory, because of the rather startling phenomenon known as anomalous transmission or the Borrmann effect [1] which is only observed in highly perfect crystals. Interest in this effect was first aroused in our laboratories by Hunter [2] who showed that anomalous transmission is highly sensitive to dislocations and to lattice disturbances due to elastic strain. It therefore offered a means of assessing the perfection of the very high quality crystals which we produce [3].

The phenomenon of anomalous transmission

In studying X-ray diffraction one can distinguish between the Bragg case and the Laue case; in the first case the reflected beam leaves the crystal at the entrance surface, in the second case the beam passes through the crystal and leaves it at the other side. Only in this latter case does anomalous transmission occur. It is most easily studied if the entrance and the exit surfaces of the crystal are perpendicular to the reflecting planes. We shall confine our attention to this case, the *symmetrical* Laue case.

The absorption coefficient of germanium for CuKα X-rays is 352 cm^{-1}. This means that the intensity of an X-ray beam after transmission through a slice of germanium 1 mm thick will be reduced to exp (—35.2) or 5×10^{-16} times the intensity of the incident beam. So if one wishes to study the diffraction of X-rays by transmission through germanium crystals, then in order to have sufficient transmitted intensity, one would have to prepare slices which are at the most about 50 μm thick. However, if the crystal under examination is sufficiently perfect, and the Bragg condition is satisfied, the transmitted intensity is many orders of magnitude larger than expected. For example, if diffraction of CuKα radiation takes place at (220) planes in germanium the absorption coefficient is reduced to approximately 14 cm^{-1}.

The effect can be readily demonstrated by means of the arrangement shown diagrammatically on the left-hand side of *fig. 1*. The specimen is a slice of a highly perfect germanium crystal which has been carefully prepared by suitable etching so that all surface damage due to cutting and grinding has been removed. The plane of the slice is best chosen as either {110} or {111} so that there are sets of {220} type planes normal to the faces of the slice, which will be used as reflecting planes. The specimen is illuminated with a narrow beam of CuKα X-rays and the transmitted intensity is measured as a function of the angle of incidence Θ. It is observed that at an angle Θ_B corresponding to the Bragg angle, there is a considerable increase in the transmitted intensity and two beams emerge: a transmitted beam T, which is parallel to the incident beam I, and a reflected beam R.

If in this arrangement a film is placed behind the crystal so as to register the emergent beams and if, in addition, we allow the incident beam to mark the straight-through position (e.g. by removing the crystal), then, when the film has been developed, three images will be visible as shown on the right-hand side of fig. 1. It is seen that beams T and R are approximately of equal intensity, implying that the energy emerging from the crystal is being divided equally between them. To

Dr. Ir. B. Okkerse is with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven; Dr. Ir. P. Penning is Professor in Theoretical and Applied Physics at the Technical University of Delft; both were formerly with Philips Research Laboratories, Eindhoven.

explain the position of the images it is necessary to assume the X-ray path shown. Careful measurement will verify that the two emerging beams originate from one location at the back of the crystal; to arrive at this location the incident beam must have traversed the crystal parallel to the reflecting planes.

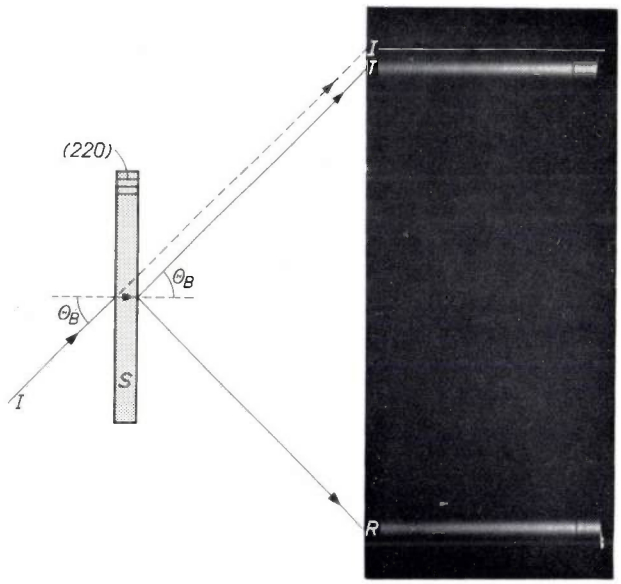


Fig. 1. Anomalous X-ray transmission. The beam geometry on the left explains the image positions seen in the photograph on the right. The X-rays appear to propagate parallel to the diffracting (220) planes. *I* incident beam. Θ_B Bragg angle. *S* perfect crystal. The intensities of the reflected beam *R* and the transmitted beam *T* are much greater than expected from the kinematical diffraction theory.

Dynamical explanation of X-ray diffraction in perfect crystals [4]

The diffraction of a parallel beam

Two theories are in current use to explain the intensities observed in X-ray diffraction studies: the *kinematical* theory and the *dynamical* theory. In its simplest form the kinematical theory is based on the following assumptions: a) X-rays, which are electromagnetic waves, are assumed to travel with the velocity of light in the crystal. In other words, the index of refraction is assumed to be unity. b) Radiation once scattered by the electron cloud surrounding an atom, is supposed to pass out of the crystal without further scattering. c) The decrease in intensity of both the primary (or transmitted) beam and the reflected beam is governed by the normal laws of absorption, which means that the relative decrease in intensity per unit length of path is constant and independent of the direction of the path.

These assumptions are quite justified provided one is dealing with a very small crystal, or a mosaic crystal consisting of very small crystal blocks misaligned slightly with respect to each other, or a fine powder. However, as the size of the crystal or of the perfect crystal blocks increases, the intensity expressions do not adequately describe the experimental observations. The main reason for the breakdown in validity of the kinematical theory for larger crystals is the violation of assumption (b) mentioned above.

The scattering power of a single free electron is very small. The total reflected intensity of a *small* crystal even if it contains a large number of electrons and even if all scattered wavelets are added, as is achieved in Bragg reflection, is still very small compared with the incident intensity. As an example let us consider that a slice of germanium 1 μm thick is being investigated in the arrangement of fig. 1. The addition of all the scattered wavelets results in a reflected amplitude of only about 5% of the incident amplitude. In view of this small ratio a second scattering can be neglected. If, however, the thickness had been 10 μm the reflected wave would have had an amplitude of half the amplitude of the incident wave. It will be clear that for such a thickness multiple scattering cannot be neglected. In fact, the photograph in fig. 1 was made with a slice of germanium which was no less than 1.85 mm thick.

The theory that takes the second and further scattering into account properly is the dynamical theory. For the scattering by any electron located at $\mathbf{r}_e$ this theory takes into account the *total wave-field*, amplitude $\mathbf{D}(\mathbf{r}_e)$, at the location of that electron, and not the incident wave-field, amplitude $\mathbf{D}_i(\mathbf{r}_e)$, alone. This means that the correct solution must be *self-consistent*: the total amplitude $\mathbf{D}(\mathbf{r})$ at any point r of space is the sum of the amplitude $\mathbf{D}_i(\mathbf{r})$ of the incident wave plus the amplitude of the waves set up by scattering, at the electrons, of the total wave-field of amplitude $\mathbf{D}(\mathbf{r})$, assumed to be known already. (The term amplitude here refers both to the magnitude and the phase.)

In the dynamical theory the self-consistent solutions in unbounded crystals are considered first. We shall refer to these solutions as *modes of propagation*. For crystals of finite size, application of the proper boundary conditions then determines which modes of

[1] The phenomenon of anomalous transmission was first observed in 1941 by Borrmann when investigating the transmission of X-rays through certain selected calcite and quartz crystals; see for instance G. Borrmann, Phys. Z. **42**, 157, 1941.

[2] L. P. Hunter, Proc. Kon. Ned. Akad. Wetensch. **B 61**, 214, 1958.

[3] B. Okkerse, Philips tech. Rev. **21**, 340, 1959/60.

[4] W. H. Zachariasen, Theory of X-ray diffraction in crystals, Wiley, New York 1946; R. W. James, Solid State Physics **15**, 53, 1963; B. W. Batterman and H. Cole, Revs. mod. Phys. **36**, 681, 1964; P. Penning, thesis, Delft 1966 (also published in Philips Res. Repts. Suppl. 1966, No. 5.)

propagation are activated by the incoming beam and to what extent.

Rather than treating the dynamical theory in detail we shall illustrate the procedure with the special example considered above. We assume at first that the medium is non-absorbing. The incoming beam is assumed to be parallel and linearly polarized such that the electric vector **D** is parallel to the reflecting planes (σ-polarization). (Waves with **D** normal to the reflecting planes, i.e. π-polarized waves, will not be discussed since these do not contribute to the anomalous transmission.) According to the dynamical theory two modes of propagation are activated by a σ-polarized wave inside the crystal. Both modes again have their vector **D** parallel to the reflecting planes. For a special direction of incidence, which as we shall show later satisfies the Bragg condition, the wave fields of both modes *travel* in a direction parallel to the reflecting planes (the z-direction), while at the same time they form *standing* waves in the x-direction, i.e. perpendicular to the reflecting planes. One mode, A_σ, has its nodes in the reflecting planes. The other mode, B_σ, has its antinodes there. In *fig. 2* the amplitudes are shown as a function of x. The reflecting planes are located at $x = nd$, where d is the spacing of the planes and n an integer 0, 1, 2, ... In mathematical terms the amplitudes **D** are given

for mode A_σ by $\mathbf{D}_A \sin(\pi x/d) \exp(-jk_A z)$, and
for mode B_σ by $\mathbf{D}_B \cos(\pi x/d) \exp(-jk_B z)$.

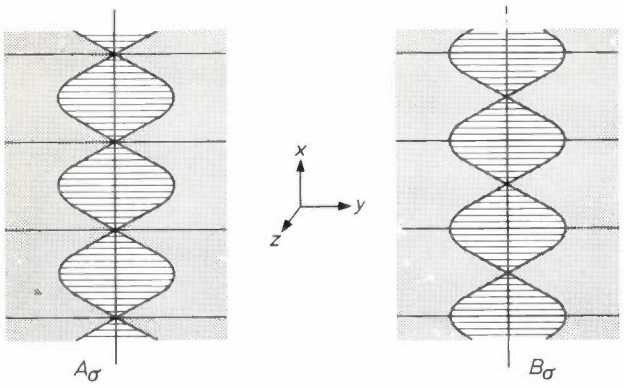


Fig. 2. The two modes A_σ and B_σ, activated by a plane wave linearly polarized in the y-direction (σ-polarization) and incident exactly at the Bragg angle. The reflecting planes, drawn as thick horizontal lines, are normal to the surface. The modes, which are similarly polarized in the y-direction, behave as standing waves in the x-direction and as travelling waves in the z-direction.

The standing wave character in the x-direction is expressed by the sine and cosine terms, the travelling wave character in the z-direction follows from the exponential term; $k_A = 2\pi/\lambda_A$ is the magnitude of the wave vector of mode A, and similarly $k_B = 2\pi/\lambda_B$ for mode B.

Let us consider in particular the case in which the set of reflecting planes contains all the atoms of the crystal lattice, as with the (220) planes in germanium. This means that the scattering electrons are located close to the reflecting planes. The interaction between mode A_σ and the electrons will be much weaker than the interaction between mode B_σ and the electrons, since the amplitude of mode A_σ at the location of the electrons is very small. The interaction with the electrons causes the wave to move faster and hence increases the wave length and decreases the magnitude of the wave vector for a given frequency. Therefore we must expect k_A to be larger than k_B. The difference

$$\Delta k = k_A - k_B$$

is of the order of a few times $10^{-5} \times 2\pi/\lambda$ and proportional to the Fourier component with period d of the electron density in the crystal lattice [5].

The power flow associated with the wave fields of modes A_σ and B_σ follows immediately from the fact that these modes have a standing wave character in the x-direction. No power flow takes place parallel to the x-axis so that it must be parallel to the reflecting planes for both modes.

Let us now consider the matching at the entrance surface located at $z = 0$. The total wave field just inside the crystal, arising from the two modes, has to be equal in amplitude and period to the wave field just outside the crystal arising from the incident plane wave with amplitude $\mathbf{D}_0$. A possible solution is that the two modes have equal magnitude and a phase difference of 90°:

$$\mathbf{D}_A = -j\mathbf{D}_0,$$
$$\mathbf{D}_B = \mathbf{D}_0.$$

Then we obtain for the amplitude just inside the crystal:

$$\mathbf{D} = \mathbf{D}_0 \exp(-j\pi x/d).$$

The sum of both modes is now a wave travelling along the surface with wave vector π/d. This wave has to be matched with the wave just outside the crystal. The wave vector component in the x-direction is equal to $(2\pi \sin \Theta)/\lambda$ with Θ the glancing angle. This component must be equal to π/d so that:

$$\lambda = 2d \sin \Theta.$$

In other words, the case in which modes A_σ and B_σ are excited inside the crystal with equal magnitude arises if the incident plane wave satisfies the Bragg condition exactly.

The matching conditions at the front surface re-

[5] Or in other words it is proportional to the corresponding structure factor; see for instance P. B. Braun and A. J. van Bommel, Philips tech. Rev. 22, 126, 1960/61, where this factor (F) is given by formula (3) on page 129.

quires a phase difference of 90° between modes A_σ and B_σ. During passage through the crystal this phase difference changes because k_A is larger than k_B. It is convenient to introduce the parameter L, defined as the distance over which the modes A_σ and B_σ have to travel before the phase difference between the two modes changes 2π radians. Hence:

$$L = 2\pi/(k_A - k_B).$$

The value of L is of the order of 10 μm.

Let us consider a crystal slab whose thickness is exactly L. At the back surface the phase difference between A_σ and B_σ is again 90°. The situation as far as matching conditions are concerned is identical at the front and back surfaces so that in the vacuum behind the crystal there appears a wave travelling parallel to the incident wave. There is no reflected wave even though the incident beam satisfies the Bragg condition exactly.

Next we consider a crystal with a thickness of $\frac{1}{2}L$. At the back surface there is now a phase difference of 270° between modes A_σ and B_σ. Just inside the crystal we have:

$$\mathbf{D} = \mathbf{D}_0 \exp(j\pi x/d) \exp(-jk_B L/2).$$

This is again a travelling wave, but it travels in the opposite x-direction. The matching now requires that only a reflected wave is present in the vacuum behind the crystal. The results are shown diagrammatically in *fig. 3*. The case of a crystal of intermediate thickness is also shown. Now there are two plane waves in the vacuum behind the crystal, one travelling parallel to the incident wave and one parallel to the reflected wave. *Fig. 4* shows how the amplitude of the transmitted wave (parallel to the incident wave) and the amplitude of the reflected wave vary with thickness.

According to the kinematical theory the amplitude of the reflected wave varies linearly with the thickness of the crystal. This is apparently only true for a thickness much smaller than L.

This remarkable result that the distribution of the emerging intensity over the transmitted and the reflected beam varies periodically with thickness (the period being L, which is therefore often called "Pendellösung length") can be verified experimentally by recording the intensity reflected by a wedge-shaped sample. If the angle of the wedge is γ, L is given by sin γ times the distance between fringes. *Figs. 5a* and *b* show such a record for a germanium wedge, with CuKα and MoKα radiation respectively. The reflecting planes are (220). From the figures it follows that L is approximately 8 μm for CuKα radiation and approximately 17 μm for MoKα.

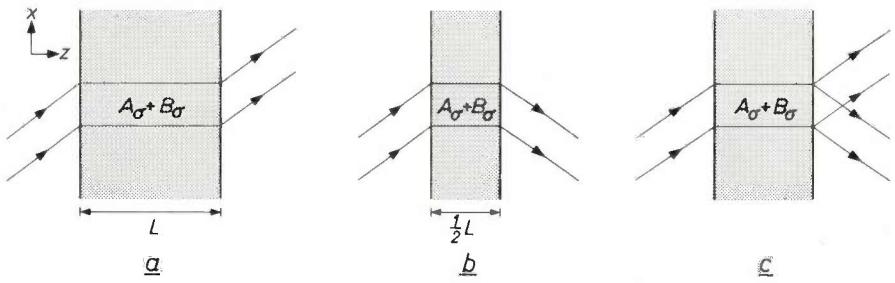


Fig. 3. Wave behaviour in non-absorbing crystals. When an incident σ-polarized plane wave satisfies the Bragg condition exactly, two modes A_σ and B_σ are excited in the crystal; these travel with a small difference in velocity, L being the distance over which their phase difference changes by 2π (of the order of 10 μm). When the thickness h is a multiple of L, the two modes set up a single transmitted wave parallel to the incident wave, behind the crystal (case *a*); when $h = (n + \frac{1}{2})L$, n being integer, a single reflected wave emerges (case *b*). For an intermediate thickness both waves are present (case *c*).

It is a well known fact that the X-rays may also

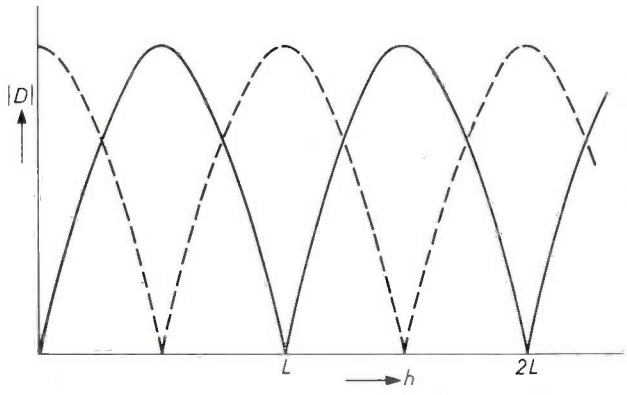


Fig. 4. Amplitude variation |**D**| of the reflected beam (solid line) and the transmitted beam (dashed line) as a function of the crystal thickness h.

interact with electrons in such a way that the incident X-ray energy is transformed into other forms of energy and is lost for diffraction. The amount of energy *absorbed* in this way will be proportional to the square of the amplitude of the wave field at the location of the absorbing electrons. From the structure of the modes A_σ and B_σ as drawn in fig. 2 it follows immediately that mode A_σ is absorbed much less than mode B_σ. If the crystal is sufficiently thick only mode A_σ will remain. At the back surface the matching now always gives two waves, one in the transmitted direction and one in the reflected direction (*fig. 6*). This can easily be seen if the amplitude distribution in mode A_σ is written in a slightly different way:

$$\mathbf{D} = \frac{1}{2\mathrm{j}}\mathbf{D}_A\left[\exp\mathrm{j}\left(\frac{\pi x}{d} - k_A z\right) - \exp\mathrm{j}\left(-\frac{\pi x}{d} - k_A z\right)\right].$$

This shows that the two emerging waves have equal amplitudes.

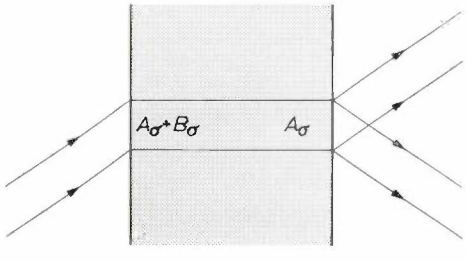


Fig. 6. Diagram showing how in a thick crystal, for the exact Bragg condition, the mode A_σ remains and splits into two waves of equal intensity on reaching the back surface.

The explanation of the phenomenon of anomalous transmission is now clear. The incident wave that satisfies the Bragg condition exactly, excites modes A_σ and B_σ inside the crystal. Mode B_σ is absorbed completely, while mode A_σ only undergoes a small absorption if the crystal is perfect or near-perfect, whereas in non-perfect crystals both modes are strongly absorbed. The power flow in mode A_σ is parallel to the reflecting planes. At the back surface it breaks down into two plane waves of equal amplitude. In fig. 1 the same conclusions have been drawn from experimental evidence.

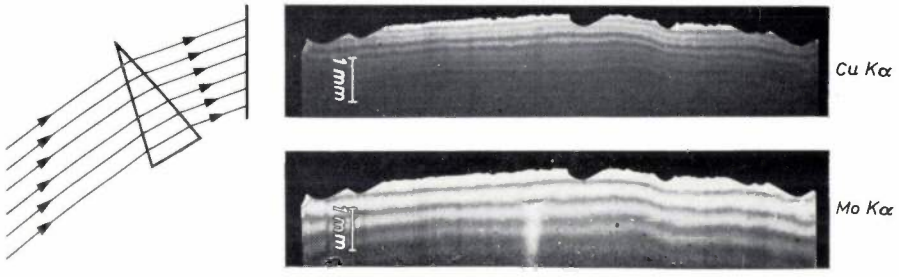


Fig. 5. Fringes corresponding to fig. 4, visible in the intensity reflected from a thin {111} germanium wedge using {220} reflecting planes and CuKα or MoKα radiation. The fading out of the fringes is due to absorption and the divergence of the incident beam. The angle of the wedge is 1°55′.

The diffraction of a divergent beam

In actual experiments the beam is more or less divergent, and we have to consider now what happens to an incident wave that does not satisfy Bragg's law exactly. In the dynamical theory it is shown that in these cases the wave field is no longer a standing wave in the direction perpendicular to the reflecting planes; a travelling wave is added to it as well. The more the incident wave deviates from the exact Bragg angle, the larger the travelling wave will become in comparison with the standing wave. Well away from the exact Bragg angle the standing wave is zero and the total wave field corresponds to a single travelling wave. Its wave vector is now close to the wave vector of the incident wave. The deviation of the refractive index from unity is small ($\approx 3 \times 10^{-6}$) and arises from the overall polarizability of the electrons. The intermediate cases where in the direction perpendicular to the reflecting planes the wave is neither a pure standing wave nor a pure travelling wave are found only in a very narrow angular range around the exact Bragg angle. The width is of the order of a minute of arc. Because there is now also a travelling wave in the direction perpendicular to the reflecting planes the nodes no longer have zero amplitude. The ratio of the amplitude at the nodes (the minimum amplitude) to the amplitude at the antinodes (the maximum amplitude) rises steadily from 0 to 1, as the deviation from the Bragg angle increases. For modes A_σ' the minimum amplitude lies at the reflecting planes whilst for modes B_σ' the amplitude is a maximum there.

A first consequence of the presence of a travelling wave in the direction perpendicular to the reflecting

planes is that the direction of power flow is no longer parallel to the reflecting planes. The larger the deviation from the Bragg angle the more the power flow deviates from being parallel to the reflecting planes. An example is given in *fig. 7*. The direction of incidence ($\Theta = \Theta_B - \delta$) is a few seconds of arc smaller than the exact Bragg angle. Two modes are activated inside the crystal. The power flow in mode A_σ' is tilted upwards over an angle α and that in mode B_σ' over the same angle downwards. Decreasing the angle of incidence still more leads to an increasing angle α. When Θ is a few minutes of arc off Θ_B, the power flow for A_σ has become parallel to the incident beam and for B_σ (which is not excited any more because of the boundary conditions) parallel to the reflected beam. The entire "fan" $2\Theta_B$ in possible directions of power flow is scanned by changing the direction of the incident wave by only a few minutes of arc.

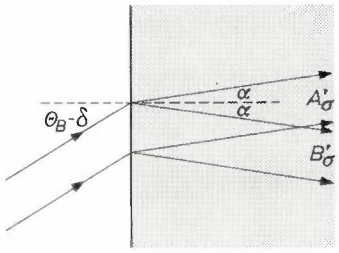


Fig. 7. As the angle of incidence deviates from the Bragg angle Θ_B (deviation angle δ) two modes are still excited in the crystal but the power flow in each makes an angle α ($\leq\Theta_B$) with the diffracting planes.

There is a second consequence of the presence of a travelling wave in the direction perpendicular to the reflecting planes, connected with the absorption. Mode A_σ is absorbed only very slightly because its amplitude is small at the location of the absorbing electrons. For modes A_σ' activated by a beam incident at an angle slightly different from the Bragg angle, the nodes no longer have zero amplitudes although the minimum is still located at the absorbing electrons. Accordingly these modes still undergo little absorption but the absorption will be greater for increasing deviations from the Bragg angle, until for deviations of more than a few minutes of arc from the exact Bragg angle the absorption coefficient reaches its normal value μ_0. For modes B_σ' the amplitude is a maximum in the reflecting planes. Hence they will suffer enhanced absorption, but it decreases with increasing deviations from Θ_B. The limit is again the normal absorption coefficient μ_0 (which corresponds to a wave interacting with the *average* electron density).

Fig. 8 shows what happens inside the crystal and at the back surface if a wave slightly off the Bragg angle strikes the crystal. The incident wave is considered to be a rather narrow beam and is indicated by a straight line. The two modes that are activated inside the crystal are narrow also and propagate in the direction of their respective directions of power flow [6]. Mode B_σ' is absorbed strongly and the beam does not reach the back surface for sufficiently thick crystals. Mode A_σ' may reach the back surface. There it breaks down again into a transmitted and a reflected wave. The distribution of intensity over the two beams is such that the vector sum of the power-flow vectors is parallel to the power flow in mode A_σ' and that the absolute sum is equal to the magnitude of the power flow in this mode on arrival at the back surface.

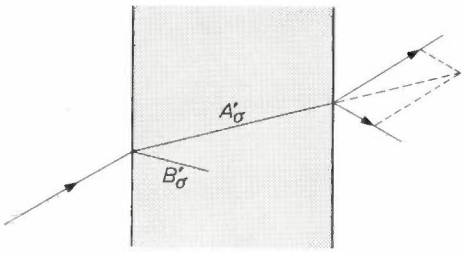


Fig. 8. The ratio of the transmitted and reflected intensities for a beam incident at an angle slightly different from the Bragg angle can be determined from the vector parallelogram of the energy flow vectors.

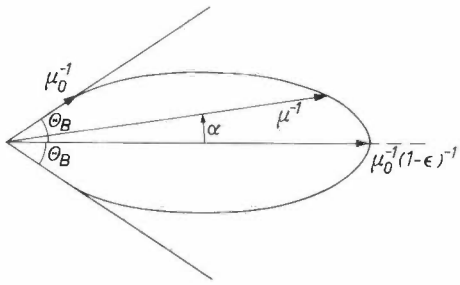


Fig. 9. Polar diagram showing the variation of the reciprocal absorption coefficient μ^{-1} with the angle α which the power flow makes with the reflecting planes for mode A_σ.

We have seen above that the absorption depends on the direction of power flow. *Fig. 9* is a polar diagram in which the reciprocal of the absorption coefficient for mode A_σ' is plotted against its direction of power flow. From this we conclude that as the crystal thickness is increased in a transmission experiment,

[6] Because they are not parallel, the beams become separated in space. The interference that leads to the fringes (fig. 5) therefore no longer occurs.

more and more of the rays in the fan $2\Theta_B$ are absorbed until eventually only those rays travelling almost parallel to the lattice planes will remain. These will emerge at the exit surface in the form of rays in the transmitted and reflected directions. A thick crystal (i.e. $\mu_0 h > 10$, where μ_0 is the normal absorption coefficient and h the thickness) therefore acts as a kind of collimator by selecting for transmission only those rays from the divergent incident beam which can undergo anomalous transmission. The beams emerging from such a crystal therefore have very little divergence. A typical half-width is in fact only a few seconds.

In *fig. 10* are given the results of calculations according to the dynamical theory, which show the sharpness of an anomalously transmitted beam compared with that of a beam which has undergone diffraction in the symmetrical Bragg case (i.e. with the crystal surface parallel to the reflecting planes). It is also interesting to note that according to these calculations the regions of diffraction do not coincide, and that the angles of incidence at which there are diffraction maxima differ by about 10 seconds. In the next section we shall discuss experiments which confirm these theoretical results.

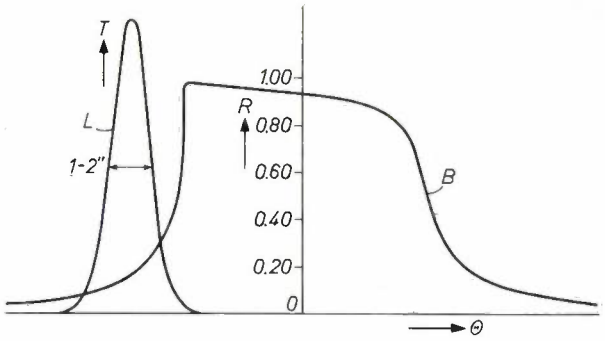
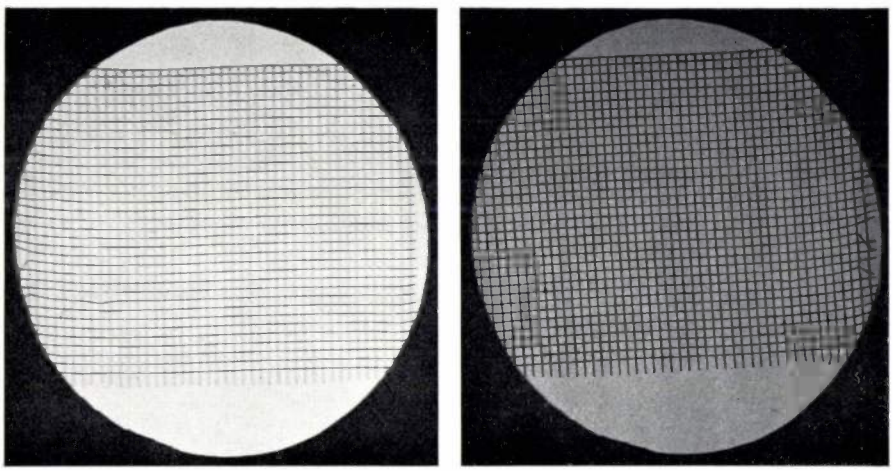


Fig. 10. Result of calculations based on the dynamical theory of X-ray transmission and reflection in a perfect crystal, where it is assumed that the set of diffracting planes is the same in transmission as in reflection, and that the incident beam is a plane wave. Curve L is calculated for the case where the diffracting planes are perpendicular to the entrance and exit surfaces of the crystal (symmetrical Laue case), curve B for the case of diffracting planes parallel to the crystal surfaces (symmetrical Bragg case). T ratio of transmitted and incident intensity. R ratio of reflected and incident intensity. Θ angle of incidence. It is seen that the intensity maxima occur at angles of incidence differing by an amount of the order of 10 seconds of arc, and that Laue diffraction occurs in a much smaller angular region than Bragg diffraction.



a *b*

Fig. 11. Illustration of the Borrmann delta effect. *a*) A grid of 25 μm wires when placed on the front surface of the crystal (facing the X-ray source) produces an image in which the shadow of the vertical wires, which are parallel to the diffracting planes, is diffuse. In (*b*) the grid is on the back surface and the shadow of the wires is sharp.

The presence of the fan (or Borrmann delta as it is sometimes called) can be clearly demonstrated by taking anomalous transmission pictures of a parallel-sided slab of dislocation-free germanium, first with a grid of fine copper wires on the front surface, i.e. between specimen and X-ray source, and secondly with the grid between specimen and film (*fig. 11*). It is seen that in (*a*) the shadow images of the vertical wires, which are parallel to the net planes, are diffuse whereas in (*b*) they are quite sharp. Reference to *fig. 12* readily shows why this is so. The absorption by a wire on the front face eliminates the complete "fan" arising at its location so that a broad shadow image is formed whereas a wire on the back absorbs only the emergent parallel rays. The influence of the Borrmann delta can also be seen in fig. 1 where the narrow incident beam I after transmission through the crystal gives relatively wide images T and R on the film.

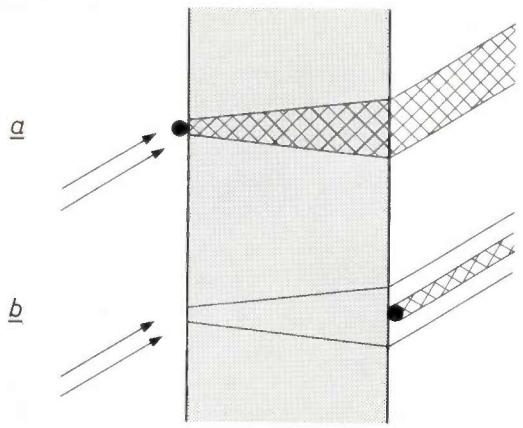


Fig. 12. Ray diagram showing formation of diffuse (*a*) and sharp images (*b*) for wires on front and back surfaces respectively.

Evaluation of the minimum in absorption

The minimum value of the absorption coefficient is usually written in the form $\mu_0(1-\epsilon)$, where ϵ is a parameter close to unity. Its actual value depends on a number of factors. The first factor is the type of reflecting plane that is used. In discussing the absorption of mode A_σ we assumed that all atoms are in the reflecting planes; only in that case will the minimum in absorption coefficient be low. This assumption, however, is not valid for all the permitted reflections in the diamond lattice of germanium. In *fig. 13* a view of the lattice projected on to the $(1\bar{1}0)$ plane is given. It shows that for the (220), $(22\bar{4})$ and the (004) reflections (and in general all the reflections with even indices whose sum is a multiple of 4) the set of reflecting planes contains all the atoms. For the (111) reflections this is not the case. Hence we expect that for the (111) reflections (and in general all the reflections with odd indices) ϵ is much smaller than unity. It has been verified experimentally that these reflections do indeed show a much less pronounced Borrmann effect.

Another factor that influences the value of ϵ is the thermal motion, since the atoms are not fixed at their equilibrium positions, but vibrate around them and accordingly move out of the reflecting planes.

It is found that the temperature dependence of ϵ can be described by a Debye-Waller factor, i.e. $\epsilon = \epsilon_0 \exp(-M)$, similar to the temperature dependence for the structure factor. The value of M is related to the root-mean-square amplitude of the vibrations of the atom and hence depends on the Debye temperature of the crystal. The value of ϵ_0 is related to the distribution of electrons around the nucleus and to the order of the reflection.

The value of $1-\epsilon$ has been determined for the three lattice planes (220), (400) and (422) of germanium by measuring the anomalously transmitted intensity through a wedge at constant temperature [7]. The results are given in *Table I*.

Table I. Experimental values of $1-\epsilon$, at 25 °C, for the reflections (220), (400) and (422) in germanium.

Reflection	$1-\epsilon$	$(1-\epsilon)/(h^2+k^2+l^2)$
220	0.0408	0.00510
400	0.0812	0.00508
422	0.1216	0.00507

It is apparent that the absorption coefficient is largest when the distance between the reflecting planes is smallest. This is to be expected since the effect of the atomic size and thermal vibration will be greater. In practical terms this means that the anomalous transmission in germanium is greatest for (220)-type planes.

It is also to be noted from the last column that $1-\epsilon$ is proportional to $h^2+k^2+l^2$. This means that if $1-\epsilon$ has been determined for one reflection it can be calculated for other reflections.

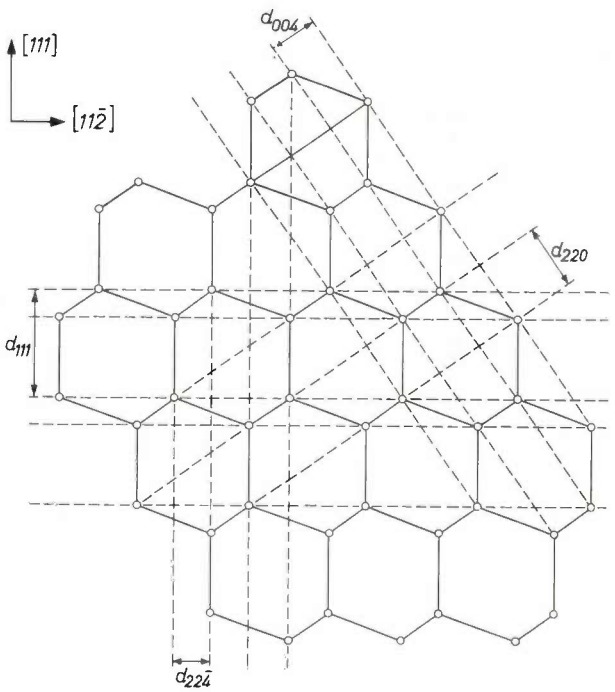


Fig. 13. Projection of a germanium lattice on to a $(1\bar{1}0)$ plane. Planes like (220), $(22\bar{4})$ and (004) contain all the atoms present in the lattice whereas (111) planes do not.

Experimental investigations of anomalous transmission

Consecutive reflections in the same perfect crystal

The sharpness of the anomalously transmitted beam enables us to measure extremely small differences in angle. For example, an anomalously transmitted beam from one part of the crystal can be made to undergo a

[7] B. Okkerse, Philips Res. Repts. **17**, 464, 1962.

further reflection in another portion by utilizing special shapes of crystals. Two possible arrangements are shown in the diagram of *fig. 14*.

The L-shaped configuration in fig. 14*a* can be used to demonstrate the result of the dynamical theory that the diffraction angle in the case of Laue reflection (R_1) is different from that in the case of Bragg reflected beam. If no load is used, and the reflecting planes of the two parts of the L may be assumed to be parallel the reflected intensity of the Bragg reflection is relatively weak; it increases on varying the angle of incidence and reaches a maximum at the angle corresponding to the maximum of the Bragg peak in fig. 10. In this way we have determined the intensity variation

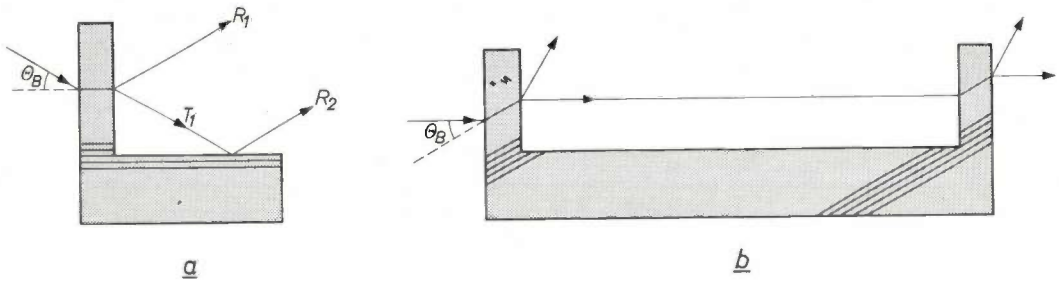


Fig. 14. Anomalous transmission experiments on single crystals of special shape. *a*) L-shaped configuration cut with (220) planes normal to the surface on which the X-rays are incident. The beam is first transmitted anomalously through the vertical part of the crystal and then the transmitted beam T_1 undergoes Bragg reflection (R_2) against its horizontal part. *b*) U-shaped configuration in which the (220) planes now make an angle Θ_B with the normal to the entrance surface. The beam is transmitted through both arms of the U-shaped crystal.

Variations in the relative orientation of the two parts of the L and the two arms of the U result in intensity variations of the second diffraction, since the Bragg condition is no longer satisfied exactly.

flection (R_2), for the same type of reflecting planes, as indicated in fig. 10. This is done experimentally by varying the relative orientation of the two sides of the L, using different loads as is illustrated in *fig. 15*, in this way varying the angle of incidence of the Bragg

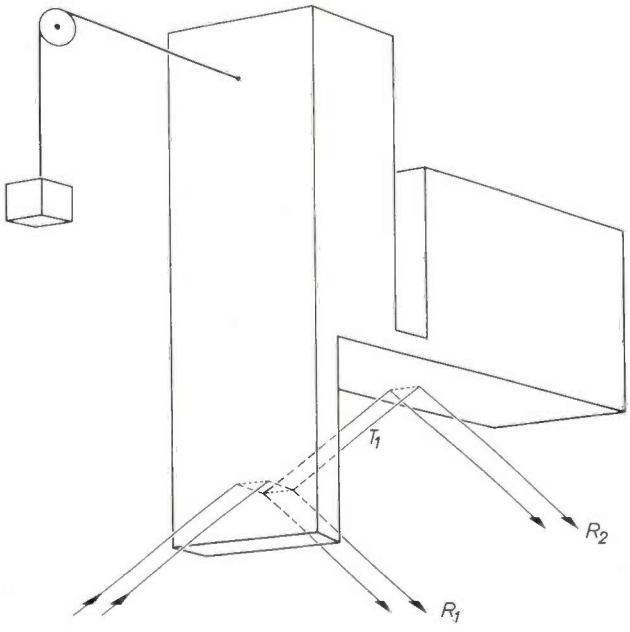


Fig. 15. Special crystal form for experiments as indicated in fig. 14*a*. By applying different loads the narrow bridge of the crystal can be bent elastically and in this way the angle of incidence on the base of the L is varied. (The regions of the crystal where the diffraction takes place are *not* deformed.)

of (220) and (400) reflections of germanium as a function of the angle of incidence. They showed good agreement with the intensities that can be calculated when "scanning" the whole region of diffraction angles with the Laue peak. These results provide a convincing proof of the validity of the dynamical theory [8].

The method of consecutive reflections is extremely sensitive to lattice disturbances: angular changes as small as 10^{-7} radians can be detected. This means that the method can detect a bending of the lattice planes in the U-shaped crystals with a radius of curvature of 500 km. The method of consecutive reflections can also be utilized to measure the refractive index [9]. This has been done for a number of plastic materials using CuKα radiation. The technique is illustrated in *fig. 16*. The values are found to differ from unity by only a few parts in a million. (In order to explain the figure it should be recalled that the refractive index for X-rays is smaller than unity so that these rays are refracted *away* from the normal when travelling from air into a medium that is denser for X-rays.)

Another application of consecutive reflections is the X-ray interferometer, as first described by Bonse and Hart in 1965 [10]. It is illustrated diagrammatically in *fig. 17*.

The device is manufactured from a large, highly perfect single crystal of silicon cut in the form of a letter E [11]. Each arm is thick enough for anomalous transmission of CuKα radiation to take place along the (220) planes. The principle of its operation

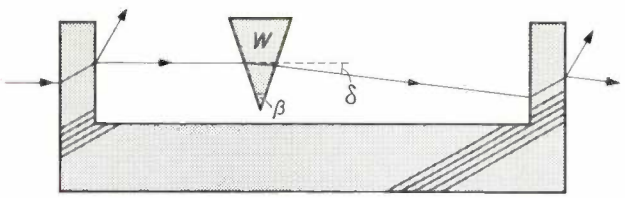


Fig. 16. Owing to refraction in the prism W the transmitted beam changes in direction by an angle δ. Because of the change in direction the Bragg condition is no longer satisfied exactly. The resulting decrease in intensity can be compensated by bending the U-shaped crystal, which gives us the value of δ. The refractive index n is calculated from the standard formula for minimum deviation [9]: $n = \sin\frac{1}{2}(\beta - \delta)/\sin\frac{1}{2}\beta$. (The change of direction is exaggerated in the figure: for instance for a prism made from plastic, δ is less than one second of arc. Since this angle is so small the direction of the power flow in the second arm of the crystal can be made parallel to the reflecting planes by bending the horizontal part of the crystal.)

is qualitatively as follows. The X-ray beam incident on the arm S of the device is split by Laue diffraction into two coherent beams $D^{S}{}_{T}$ and $D^{S}{}_{R}$. On reaching the "mirror crystal" M they each undergo further Laue diffraction to produce beams $D^{M}{}_{TT}$, $D^{M}{}_{TR}$, $D^{M}{}_{RR}$ and $D^{M}{}_{RT}$. The geometry is so arranged — M being equidistant between A and S — that the beams $D^{M}{}_{TR}$ and $D^{M}{}_{RR}$ are incident on the same area of the analyser crystal A. In the region where these two beams overlap interference occurs and two waves emerge from A, $D^{A}{}_{T} = D^{A}{}_{TRR} + D^{A}{}_{RRT}$ and $D^{A}{}_{R} = D^{A}{}_{TRT} + D^{A}{}_{RRR}$. Each consists of two anomalously transmitted beams of coherent waves which are in phase. If a phase shift is introduced into *one of the beams* before A, for example $D^{S}{}_{T}$, by placing a piece of material in the beam, then there will be a corresponding phase shift in the beam $D^{M}{}_{TR}$ (and $D^{M}{}_{TT}$) and consequently also in $D^{A}{}_{TRR}$ and $D^{A}{}_{TRT}$. This will result in a reduced intensity in the beams $D^{A}{}_{T}$ and $D^{A}{}_{R}$. Since anomalously transmitted beams show practically no divergence, variations of the phase shift in the cross-section of such a beam will result in local intensity variations. A photographic plate on which either the $D^{A}{}_{T}$ or $D^{A}{}_{R}$ beam is incident, therefore, will record the phase variations due either to path length differences or to inhomogeneities of the material placed in the beam.

An interference pattern will also be observed if there is a misalignment of any of the three arms. This can be utilized in a variety of ways [12].

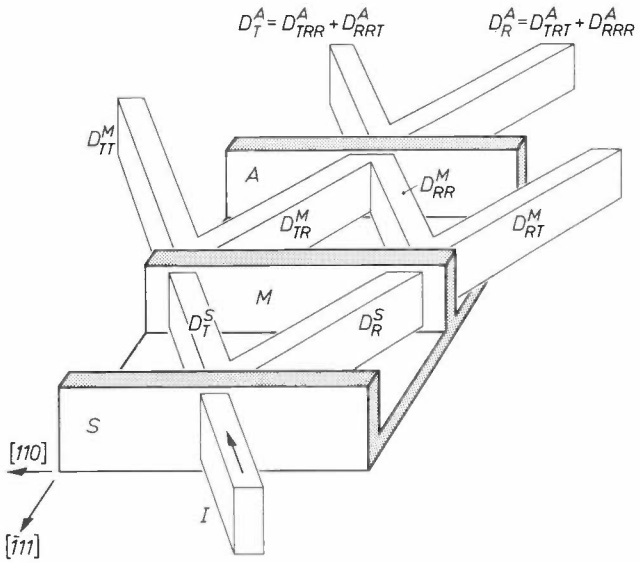


Fig. 17. Illustrating the principle of the X-ray interferometer.

Influence of lattice disturbances on anomalous transmission

The dynamical diffraction theory discussed so far has only been concerned with perfect crystals. Real crystals usually are not perfectly periodic but possess a variety of defects which affect the periodicity to a greater or lesser extent and thereby influence the diffraction properties. For example, a large number of small angle grain boundaries, as in a mosaic crystal, destroys the long-range periodicity so that there is a complete breakdown of the conditions necessary for anomalous transmission. On the other hand, one would expect that in lightly deformed crystals the influence of the deformation could be determined in much the same way as in geometrical optics where beam propagation through a medium of slowly varying refractive index is treated.

An extension of the dynamical theory has in fact been developed based on this analogy [13]. The deformation in the crystal is described by a slowly varying reciprocal lattice vector, and it is assumed that a mode of propagation remains a single mode as it progresses through the crystal. To comply with the changing lattice parameter, the mode of propagation must gradually change its wave character and hence also its direction of power flow. In general, it is found that as the beam progresses through the crystal its path becomes curved. At the exit face transmitted and reflected beams emerge as in an unstrained crystal. The ratio of their intensities depends on the angle the beam in the crystal makes with the back surface.

We now give some examples of different types of deformation and their effect on anomalous transmission.

Example 1. If all the atoms in the reflecting planes are moved parallel to these planes no effect is expected since the distance between the reflecting planes remains unchanged. All the atoms still lie on the nodal planes of mode A. Again, subjecting the crystal to uniform expansion or contraction does not change the transmission since the periodicity is maintained; however, the Bragg angles change slightly.

Example 2. A crystal in which the reflecting planes are perpendicular to the entrance and exit surfaces is

[8] B. Okkerse, Philips Res. Repts. **18**, 413, 1963.
[9] B. Okkerse, Philips Res. Repts. **20**, 377 and 389, 1965.
[10] U. Bonse and M. Hart, Appl. Phys. Letters **6**, 155, 1965.
[11] U. Bonse and M. Hart, Appl. Phys. Letters **7**, 99, 1965.
[12] U. Bonse and M. Hart, Z. Physik **190**, 455, 1966.
[13] P. Penning and D. Polder, Philips Res. Repts. **16**, 419, 1961.

bent so that there is a tapering of these planes which, however, still remain flat. This can also be achieved by means of a temperature gradient parallel to the reflecting planes. The situation is shown diagrammatically in *fig. 18*. In this example as well the anomalously transmitted beam is unaffected by the distortion. However, the diffraction angle at the entrance surface is different from the diffraction angle at the exit surface [14]. The distortion has no effect because modes of propagation can be found which consist entirely of cylindrical (or spherical) waves, whose origin is situated in the axis of bending: all the atoms in the lattice then lie once more on the nodal planes.

Example 3. If the deformation is such as to cause a bending of the lattice planes, a large change in the transmitted and reflected intensity can be observed [15]. Such a situation occurs when the lattice planes of the last example make an angle other than 90° with the surface as in *fig. 19*, or when a temperature gradient is applied to the crystal perpendicular to the reflecting planes (*fig. 20*). The path of the X-ray beam in the crystal is now strongly curved. This curvature is very much greater (roughly 10^4 times greater) than that of the bending of the lattice planes, and so for most of its path the beam will cut across them and thereby undergo considerable absorption. The intensities of the transmitted and reflected beams emerging from the crystal depend on the angle the beam in the crystal makes with the exit surface. The path of the beam is also dependent on the angle of incidence.

Example 4. The effect of the strain field around a *scratch* is illustrated in *fig. 21*. The lattice distortion surrounding the scratch causes the intensity of the transmitted and reflected beams to decrease so that the scratch is visible as a region of reduced intensity. When the distorted region is removed, for example by etching, the scratch is no longer visible in X-ray photographs although the trace of the scratch can often be seen optically as a groove. Bending the crystal gives rise to stress concentrations around the groove so that the "scratch" is again visible by X-rays.

Example 5. Dislocations represent another disturbance of the lattice, in the interior of the crystal. Here, the exact matching of the atomic positions with the nodes of the wave field is disturbed around the dislocation so that the anomalous transmission pictures of a parallel beam incident at the Bragg angle will contain the *shadows* of dislocations. This property has been developed into a method of revealing dislocations in highly perfect crystals which will be discussed now. The method is generally referred to as the Barth-Hosemann technique [16].

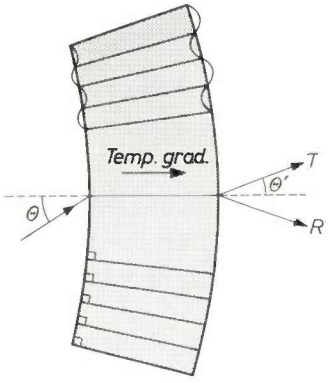


Fig. 18

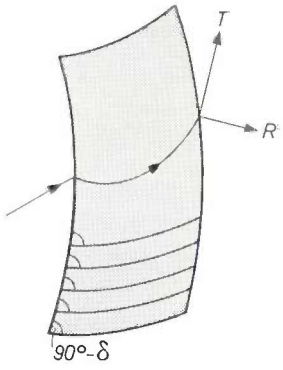


Fig. 19

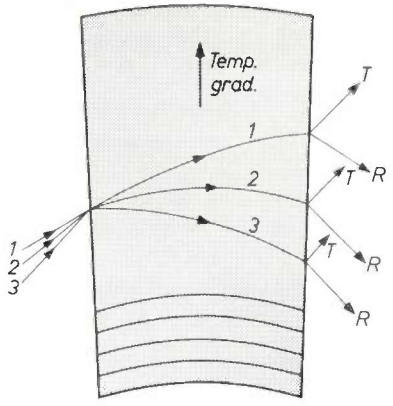
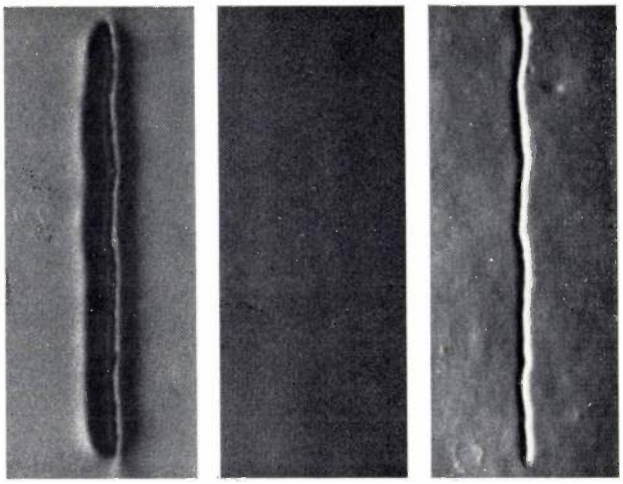


Fig. 20

Fig. 18. In a crystal deformed so that the lattice planes taper while remaining undistorted, the nodes of a cylindrical standing wave can be matched to the tapering planes, which means that the absorption of the transmitted beams is still small. The transmitted and the reflected beam now emerge at an angle Θ' which is smaller than the angle of incidence [14]. This type of deformation can be produced by applying a temperature gradient parallel to the reflecting planes.

Fig. 19. The path of the X-ray beam in a deformed crystal with a bending of the lattice planes. The beam curvature is very much greater than the lattice curvature (roughly 10^4 times). Additional absorption takes place. Minute curvatures can be detected in this way.

Fig. 20. Crystal with a temperature gradient normal to the reflection planes. Changing the angle of incidence varies the path in the crystal. Path *1*: the power flow is parallel to the reflecting planes at the exit surface. Path *2*: the same in the middle of the crystal. Path *3*: the same at the entrance surface. Path *2* is the path of minimum absorption. For path *1* the intensity of *R* is approximately equal to that of *T*; for paths *2* and *3* the intensity of *R* is greater than that of *T* [15].



a *b* *c*

Fig. 21. *a*) The influence on anomalous transmission of a *scratch* on the back surface of a germanium crystal. In (*b*) the scratch is removed by etching and in (*c*) the crystal is bent subsequently. In (*a*) a lattice distortion is present around the scratch, in (*c*) the stress around the remaining groove is higher than in the bulk of the bent crystal. The strong contrast in (*c*) occurs because of the different bending of the lattice planes on the two sides of the groove causing the reflected intensity to be greater than the transmitted intensity in one case and smaller in the other case. The pictures were made with the reflected beams.

The Barth-Hosemann technique resembles the Lang technique described earlier in this journal [17]. An essential difference is however that in the Lang technique only very thin crystals ($\mu_0 h < 1$), whereas in the Barth-Hosemann technique only thick crystals can be used ($\mu_0 h > 10$). As a consequence the dislocations show up in the Lang technique as regions of *enhanced* transmission.

Revealing dislocations by anomalous transmission

The principle of the Barth-Hosemann technique is shown diagrammatically in *fig. 22*. In a highly perfect crystal of sufficient thickness ($\mu_0 h > 10$) only the rays nearly parallel to the reflecting planes can reach the back surface. Therefore, the emerging reflected and transmitted beams will produce on a film two equal projection pictures (topographs) of the part of the crystal traversed. In these pictures the images of all points lying behind each other in the direction parallel to the reflecting planes will coincide. It is common practice to use the reflected beam only; the transmitted beam is then intercepted by a lead screen. Because the crystal itself acts as a plane collimator the incident beam may be rather divergent in one direction, namely the direction perpendicular to the reflecting planes. The only disadvantage of such a divergent beam is that rays incident at the Bragg angle corresponding to the $K\alpha_1$ and also the $K\alpha_2$ wavelength (and possibly the $K\beta$ wavelength) will be present, so that multiple images result, which will affect the resolution of closely spaced dislocations in the direction perpendicular to the reflecting planes.

The resolution in the direction *parallel* to the reflecting planes is governed entirely by the size of the X-ray focus in this direction. The crystal does not give any collimation in this direction. The smaller this size of the focus, the better the resolution.

To avoid images due to $CuK\beta$, a nickel filter can be placed over the exit window of the X-ray tube. Alternatively, the $K\beta$ images can be avoided by previous collimation so that the angular spread of the incident beam is less than the difference in Bragg angles of the $K\alpha$ and $K\beta$ components of the radiation used. This has in fact been done in the experimental arrangement used at the Philips Research Laboratories in

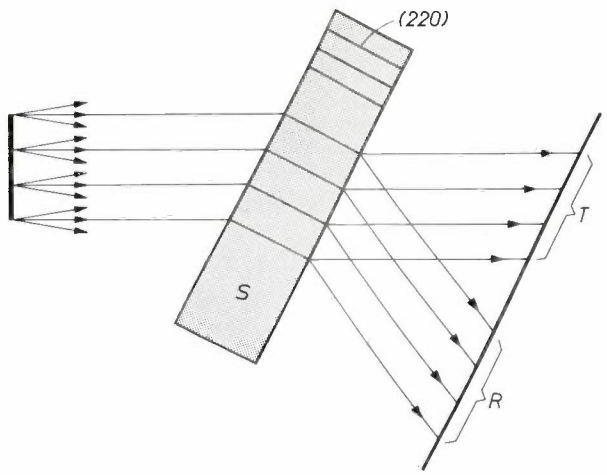


Fig. 22. Principle of the Barth-Hosemann technique for revealing dislocations. A divergent X-ray beam is incident on the crystal *S*. Only those rays which are incident at an angle differing not more than a few seconds from the Bragg angle will be anomalously transmitted; all the other rays will be strongly absorbed in the thick crystal. Dislocations situated in the illuminated region of the crystal are visible on the film in both the *T* and *R* images as regions of low intensity.

Eindhoven, shown in *fig. 23*, in which an effective X-ray focus of 1×1 mm is available. Since the beam is small in the direction perpendicular to the reflecting planes we need to scan the crystal in this direction, in order to get a complete picture of the crystal. As we move the crystal and the film together in the same direction the topographical relationship between crystal and film is preserved in the overlapping diffraction images produced in the successive positions. This scanning movement gives the additional advantage of smoothing out the local intensity variations of the X-ray source.

[14] H. Cole and G. E. Brock, Phys. Rev. **116**, 868, 1959.
[15] B. Okkerse and P. Penning, Philips Res. Repts. **18**, 82, 1963.
[16] H. Barth and R. Hosemann, Z. Naturf. **13a**, 792, 1958.
[17] A. E. Jenkinson, Philips tech. Rev. **23**, 82, 1961/62.

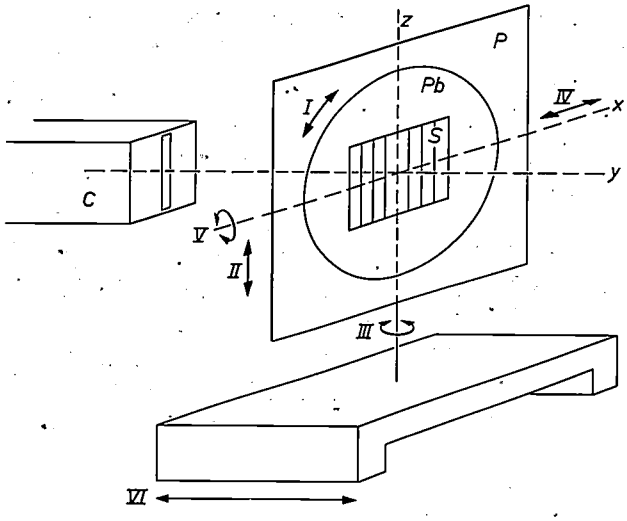


Fig. 23. Diagram of the equipment used at the Philips laboratory in Eindhoven for recording topographs by means of the Barth-Hosemann technique. A suitable X-ray generator is the standard Philips PW 1010 with a copper target. The best operating conditions for this work are 26 kV and 36 mA. At this rating the ratio of characteristic to continuous radiation is favourable. The specimen S is attached to a lead plate Pb which has a suitable hole. Both are then attached to the plate P of the goniometer. The diffracting planes are parallel to the plane of the ribbon-shaped X-ray beam from the collimator C. To obtain this orientation, the specimen can be rotated about the axis y (*I*). Means of adjustment in the vertical direction is also provided (*II*). To set to the appropriate Bragg angle, corresponding rotation is provided (*III*). A reciprocating motion (*IV*) along the x-axis ensures that the whole crystal is scanned by the beam. A further degree of freedom is also provided around the x-axis (*V*). Finally the whole goniometer can be displaced bodily in the y-direction (*VI*). The film holder (not shown) is mounted on the base of the goniometer and undergoes the same reciprocating motion as the specimen. It can be set at any angle with respect to the specimen but is usually best set parallel to the specimen. One beam can be suppressed.

With reference to *fig. 24* we now calculate the distance δx between the $K\alpha_1$ and the $K\alpha_2$ images to learn to what degree it affects the resolution. This distance on the film is given by

$$\delta x = \frac{S\,\delta\Theta}{\cos^2\Theta}.$$

Since $\lambda = 2d \sin\Theta$ it follows that

$$\delta x = \frac{S \sin\Theta}{\cos^3\Theta} \cdot \frac{\delta\lambda}{\lambda}.$$

The minimum distance δx_m of the images occurs when we place the film as close to the crystal as possible, i.e. at a distance $S_m = l/2 \sin\Theta$. As in this case $l = 1$ mm, we can calculate that for (220) reflections in germanium $\delta x_m = 1\frac{1}{2}$ μm, which is of little practical consequence.

We should like to mention that in the X-ray unit used, there is also an effective focus of 10×0.1 mm. As the beam is now much wider in the direction perpendicular to the reflecting planes we do not need to scan the crystal (which has however the disadvantage that the intensity variations of the X-ray source are not smoothed out). Since l is 10 mm in this arrangement, a larger distance S_m would be required, and we can calculate that the $K\alpha_1$-$K\alpha_2$ splitting does affect seriously the resolution. But there is a way out; and that is by placing the film in contact with the specimen at the exit surface, and so reducing the $K\alpha_1$-$K\alpha_2$ splitting to almost zero, with the additional advantage that both transmitted and reflected beams now contribute to the image so that the exposure time is reduced. For this latter arrangement to be effective the crystal must be opaque to all radiation other than at the Bragg angle. The crystal must therefore act as its own filter and not pass, for example, short-wavelength white radiation. (As this radiation usually emerges at angles quite different from the Bragg angle for CuKα radiation, it is a nuisance only in this arrangement where the film is placed close to the crystal.) Crystals which possess the required self-filtering property when examined with CuKα radiation are germanium and gallium arsenide. These materials have an X-ray absorption edge such that both long and short wavelength radiation undergo appreciable absorption. In silicon, on the other hand, the absorption edge occurs in a position where the short wavelengths are not very much attenuated so that appreciable "blackening" of the film will occur.

It is important to ensure that all surface damage is removed since the property of anomalous transmission is equally sensitive to surface strain. This can be done by suitable etching. The final specimen thickness should satisfy the criterion $\mu_0 h > 10$ so that with CuKα radiation a suitable specimen thickness for germanium would be about 1 mm, for silicon about 2 mm, and for

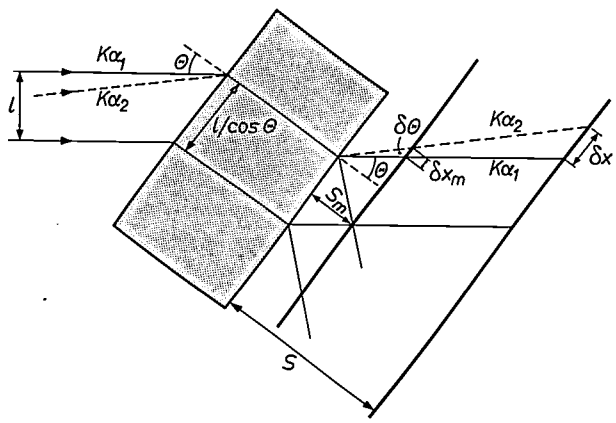


Fig. 24. Diagram to illustrate calculation of the minimum distance δx_m of CuKα_1 and Kα_2 images on the film.

gallium arsenide about 1 mm. The slices in our experiments are orientated to better than 1° and have as main faces {111}, {110} or {112}, so that the reflecting planes {220}, {400} and {422} are normal to them.

Topographs and their interpretation

Fig. 25a, b and *c* show three topographs of a gallium arsenide crystal [18] containing a few dislocations, which are made with $(2\bar{2}0)$, $(\bar{2}02)$ and $(02\bar{2})$ as reflecting planes. As expected they are visible as regions of reduced transmission.

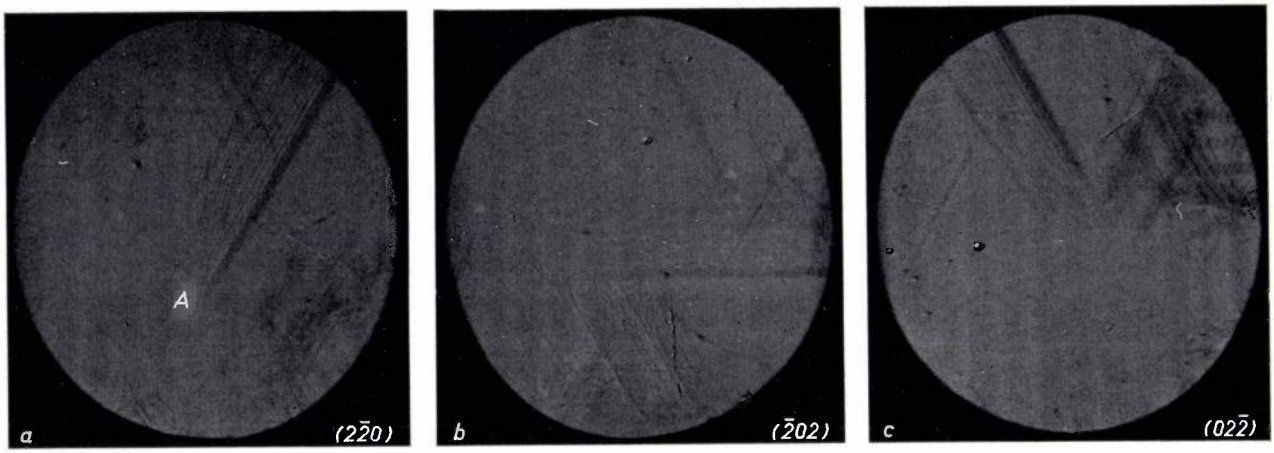


Fig. 25. Anomalous transmission topographs of a GaAs crystal, made with (2$\bar{2}$0), ($\bar{2}$02) and (02$\bar{2}$) respectively as reflecting planes. All the topographs are shown here with the normal to the reflecting planes oriented horizontally. The row of dislocations clearly visible at *A* lie in the (11$\bar{1}$) plane, a slip plane. We shall indicate by way of example how the direction of the Burgers vector can be deduced. We may assume the Burgers vector in gallium arsenide to have one of the six ⟨110⟩ directions (and to have a magnitude of $\frac{1}{2}a\sqrt{2}$). The Burgers vector should lie in the (11$\bar{1}$) plane, so that there are three possible directions: [$\bar{1}$10], [101] and [011]. We may make a second assumption, namely that the contrast of the dislocation is good if its Burgers vector is perpendicular to the reflecting planes and is poor if the Burgers vector is in the reflecting planes. The high contrast in the (02$\bar{2}$) topograph eliminates [011] (because it lies in the (02$\bar{2}$) plane). If [$\bar{1}$10] were the direction of the Burgers vector, then the contrast in the (2$\bar{2}$0) topograph (*a*) would be strong (because [$\bar{1}$10] is perpendicular to the (2$\bar{2}$0) plane), and further the (poor) contrast in (*b*) and (*c*) would be equal (since [$\bar{1}$10] makes the same angle with the ($\bar{2}$02) and the (02$\bar{2}$) planes). However, there is clearly a difference in contrast between (*b*) and (*c*). [101] remains as the only possibility, and we do indeed observe a poor contrast in the ($\bar{2}$02) topograph ([101] lying in the ($\bar{2}$02) plane), while (*a*) and (*c*) show an equally strong contrast.

An interesting feature of these pictures is the shape of the dislocation image. Dislocations running from the front to the back surface of the specimen give rise to an image which is delta-shaped, the sharp pointed end of the image corresponding to the part near the exit surface of the specimen. This is a consequence of the Borrmann delta effect, which we have already mentioned (p. 119). Experimentally it is observed that the image width is also affected by the density of the dislocations. The greater this is, the narrower the image of each dislocation.

The contrast of the dislocation image depends on the angle between the Burgers vector of the dislocation and the reflecting planes. The usefulness of this information can be illustrated using fig. 25 as an example. In the topographs (*a*) and (*c*), which are made with (2$\bar{2}$0) and (02$\bar{2}$) respectively as reflecting planes, we observe a strong contrast, while in the topograph (*b*), where ($\bar{2}$02) are the reflecting planes, the contrast is poor. From such differences in contrast one can deduce the direction of the Burgers vector — in this case the [101]-direction.

[18] This crystal was produced by the Battelle Memorial Institute, Geneva.

Summary. In studying the X-ray transmission of a highly perfect and sufficiently thick crystal (germanium for instance) it is found that on irradiation at the Bragg angle for particular sets of reflecting planes: a) the transmitted intensity is many orders of magnitude larger than would be expected from the kinematical theory, usually applied for non-perfect crystals; b) the power flow is parallel to the reflecting planes; c) a transmitted and a reflected beam of equal intensity emerge at the other side of the crystal, and d) the divergence of the emerging beams is extremely small (much less than a minute of arc). In order to explain these phenomena, known as anomalous transmission, the dynamical theory of X-ray diffraction is discussed. Property (d) can be used for the measurement of very small angles, of the order of 10^{-7} radians. With this method it is clearly established that the diffraction angles are slightly different when the entrance and exit surfaces of the crystal are parallel to the reflecting planes (symmetrical Bragg case) and when they are perpendicular to these planes (symmetrical Laue case). The anomalous transmission is extremely sensitive to dislocations and to lattice disturbances caused by elastic strain, and it therefore offers a means of assessing the perfection of a crystal. Bending of lattice planes with a radius of many hundred kilometres can be detected. The Barth-Hosemann technique based on anomalous transmission is discussed as a means of making dislocations visible.

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, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany *A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, 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.

G. A. Acket: Microwave Hall mobility of hot electrons in gallium arsenide.
Philips Res. Repts. **22**, 541-552, 1967 (No. 6). *E*

N. W. H. Addink & L. Bastings: Chemical inactivation *in vitro* of carbonic anhydrase by carcinogenic substances.
Nature **216**, 72-73, 1967 (No. 5110). *E*

J. Basterfield & M. J. Prescott: X-ray-diffraction topographic studies of magnetic-domain configurations in terbium iron garnet.
J. appl. Phys. **38**, 3190-3192, 1967 (No. 8). *M*

H. Bosma: On the theory of linear noisy systems.
Thesis, Eindhoven 1967. *E*

J. C. Brice: The kinetics of growth from solution.
J. Crystal Growth **1**, 218-224, 1967 (No. 4). *M*

G. Engelsma: Photoinduction of phenylalanine deaminase in gherkin seedlings, II. Effect of red and far-red light.
Planta **77**, 49-57, 1967 (No. 1). *E*

R. Genève: Modulation de la lumière.
Rev. HF **7**, 1-20, 1967 (No. 1). *L*

G. Klein & C. W. Bodenhausen: Stabilized power supply for 3 kV, 60 mA.
Electronic Engng. **39**, 609-611, 1967 (No. 476). *E*

H. de Lang: Polarization properties of optical resonators passive and active.
Thesis, Utrecht 1966. *E*

R. Mühlstroh & H. G. Reik: Optical absorption by small polarons in *p*- and *n*-type lanthanum cobaltite.
Phys. Rev. **162**, 703-709, 1967 (No. 3). *A*

B. J. Mulder: Sensitized photoconduction of anthracene.
Philips Res. Repts. **22**, 553-567, 1967 (No. 6). *E*

H. Rau: Thermodynamik des Selendampfes.
Berichte Bunsenges. phys. Chemie **71**, 711-715, 1967 (No. 7). *A*

H. Rau: Das Bildungsgleichgewicht des Selenwasserstoffs.
Berichte Bunsenges. phys. Chemie **71**, 716-719, 1967 (No. 7). *A*

J. F. Schouten (Instituut voor Perceptie Onderzoek, Eindhoven): Subjective stroboscopy and a model of visual movement detectors.
Proc. Symp. on models of the perception of speech and visual form, Boston 1964, p. 44-55; 1967.

G. Schulten: Simple molecular frequency standards for millimeter waves.
Microwave J. **10**, No. 12, 71-73, 1967. *H*

K. van Steensel: Vacuum-deposited thin-film capacitors of silicon monoxide.
Microelectronics and Reliability **6**, 261-268, 1967 (No. 4). *E*

T. L. Tansley: Spectral response of *p-n* heterojunctions.
Phys. Stat. sol. **23**, 241-252, 1967 (No. 1). *M*

G. Vandewiele (Université Libre de Bruxelles) **& M. Noé:** An inequality concerning tests of fit of the Kolmogorov-Smirnov type.
Ann. math. Statistics **38**, 1240-1244, 1967 (No. 4). *B*

J. Verweel: Magnetic properties of some ferroxplana single crystals.
Thesis, Amsterdam 1966. *E*

S. B. Weinstein: Optimum radar-antenna-gain patterns for attenuation of fixed background clutter.
Philips Res. Repts. **22**, 568-576, 1967 (No. 6). *E*

Volume 29, 1968, No. 3/4 pages 69-128 Published 15th April 1968

PHILIPS TECHNICAL REVIEW

VOLUME 29. 1968, No. 5

Optical investigations of semiconductor surfaces

K. H. Beckmann

We know today that much of our technology is critically dependent on surface phenomena. In some situations these phenomena are essential, whereas in others they are a nuisance and have to be effectively rendered harmless. The technological handling of surface problems is naturally related to their investigation in the laboratory, and two rather sophisticated methods employed in such investigations are discussed in this article.

The physical properties of solid surfaces are often different from those of the bulk material. The reason for this behaviour is to be seen primarily in the fact that the symmetry of the crystalline lattice is interrupted at the surface. This reduction in symmetry leads to the formation of a transition region between the surface and the undisturbed bulk material, since the absence of nearest neighbours on one side of the atoms at the surface does not only affect the surface atoms but creates a disturbance in the crystal field which extends several atomic layers into the solid. If the solid is not surrounded by high vacuum but is in contact with another medium which may be a gas or a liquid, adsorptions or chemical reactions may occur at the surface, leading to the formation of a third phase on the solid which consequently produces additional changes of the surface properties.

These surface properties are of great importance in the electrical behaviour of semiconductors and semiconductor devices. The number and type of the charge carriers within the space charge region below the surface can easily be changed by different surface treatments. Atoms adsorbed on the surface or incorporated into surface films may act as recombination centres, thus influencing strongly the lifetime of minority carriers. Furthermore, chemical processes such as etching and electropolishing play an important role in device fabrication [1], giving yet another incentive in the search for information about the nature of the chemical reactions between semiconductor surfaces and other agents.

For such investigations there are two optical methods which can furnish useful direct information about the presence, thickness, and chemical composition of films on the surface. This paper will deal with these methods: infra-red absorption spectroscopy by the technique of internal reflection, and ellipsometry. Both the theoretical background and the experimental realization of these procedures will be discussed, and particular attention will be given to results which have been obtained at the Philips laboratories in Hamburg and Eindhoven. The ellipsometric work is being carried out at Eindhoven by Dr. Bootsma and Dr. Meyer and at Hamburg by the author, working in close consultation.

Internal reflection spectroscopy (IRS)

If electromagnetic radiation is incident on the boundary between a denser medium *1* (with index of refraction n_1) and a rarer medium *2* (with smaller refractive index n_2), total reflection occurs if both media are non-absorbing and the angle of incidence is greater than the "critical angle", whose magnitude φ_{1c} is given by the relation:

$$\sin \varphi_{1c} = \frac{n_2}{n_1}.$$

It is known that in total reflection the radiation pene-

Dr. K. H. Beckmann is with the Hamburg laboratory of Philips Zentrallaboratorium GmbH.

[1] J. J. A. Ploos van Amstel, Philips tech. Rev. **22**, 204-214, 1960/61.

trates a small distance into the rarer medium. This is a complicated process and the calculation for the case of an infinitely wide beam totally reflected at a boundary between two non-absorbing media shows that inside the rarer medium there is an energy flow parallel to the boundary. The time average of this energy flow decreases exponentially with the depth z (see *fig. 1*). An additional flow across the interface with equal magnitudes in the directions $1 \rightarrow 2$ and $2 \rightarrow 1$ occurs at certain places, the whole pattern moving with velocity $c_1/\sin \varphi_1$ (c_1 velocity of light and φ_1 angle of incidence in the denser medium) parallel to the boundary. Thus both the total and the local time average values of the flow through the interface are zero. In the case of a

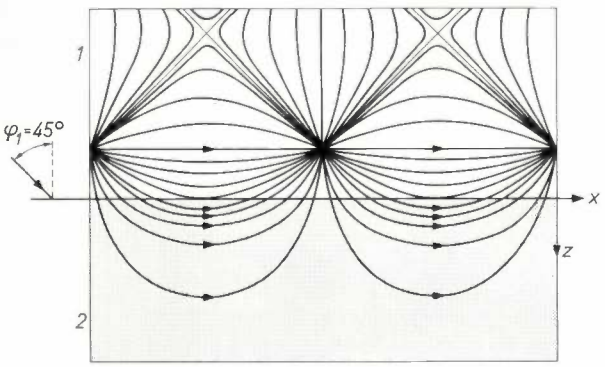


Fig. 1. Lines of energy flow at total reflection of an infinitely wide beam of light. The boundary between media *1* (the denser one) and *2* contains the x-axis and is normal to the x,z-plane. The density of the lines at each point is proportional to the local magnitude of the energy flow (instantaneous values). The indices of refraction and the angle of incidence φ_1 are chosen such that $\varphi_1 = 45° = \varphi_{1c}$ (the critical angle). The whole pattern travels with a velocity $c_1/\sin \varphi_1$ in the x-direction. (After A. Eichenwald [2].)

beam of finite width, however, there is a net flow of energy into the rarer medium in one part of the beam and a net flow from this medium in another part, but the time average of the energy flow across the interface integrated over the whole area will remain zero. If now the rarer medium is one which absorbs the radiation, the time average of the integrated energy flow into the second medium will be positive and the reflectivity will be less than unity.

Fahrenfort [3] has used this effect, which is called attenuated total reflection, for infra-red absorption measurements of materials which cannot be prepared for normal transmission measurements. Since at total reflection the penetration of the radiation into the second medium is only small, even with strongly absorbing materials, reflected intensities of appropriate magnitude are obtained. This method of measuring the absorption is now known as internal-reflection spectroscopy (IRS).

On the other hand, N. J. Harrick [4] of Philips Laboratories, Briarcliff Manor, U.S.A., has shown that the same effect can also be used for measuring the infra-red absorption of films which are too thin for the absorption to be measured by the standard transmission methods. The basic scheme is shown in *fig. 2*. The

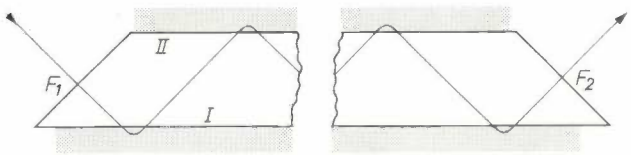


Fig. 2. Internal-reflection element, with material whose infra-red absorption is to be measured by internal-reflection spectroscopy (rarer medium, shaded) in contact with faces *I* and *II*. The diagram also shows the path of the light which arrives at F_1 from a monochromator and goes from F_2 to the detector. The fact that the radiation penetrates a small distance into the rarer medium is indicated here by the "bending" of the rays in the rarer medium. The element can be designed for up to 200 reflections.

material or the thin film to be measured is in contact with one or both sides (*I* and *II* in fig. 2) of a specially prepared internal-reflection element. A monochromatic beam of light enters the element through the entrance face F_1. An essential condition is that this element must be transparent for the radiation used. The angle of incidence on face *I* should be such as to give total reflection at this interface. With reflection elements consisting of semiconducting materials these conditions are satisfied for a wide range of angles of incidence since these materials are transparent and their indices of refraction are very high in the infra-red (e.g. $n = 4$ for germanium and $n = 3.4$ for silicon). If the material in contact with the element is one which absorbs, there is absorption on total reflection at face *I* and the unabsorbed radiation then arrives at face *II*, where the same process is repeated. Thus a greater attenuation is obtained by the repetition of the absorption process at every reflection, and after the light has traversed the element and emerged from the exit face F_2 it strikes the detector.

The radiation is only distributed homogeneously in the medium to be measured if the film is thinner than about 10 nm. With thicker films the limited penetration depth of the radiation may become apparent, and depending on the film thickness (extent of material *2* in the z-direction, see fig. 1) only part of this material will

be penetrated by the radiation. Since the time average of the energy flow density in the x-direction in the absorbing medium *2* decreases exponentially with z, it follows that the amount of the radiation absorbed also decreases exponentially with increasing z, if medium *2* is homogeneous. If now the chemical composition of this medium is not homogeneous in the direction of z the interpretation of the absorption spectrum will be very difficult.

The path of the light inside the internal-reflection element should have the same angle of incidence at every reflection, and it should remain in the plane of entry. For these conditions to be met the geometry of the element must be very accurate. The two faces *I* and *II* must be flat to within a fraction of a wavelength, and they must be exactly parallel. This can be achieved by polishing techniques. After polishing, however, the surface is coated with a thin film of grinding material and fragments of the semi-conductor. This undesirable coating can be removed from germanium reflecting elements by using a special etching process, and for silicon elements it can be thermally oxidized to an oxide film 1 μm thick, which is then etched away by concentrated hydrofluoric acid. In both cases the transmission of the elements increases although the flatness deteriorates slightly. This improvement probably comes about because there was some scattering of the radiation inside the surface layer caused by the polishing. With germanium and silicon elements 200 successive total reflections can be obtained by using an angle of incidence of 45° and elements 60 mm long and 0.3 mm thick. However, with so long a path length for the radiation inside the semiconductor the measurements can only be extended to 8-10 μm wavelength since the transmission of the element itself decreases rapidly with increasing wavelength owing to the lattice absorption.

Harrick [5][6] has worked out several other geometrical arrangements which may be of use not only in the measurements of thin films but also with solids or liquids. In order to obtain a suitable amount of absorption, it may only be necessary to cover a part of the reflecting surfaces with the material to be measured. *Fig. 3a* shows a double-pass multiple-reflection element which gives twice the number of reflections for a given length of element. This kind of element is easily dipped to the required depth into liquids for investigations of their chemical composition or of the interaction between the liquid and the element. Fig. 3*b* shows a similar arrangement which can be used for measurements at different angles of incidence.

Such a variation of the angle of incidence may be necessary if measurements of the highest possible sensitivity are required. There are several important parameters that change with the angle of incidence. The depth of penetration d_p of the radiation into the rarer medium — i.e. the distance from the totally reflecting interface at which the electric field vector of the radiation has decreased to 1/e of its original value — decreases with increasing angle. The amplitude of the electric field in the rarer medium changes with increasing angle of incidence and the area illuminated by a beam of given diameter increases. For materials of low absorption, the influence of these parameters and that of the refractive index on the degree of absorption can be expressed in an approximate formula for the effective thickness d_e [7]. Provided that the thickness of the absorbing medium is much greater than the penetration depth d_p, the effective thickness d_e can be defined as the thickness of the material which, in conventional transmission at normal incidence, would give the same absorption as a single reflection in the IRS method. *Fig. 4* shows that the value of d_e depends on the polarization of the radiation, and furthermore that it is not only primarily the influence of the variation of the depth of penetration d_p that leads to the decrease of the effective thickness with increasing angle of

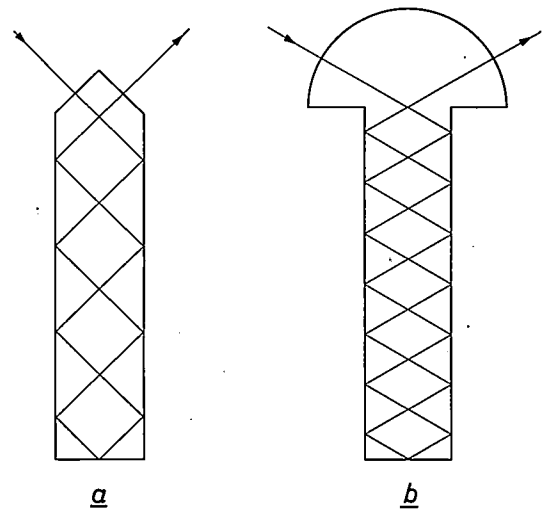


Fig. 3. Double-pass internal-reflection elements developed at Philips Laboratories, Briarcliff Manor (U.S.A.) by N. J. Harrick [5]. The element (*b*) is a more versatile type than (*a*) since the angle of incidence can be varied.

[2] Published in: Cl. Schaefer and G. Gross, Ann. Physik (4) **32**, 648-672, 1910.
[3] J. Fahrenfort, Spectrochim. Acta **17**, 698-709, 1961.
[4] N. J. Harrick, J. phys. Chem. **64**, 1110-1114, 1960.
[5] N. J. Harrick, Analytical Chem. **36**, 188-191, 1964.
[6] N. J. Harrick, Internal reflection spectroscopy, Wiley, New York 1967.
[7] N. J. Harrick and F. K. du Pré, Appl. Optics **5**, 1739-1743, 1966.

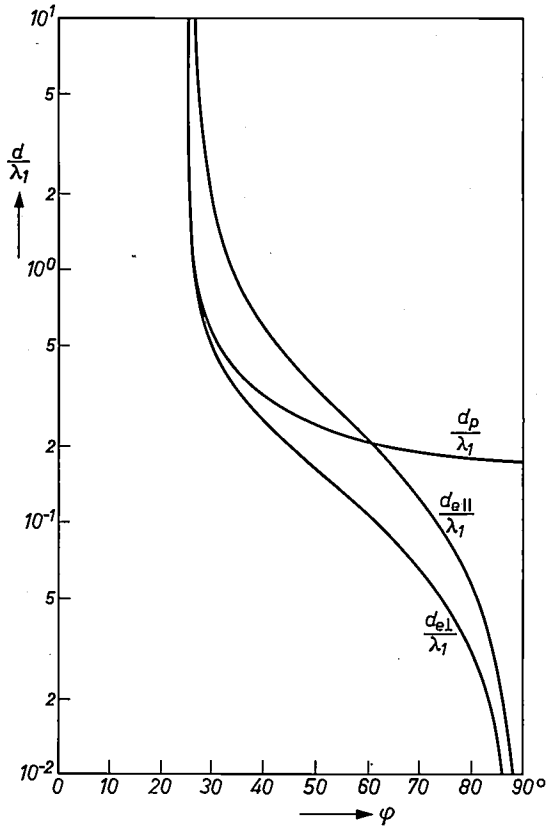


Fig. 4. Relative penetration depth into the weakly absorbing rarer medium d_p/λ_1 and effective thickness d_e/λ_1 plotted against angle of incidence φ, for an interface with a relative index of refraction $n_{21} = n_2/n_1 = 0.423$. (The subscript 1 refers to the denser medium, from which the beam is incident, and 2 to the rarer medium; λ_1 is the wavelength of the incident radiation.) The effective thickness has been defined as the thickness of the material which, in a single conventional transmission, would give the same absorption as a single total reflection in the IRS method. The value of d_e for polarization of the radiation parallel to the plane of incidence ($d_{e\parallel}$) is different from that for polarization perpendicular to the plane of incidence ($d_{e\perp}$). The plot gives an indication of the strength of the coupling between the radiation and the absorbing molecules, since the effective thickness is proportional to this coupling. (After N. J. Harrick and F. K. du Pré [7].)

incidence. *Figs. 5a* and *b* show the results of approximate calculations for different angles of incidence and refractive indices. The functions shown here are of importance in the quantitative evaluation of infra-red absorption measurements performed by IRS on bulk materials since they allow the calculation of the absorption coefficient α from such measurements. To a first approximation the intensity I at the detector is:

$$I = I_0(1 - R)^2 \exp(-\alpha N d_e),$$

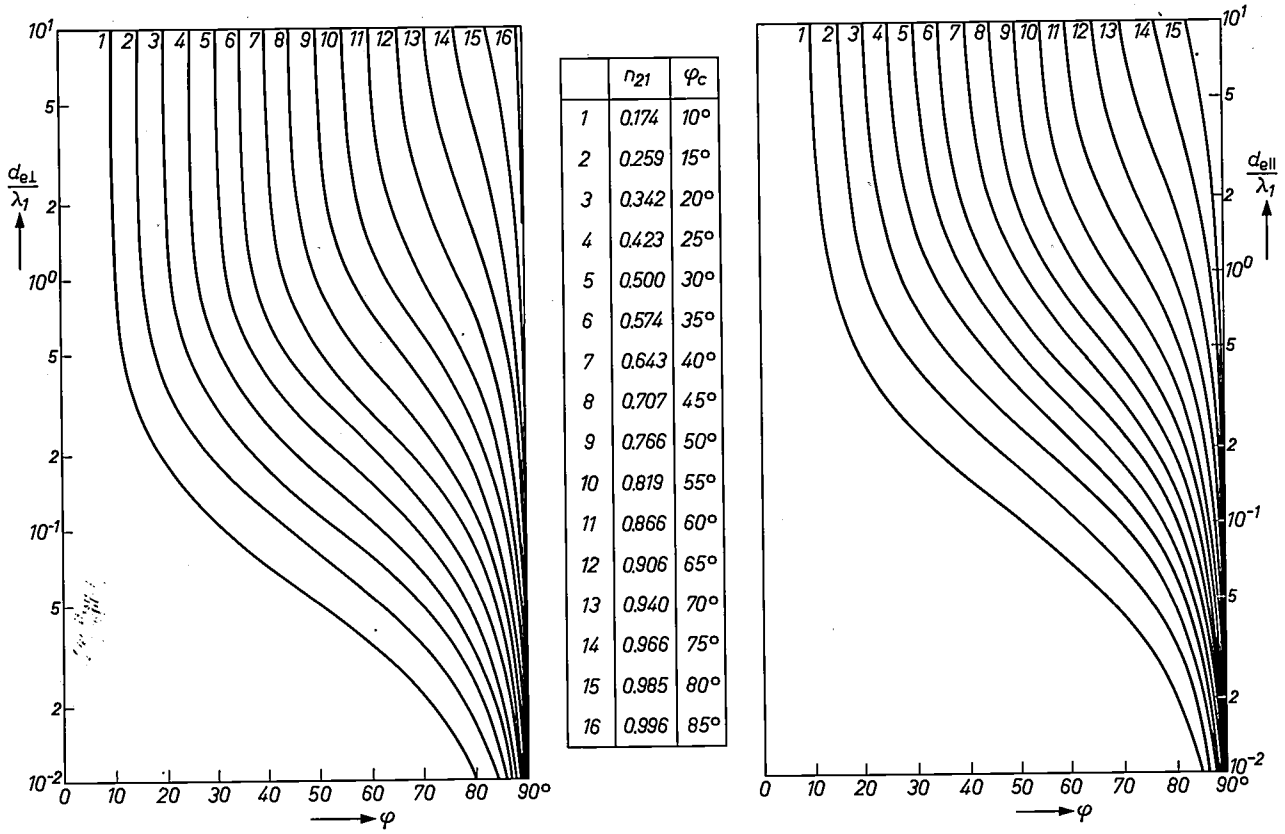
where I_0 is the intensity incident on the entrance face of the element, R the reflectivity at the entrance and exit faces, N the number of total reflections at the interfaces between the element and the absorbing material, and d_e the effective thickness.

When the spectral dependence of absorption is examined by IRS, the positions and heights of the absorption peaks may differ from those found by a single ordinary transmission. These differences arise because the refractive index of the material being examined changes with wavelength. We have seen above that one of the quantities which determine the effective thickness is the refractive index. In pure materials there may be large changes of refractive index, particularly in the range of anomalous dispersion near the wavelength of the fundamental "stretching vibrations" of molecules. An example of what may happen appears in *fig. 6* which shows the spectral dependence of the intensity R reflected from the silicon/silicon dioxide interface as calculated by a computer. The calculation was performed for a single reflection, and the rarer medium (silicon dioxide) was assumed to extend to infinity. The real and imaginary parts (n and k respectively) of the optical constant of crystalline quartz [8] for light polarized perpendicularly to the optical axis, which were used in this calculation, are also given in fig. 6. The relation between k and the absorption coefficient α is:

$$k = \alpha\lambda/4\pi.$$

In the calculated plot of R against λ there is seen to be a dip in R at 8.6 μm. This dip is due to the peak of n at the same wavelength which gives a peak in the effective thickness. The position of the maximum of k differs markedly from that of the minimum of R between 9 and 10 μm, which is due to the marked increase in the effective thickness above 9 μm wavelength resulting from the increase of n. For the example given in fig. 6 with an angle of incidence of 45° the critical angle is already exceeded when the refractive index of the oxide reaches a value of 2.5. Near 9.7 μm R is equal to zero although k has already vanished. The reason for this is that the real parts of the optical constants of both media are equal at that wavelength, so that no reflection occurs at the interface. At longer wavelengths R increases with decreasing n for the oxide, since total reflection is now being re-established. It should be recalled, however, that the decrease of R to zero only comes about because the rarer medium has been assumed to be of infinite extent in this calculation. If there is only a thin film of oxide present, then if the film has a high refractive index there will be total reflection at the back of the film and the reflectivity will in general be greater than zero. If the film thickness is much smaller than the wavelength, R is much less affected by the variation of n, since in this case the distribution of the radiation inside the thin film remains practically unaffected by the variation of the depth of penetration.

From figs. 4 and 5 it can be seen that the greatest sensitivity is obtained if the angle of incidence is near



	n_{21}	φ_c
1	0.174	10°
2	0.259	15°
3	0.342	20°
4	0.423	25°
5	0.500	30°
6	0.574	35°
7	0.643	40°
8	0.707	45°
9	0.766	50°
10	0.819	55°
11	0.866	60°
12	0.906	65°
13	0.940	70°
14	0.966	75°
15	0.985	80°
16	0.996	85°

Fig. 5. Relative effective thickness d_e/λ_1 plotted against angle of incidence φ for different relative indices of refraction n_{21}. These curves permit the proper choice of the angle of incidence to obtain the desired degree of attenuation of the radiation, and they are useful in quantitative evaluations of absorption measurements by IRS. *a*) Polarization perpendicular to plane of incidence and *b*) parallel to plane of incidence. The critical angle φ_c for each value n_{21} is also indicated. (After N. J. Harrick and F. K. du Pré [7].)

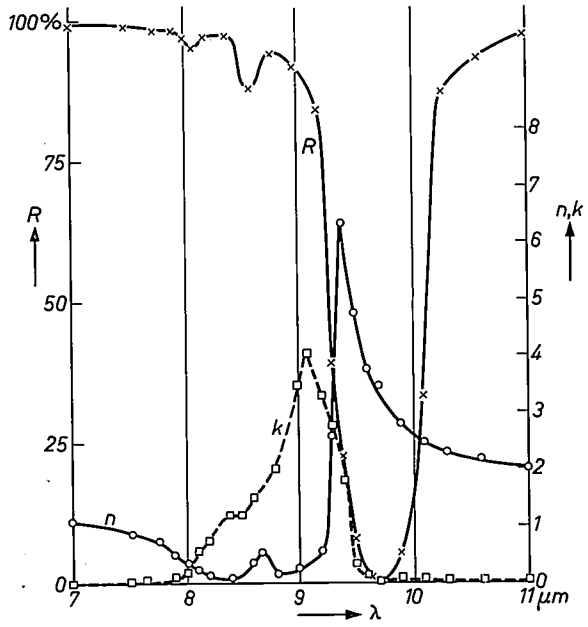


Fig. 6. Change of n and k with wavelength for α-quartz (from [8]). The reflectivity R has been calculated from these values as a function of wavelength [9] for total reflection at the boundary between silicon and silicon dioxide with an angle of incidence of 45°. The rarer medium (silicon dioxide) has been assumed to extend to infinity.

to the critical angle. Another way of increasing the degree of absorption is by using reflecting elements of such dimensions that the radiation forms a standing wave with high amplitude of the electric field at the boundary of the absorbing medium [10]. A greater absorption can also be attained if the amplitude and phase of the reflected beam in the denser medium are matched to those of the "transmitted" beam in the rarer medium. The matching can be achieved by coating the internal reflection element with three layers: first, an absorbing film for amplitude matching, next, an interference film for phase matching, and then a layer of the absorbing material to be investigated. This matching technique, however, is only effective for one wavelength [11].

The experimental techniques which can be used for measurements of infra-red absorption by internal-reflection spectroscopy are the same as those used for normal transmission measurements. Single or double

[8] W. G. Spitzer and D. A. Kleinman, Phys. Rev. **121**, 1324-1335, 1961.
[9] The formulae for this calculation are given by K. Gottlieb and B. Schrader, Z. anal. Chemie **216**, 307-316, 1966.
[10] F. Berz, Brit. J. appl. Phys. **16**, 1733-1738, 1965.
[11] N. J. Harrick and A. F. Turner, J. Opt. Soc. Amer. **56**, 553-554, 1966.

beam operation is possible, and differential techniques can also be used (see *fig. 7*). The high sensitivity of this technique can be demonstrated by exposing one of the two internal reflection elements in the arrangement shown in fig. 7 to methanol vapour. The thin invisible film of methanol then produced by adsorption of vapour molecules on the element introduces an absorption due to the OH stretching vibration, and this absorption gives a full-scale deflection of the recorder pen, without unusually high amplification being required (signal-to-noise ratio of 15).

Fig. 7. Arrangement for the measurement of the infra-red absorption of thin films by internal-reflection spectroscopy by a differential technique. The monochromatic beam from the monochromator *Mon* is focused by the spherical mirror M_1 and the plane mirrors M_2, which also serve as beam splitter, on the entrance faces of the two internal-reflection elements $IR_{1,2}$. In the case shown these elements consist of single-crystal silicon and their dimensions are $0.3 \times 15 \times 60$ mm. One of these elements is coated with the film under investigation, and the other is left uncoated. The radiation leaving each element is focused on to one of the thermocouples *Th* by means of ellipsoidal mirrors M_3. The difference between the two thermoelectric outputs produced at the thermocouples by the incident radiation is then amplified and displayed on a recorder.

IRS is particularly useful for novel applications of what is known as *effect modulation*. Modulation of the light intensity is in general a valuable means of eliminating background and other disturbances in absorption measurements, since with modulated light an a.c. amplifier tuned to the modulation frequency only amplifies the signal due to the change of intensity of the light beam. Periodic chopping of the incident light beam is often used for achieving a modulation; periodic variation of a parameter on which the absorption depends — if such a parameter exists — may also be used, and this is what is called effect modulation. Since such a process alters the degree of absorption (chopping of course does not do this) a signal is obtained which is directly proportional to the absorption to be studied. If this method is applied with IRS, minute relative changes of the absorption can be amplified by using a large number of reflections. This signal can be amplified further by a tuned or lock-in amplifier. At wavelengths where there is no absorption to be modulated the signal will be zero. Thus a large gain in sensitivity over the simple chopping method is obtained, since with simple chopping of the beam there is a large signal proportional to the intensity incident on the detector, and small absorptions give only small changes in this

intensity. Effect modulation can be used in examining either the dependence of the absorption on the modulating parameter itself, or, indirectly, its dependence on (say) wavelength.

As we have seen above, one of the most interesting features of the IRS technique is that it can be used for measuring the infra-red absorption of thin films and of materials which cannot easily be prepared for conventional transmission measurements. There may, however, be small errors in the determination of the position of the absorption peaks if the dispersion curve of the material being investigated is not known, and this will be very often the case. A knowledge of the refractive index is also necessary if a quantitative evaluation of the absorption measurements is required in terms of the number of molecules forming a surface film or of the film thickness. Such a quantitative evaluation is required, for example, when studying the kinetics of surface coverages and film growth. In these cases it is more convenient to use the method of ellipsometry which will be described in the next section.

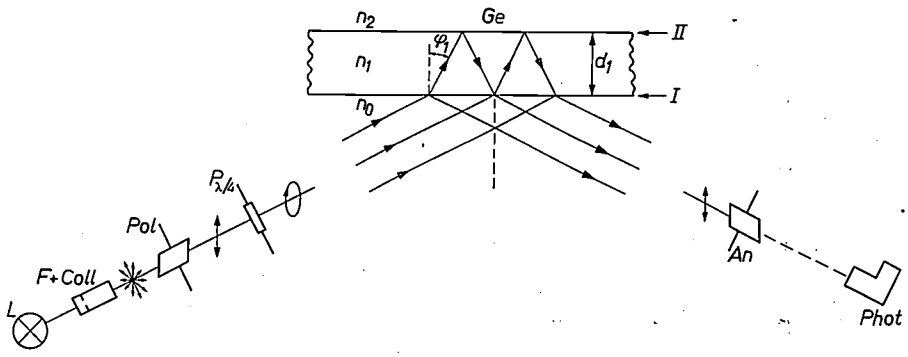


Fig. 8. Arrangement for ellipsometric measurements as used at Philips Research Laboratories in Eindhoven [12]. A parallel beam of linearly polarized monochromatic radiation obtained from a mercury lamp L with filters and collimator ($F + Coll$) is converted into elliptically polarized radiation by means of a polarizer *Pol* and a quarter-wave plate $P_{\lambda/4}$. By varying the azimuth of the polarizer any desired phase shift between the components of the light vector parallel and normal to the plane of incidence can be produced. The setting of the polarizer has to be such that the phase shift it produces is balanced by the phase shift due to the reflection at the germanium sample. Thus the reflected light is again linearly polarized, and this plane of polarization is then determined with the help of an analyser *An*. The light transmitted by *An* is detected by a photomultiplier *Phot*.

Ellipsometry

This method makes use of the change in polarization of light which occurs on reflection at a surface. The polarization is characterized by the phase difference and amplitude ratio of the two components (in and normal to the plane of incidence) into which the electric vector of the light wave may be resolved. If light is reflected from a surface coated with a film, the change of polarization is determined by the thickness and refractive index of the film. The reason for this dependence is that in a film whose thickness is much less than the length of coherence there are multiple reflections, giving discrete phase differences between beams which have traversed the thin film different numbers of times (see *fig. 8*). This will in general result in elliptically polarized light, and this is the reason why the method is called ellipsometry. The theoretical background of the method will now be discussed.

The reflection of light from a surface coated with a film can be described, just as in the case of an uncoated surface, by amplitude reflection coefficients $\mathbf{R}$ which may be complex numbers and which are generally different for the two directions of polarization parallel and perpendicular to the plane of incidence. The reflected amplitudes of both components are $\mathbf{E}^{(r)} = \mathbf{R}E^{(i)}$, where $E^{(i)}$ is the magnitude of the appropriate component of the electric vector of the incident wave, assumed for simplicity to be real. The writing of $\mathbf{R}$ or $\mathbf{E}^{(r)}$ as complex quantities indicates that there is a phase shift at this reflection, given by the argument of $\mathbf{R}$. For a thin film of thickness d_1 and refractive index n_1 whose amplitude reflection (Fresnel) coefficients for the two sides are $\mathbf{r}_I$ and $\mathbf{r}_{II}$ the amplitude reflection coefficient $\mathbf{R}$ is given by:

$$\mathbf{R} = \frac{\mathbf{r}_I + \mathbf{r}_{II} \exp\left(-\frac{4\pi j}{\lambda} n_1 d_1 \cos\varphi_1\right)}{1 + \mathbf{r}_I \mathbf{r}_{II} \exp\left(-\frac{4\pi j}{\lambda} n_1 d_1 \cos\varphi_1\right)}, \quad (1)$$

where j is the imaginary unit, λ the wavelength of the radiation *in vacuo* and φ_1 the angle of refraction in the thin film (see fig. 8). If we introduce the Fresnel coefficients for parallel or perpendicular polarization, we

obtain **R** for the appropriate components of the light vector. Even if the surrounding medium, the thin film and its substrate are non-absorbing, which implies real optical constants and Fresnel coefficients, **R** will in general be complex and there will be a phase shift due to reflection. The expression for the amplitudes of both components after reflection is (see also *fig. 9*):

$$\left.\begin{aligned}\mathbf{E}_{\perp}^{(r)} &= \mathbf{R}_{\perp}E_{\perp}^{(i)} = R_{\perp}\exp(j\delta_{\perp})\,E_{\perp}^{(i)}\\ \mathbf{E}_{\parallel}^{(r)} &= \mathbf{R}_{\parallel}E_{\parallel}^{(i)} = R_{\parallel}\exp(j\delta_{\parallel})\,E_{\parallel}^{(i)}\end{aligned}\right\}, \quad (2)$$

expressing **R** by its magnitude and argument. From eqs. (2) we notice that after reflection there exists a phase difference Δ between the components of the electric vector:

$$\Delta = \delta_{\parallel} - \delta_{\perp},$$

and this generally means that the light is elliptically polarized. The magnitudes of the components after reflection are:

$$|\mathbf{E}_{\perp}^{(r)}| = R_{\perp}E_{\perp}^{(i)} \quad \text{and} \quad |\mathbf{E}_{\parallel}^{(r)}| = R_{\parallel}E_{\parallel}^{(i)}.$$

This phase difference Δ and the ratio of the magnitudes $|\mathbf{E}_{\parallel}^{(r)}|/|\mathbf{E}_{\perp}^{(r)}|$ from which one easily obtains $R_{\parallel}/R_{\perp}$ when $E_{\parallel}^{(i)}/E_{\perp}^{(i)}$ is known, can be measured. These two measured quantities allow the determination of two unknown parameters which appear in eq. (1), for example the thickness d_1 and the refractive index n_1 of the film.

The polarization can be determined in the following way [12]. A parallel beam of monochromatic light is linearly polarized by a Glan-Thomson prism (see fig. 8) and then passes through a quarter-wave plate whose fast axis is fixed at 45° to the plane of incidence. With this arrangement the light incident on the sample is in general elliptically polarized. The phase shift Δ which occurs on reflection can be determined by compensation. The rotatable polarizer is for this purpose set so as to cause the radiation after the reflection at the sample to be linearly polarized again. The achievement of complete compensation is indicated by the best possible extinction of the radiation behind the crossed analyser being obtained. The phase difference which the quarter-wave plate introduces between the components of the light vector of the incident beam polarized in and perpendicular to the plane of incidence is then equal in magnitude and opposite in sign to the phase shift produced by the reflection. The azimuth of the restored linear polarization, which obeys the relation:

$$\tan(90° - \gamma) = \frac{|\mathbf{E}_{\parallel}^{(r)}|}{|\mathbf{E}_{\perp}^{(r)}|} = \frac{R_{\parallel}}{R_{\perp}}\frac{E_{\parallel}^{(i)}}{E_{\perp}^{(i)}} = \tan\psi\frac{E_{\parallel}^{(i)}}{E_{\perp}^{(i)}}, \quad \ldots \quad (3)$$

is then determined by the analyser setting required for extinction; the ratio $E_{\parallel}^{(i)}/E_{\perp}^{(i)}$ in the formula is given by the tangent of the azimuth of the polarizer. The whole measurement is carried out with a photomultiplier as detector.

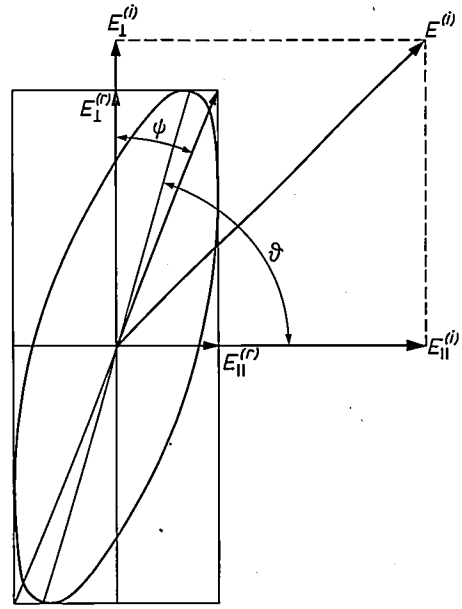


Fig. 9. Polarization before and after reflection at an interface with complex reflection coefficient. The light vector $\mathbf{E}^{(i)}$ of the incident linearly polarized wave is made up from the two components $E_{\parallel}^{(i)}$ and $E_{\perp}^{(i)}$, whose magnitudes change to $E_{\parallel}^{(r)}$ and $E_{\perp}^{(r)}$ upon reflection. Since there is a phase shift between the two components on reflection the reflected light will be elliptically polarized.

Another experimental arrangement, which is used at the Philips laboratories in Hamburg for the study of electrode reactions in electrolytes [13], is shown in *fig. 10*. A monochromatic beam of plane-polarized light falls on the surface to be investigated. In fig. 10 this surface is that of an electrode in contact with an electrolyte in a cell; this cell is therefore furnished with appropriate windows. Electrochemical polarization with the aid of a counter electrode is also possible. After reflection the light will in general be elliptically polarized. This phase shift can be cancelled by a Soleil-Babinet compensator, which consists of two quartz wedges and a single quartz plate with a certain orientation of their optical axes; the arrangement is shown in fig. 10. The larger wedge can be shifted with a precision adjustment, thus forming, with the small wedge of the same orientation, a birefringent plate of variable thickness which will give any desired phase difference between the two components. The elliptically polarized light which arrives at the compensator can thus be re-converted into linearly polarized light. Since the compensator has been calibrated, the phase shift Δ produced on reflection can be determined by this procedure. In this apparatus the measurements are performed with the aid of an elliptical half-shade device

in front of the analyser. The half-shade, when viewed through a telescope, shows clearly whether it is illuminated by linearly polarized radiation or not. Finally the azimuth of the plane of polarization of this linearly polarized light which leaves the compensator is determined with a Lippich half-shade in front of the analyser.

With this method the polarizer may be set to a fixed azimuthal angle of 45°, so that we always have (see fig. 9):

$$E_{||}^{(i)} = E_{\perp}^{(i)}.$$

After compensation of the phase difference the azimuth of the restored linear polarization as given by eq. (3) is equal to $90° - \psi$.

It follows from eqs. (3), (2) and (1) that the measured quantities Δ and ψ depend on the amplitude reflection coefficients $\mathbf{R}_{||}$ and $\mathbf{R}_{\perp}$ given in eq. (1), and these in turn depend on the optical constants of the substrate, the thin film and the ambient medium. In these investigations of films the optical constants of the substrate and the ambient medium are usually known so that two unknown quantities, e.g. the thickness and the refractive index of the film, can be determined from one measurement of ψ and Δ. But the evaluation of such a measurement is difficult, since it is impossible to insert Δ and ψ into the equation

$$\mathbf{R}_{||}/\mathbf{R}_{\perp} = \exp(j\Delta) \tan\psi, \quad \ldots \quad (4)$$

and to solve for n_1 and d_1 the equation obtained by

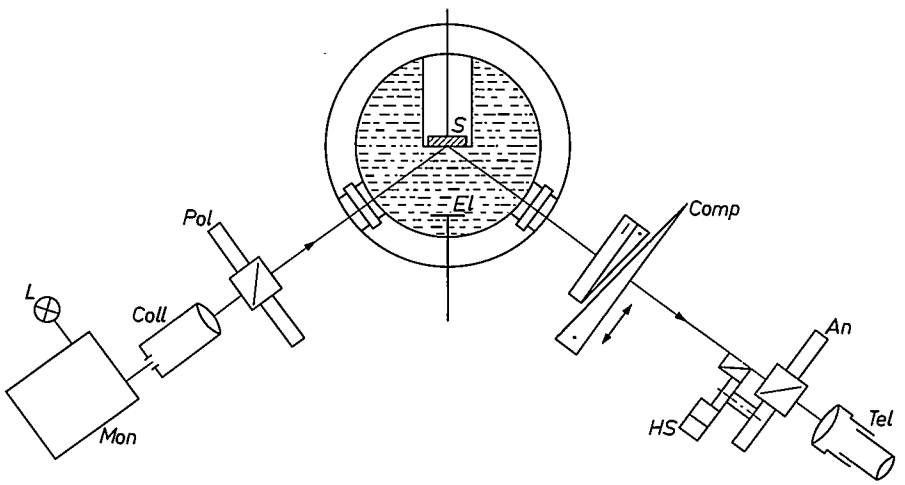


Fig. 10. Apparatus for the measurement of the polarization of light reflected from an electrode surface *S* [13]. In this arrangement a parallel beam of monochromatic *linearly* polarized light is reflected by the specimen. The polarization after reflection will in general be elliptical. This elliptically polarized light is converted into linearly polarized light by means of a compensator *Comp*. The azimuth of this restored linear polarization is then measured with an analyser *An*. *L* light source. *Mon* monochromator. *Coll* collimator. *Pol* polarizer. *HS* half-shade device, *Tel* telescope for visually checking the compensation and extinction. *El* counter electrode for performing experiments on film growth on the sample *S*.

Investigations at Philips Research Laboratories in Eindhoven have shown [12] that the phase shift can be measured with an accuracy of $\pm 0.02°$ and the azimuth to $\pm 0.01°$ in this arrangement (fig. 9) by the use of a photomultiplier as a detector. In investigations performed at the Philips laboratories in Hamburg [13] we used the visual method given in fig. 10 since the materials under investigation often tended to form inhomogeneous films, and this can be recognized when observing the half-shade which is illuminated by a parallel beam of light reflected from the sample. Then if there is inhomogeneity the measurements can be restricted to a part of the surface. The accuracy is $\pm 0.3°$ for the phase shift and $\pm 0.05°$ for the azimuth when the visual method is used.

combining (4) and (1). It is therefore necessary to calculate the values of Δ and ψ by computer, for given optical constants. These calculated results demonstrate that for given optical constants of the substrate and of the ambient medium, and for a fixed angle of incidence, n_1 and d_1 are unique functions of the measured quantities of Δ and ψ in a wide range. This is demonstrated in *fig. 11* for $1 < n_1 < 4$. If the film absorbs, the determination of a third quantity, the absorption coefficient, is necessary, but this can only be done with an additional measurement (e.g. an independent measurement of the thickness) or by measurements of Δ and ψ

[12] G. A. Bootsma and F. Meyer, Surface Sci. 7, 250-254, 1967 (No. 2).
[13] K. H. Beckmann, Ber. Bunsenges. phys. Chemie 70, 842-849, 1966.

while the film is growing, and subsequent comparison of this measured curve with calculated ones.

When very thin films ($d_1 \leqq 10$ nm) are to be investigated, approximate formulae for the dependence of Δ and ψ on n_1 and d_1 can be applied. In these approximations the changes of Δ and ψ as a result of the given set of the other parameters, shows that for very thin films (with $\delta < 10°$; for the definition of δ see fig. 11) under the condition of fig. 11 the variation of Δ with d_1 is one to two orders of magnitude greater than that of ψ. In this case, therefore, hardly any change of ψ is to be seen with increasing d_1. But since within

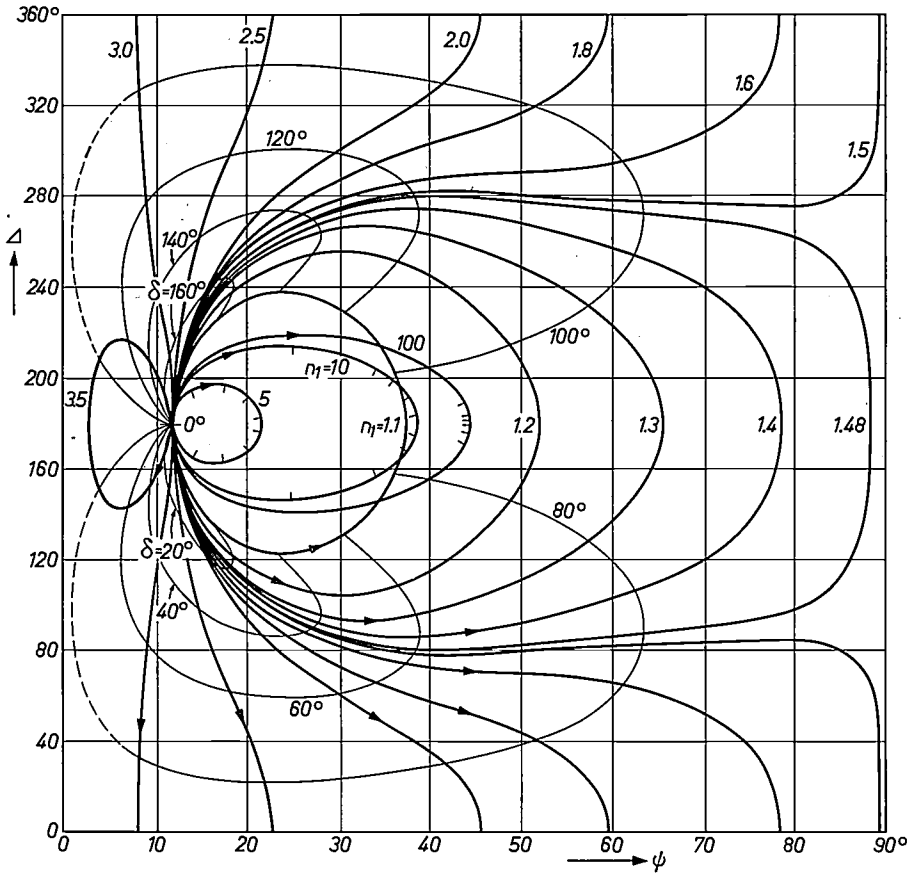


Fig. 11. Plot of Δ against ψ as calculated by computer for thin films of different indices of refraction (n_1) on silicon. The angle of incidence φ_0 is 70°, the wavelength of the radiation λ (*in vacuo*) 546 nm. The arrows indicate the direction of increasing film thickness (d_1). The variable parameter δ for each curve is the extra phase shift (in degrees) which occurs in components of either polarization $E_\perp$ and $E_\parallel$ for a single passage through the thin film: $\delta = (360/\lambda)\, d_1 \sqrt{(n_1^2 - \sin^2 \varphi_0)}$. Lines δ = constant are drawn for easier reference (thin lines). This plot shows the exact periodicity of the relation between Δ and ψ since the curve repeats itself for $\delta > 180°$. (After R. J. Archer [14].)

presence of the thin film on the substrate surface are expressed as linear functions of d_1, whose slopes α and β are related to the refractive index of the film n_1:

$$\delta\Delta = \alpha d_1, \quad \delta\psi = -\beta d_1.$$

With the parameters used in fig. 11 we obtain:

$$\alpha \approx 3°/\text{nm} \quad \text{and} \quad \beta \approx 0.03°/\text{nm} \quad \text{for} \quad n_1 = 1.55.$$

Similar approximations exist for weakly absorbing films where the slope is determined by n_1 and the absorption constant k_1. An examination of fig. 11, which shows the dependence of Δ and ψ on n_1 and d_1 for the certain limits the change of Δ with d_1 is fairly independent of n_1, the thickness of very thin films can be determined by measuring the variation of Δ alone.

If the experimental error — which determines the minimum angle that can be measured — is used to calculate the order of magnitude of the smallest film thickness that can be determined by this method, this is found to be 0.01 nm. Now it is known that even a monoatomic layer has a thickness of 0.2-0.3 nm (2-3 Å), so that a film as thin as 0.01 nm seems rather puzzling.

An appropriate model of such a "sub-monolayer" is one of a two-dimensional distribution of oscillators [15].

Within the scope of this model it is reasonable to assume [16] that an incomplete but homogeneous monoatomic film, in which only a fraction Θ of the possible sites is occupied by the adsorbed molecules, shows the same optical behaviour as a continuous film of thickness Θd_m, where d_m is the thickness of a monolayer of the closest possible packing. This means that a linear relationship can be assumed to exist between Θ and the variation of Δ. From this it can be seen that the absorption can be measured by ellipsometric methods even for such low coverages [17].

Applications of the methods described

Applications of internal-reflection spectroscopy

Internal-reflection spectroscopy has been used at the Philips laboratories in Hamburg for a study of chemical reactions between semiconductors and electrolytes. For many etching solutions the products of the reaction with a semiconductor are deposited on the surface of the semiconductor, since they are not very soluble in these solutions. In these experiments the chemical reactions occurred at the surfaces of internal-reflection elements prepared from the semiconducting material under investigation.

Fig. 12 gives a comparison of two spectra of reaction films of this type obtained by different techniques on the same element. Curve *a* was measured by normal transmission of the light through the films on both sides of the element, whereas curve *b* was obtained by IRS with 50 reflections. The film consisted of a compound containing germanium-oxide and germanium-hydride bonds which had been produced on the single-crystalline germanium reflection element by immersing it in an etching liquid consisting of 1 part of a concentrated solution of H_2O_2 and 10 parts of a concentrated solution of HF [18]. With the transmission technique only two flat absorption bands at 825 and 690 cm^{-1} can be observed, but with internal-reflection spectroscopy a greater number of absorption bands are revealed. As far as we can tell there is no difference in the position of the absorption peaks at 825 and 690 cm^{-1} measured with the two different methods. *Fig. 13* shows the absorptions obtained with different numbers of reflections and with different thicknesses of a film produced by thermal oxidation. In this case the layer to be investigated was composed of germanium dioxide which had been produced by heating the germanium internal-reflection element in air. The absorption observed is due to the GeO stretching vibration, and it has been corrected for the absorption by built-in oxygen. From the degree of absorption it can be concluded that for this case the absorption of a film of only a few molecular layers is sufficient to be recorded. Extrapolation of this result shows that for this amount of absorption per oscillator it is possible to measure the absorption spectrum of a monolayer if a higher number of reflections, e.g. 200, is used.

As an example of effect modulation we should now like to discuss the infra-red absorption by the free charge carriers in semiconductors. As has been shown

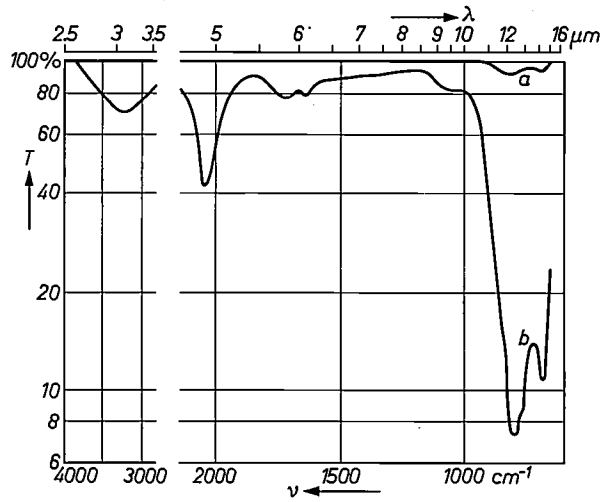


Fig. 12. The spectral dependence of transmission T of a film consisting of germanium dioxide and germanium hydride on a germanium single crystal.
a) Single transmission.
b) IRS, with 50 reflections.

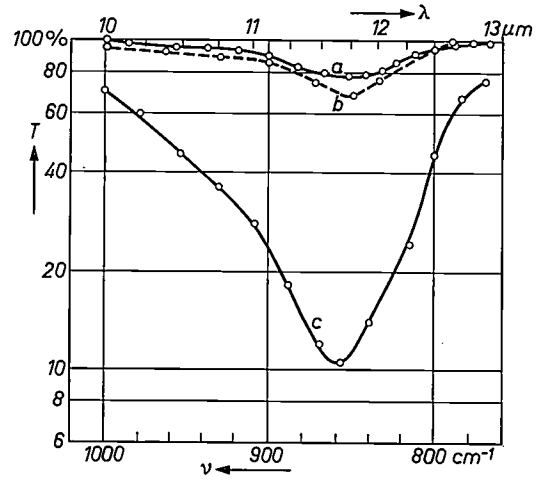


Fig. 13. As fig. 12 for thermally produced germanium dioxide films on Ge, with different conditions of measurement. The dip in the spectrum is the absorption band due to the GeO stretching vibration.
a) Thickness of the film 1.5 nm, 50 reflections.
b) Thickness 1.5 nm, 100 reflections.
c) Thickness 30 nm, 50 reflections.

[14] R. J. Archer, J. Opt. Soc. Amer. **52**, 970-977, 1962.
[15] C. Strachan, Proc. Cambr. Phil. Soc. **29**, 116-130, 1933. D. V. Sivukhin, Sov. Phys. JETP **3**, 269-274, 1956/57.
[16] R. J. Archer and G. W. Gobeli, J. Phys. Chem. Solids **26**, 343-351, 1965.
[17] For a detailed discussion of ellipsometry in the sub-monolayer region, see G. A. Bootsma and F. Meyer, to be published in Surface Science.
[18] K. H. Beckmann, Surface Sci. **5**, 187-196, 1966.

earlier [19] the density of the free electrons and holes at the surface of a semiconductor in contact with an electrolyte can be varied by a change of the potential difference between the semiconductor and a counter electrode, i.e. by applying a voltage between both electrodes. It is known that electrons and holes absorb infra-red radiation with a characteristic spectral dependence, which is different for both types of charge carrier. If we now "illuminate" a semiconductor acting as an internal-reflection element, which is immersed in an electrolyte, with a constant light level (unchopped radiation) and change the potential of this electrode by applying an a.c. voltage, the carrier density in the space-charge region of the semiconductor changes at the same frequency. This gives rise to an in-phase variation of the infra-red absorption and periodic variation in the intensity of the transmitted light. This part of the signal alone is then amplified and recorded. Curve *a* in *fig. 14* shows the absorption spectrum of holes in germanium, measured by IRS. This spectrum was obtained by applying a bias to the semiconductor electrode, which bends the energy bands upwards. The absorption spectrum of the electrons (*b*) was obtained by bending the bands downwards by applying a negative bias. In both cases a small a.c. voltage for the modulation of the carrier density was superimposed on this bias. Besides the great increase in sensitivity, effect modulation here offers the possibility of measuring the absorption of both types of charge carrier separately, which normally requires two separate samples, one of *N*-type and one of *P*-type material. Since the absorption per charge carrier, the "absorption cross-section", is known, it can be deduced from fig. 14 that the variation in the density of holes due to the a.c. voltage is 4×10^{11} cm^{-2} (curve *a*) and that of electrons is 4×10^{12} cm^{-2} (curve *b*). On the other hand, it was possible to deduce the values of the band bending from the magnitude of the applied voltages. This permits the determination of the change of the carrier density associated with the band bending and these values were found to agree fairly well with the densities determined from the optical absorption experiments.

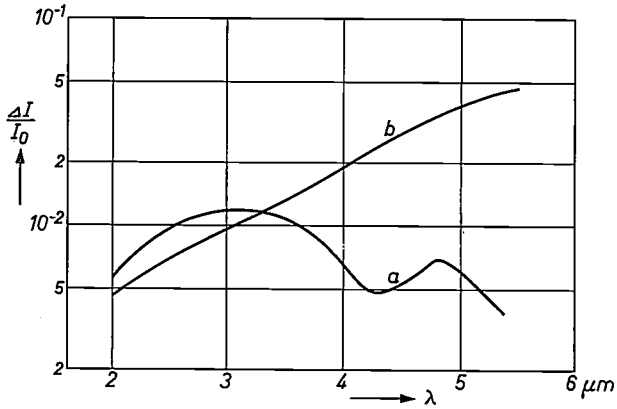


Fig. 14. Spectral dependence of the infra-red absorption of free charge carriers. (I_0 incident intensity, ΔI attenuation observed.) Measured by IRS with a modulation of the density of holes (*a*) and electrons (*b*) in the space-charge region of intrinsic germanium in contact with an electrolyte.

Applications of ellipsometry

Ellipsometric measurements of the physical adsorption of inert and other gases on faces of single-crystalline silicon and germanium under equilibrium conditions have been performed at Philips Research Laboratories in Eindhoven [20]. The number of the adsorbed atoms was determined volumetrically with a powder of the same substance. In order to maintain identical conditions the single-crystalline surface on which the ellipsometric measurements were to be performed was created by cleavage in air, and the powder for the volumetric determination by crushing single-crystalline material in air. The ellipsometric data and the amount of adsorbed gas were then determined simultaneously in the same cell at the temperature of liquid nitrogen or oxygen. The measurements were carried out under equilibrium conditions in the pressure range 10^{-4} to 10^{-1} torr. *Fig. 15* shows a representative result, namely the adsorption of krypton on germanium. The change in phase shift $\delta\Delta$, within the limits of error, is seen to be a linear function of the degree of coverage Θ which was determined volumetrically. This is in accordance with the assumption made within the scope of the two-dimensional model for a sub-monoatomic layer. The slope of the line gives information about the polarizabilities of the adsorbed atoms.

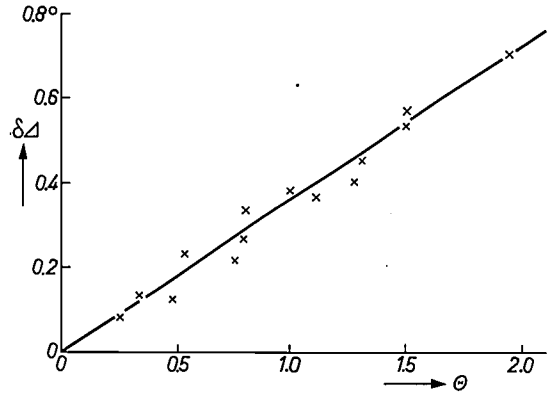


Fig. 15. Change $\delta\Delta$ of the phase shift plotted against the degree of coverage Θ (in monolayers) during the adsorption of krypton on germanium at the temperature of liquid nitrogen, as investigated at Philips Research Laboratories in Eindhoven [20]. All data were taken under equilibrium conditions by varying the gas pressure in the cell containing the germanium specimen.

[19] H. U. Harten, R. Memming and G. Schwandt, Philips tech. Rev. 26, 127-135, 1965.
[20] See the article referred to under [17].

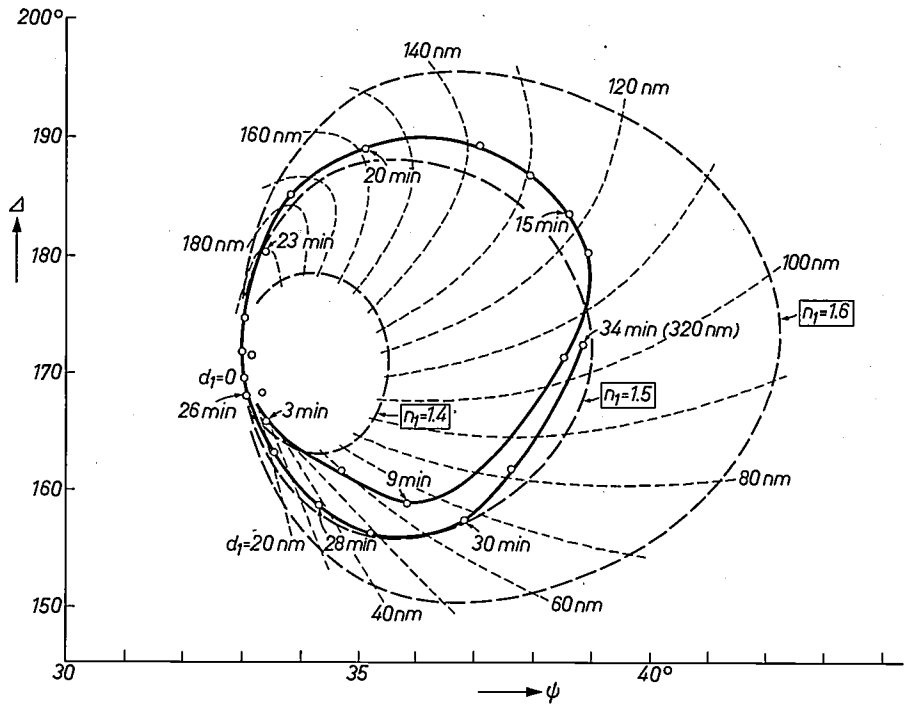


Fig. 16. Changes in the polarization of reflected light, plotted in the same way as in fig. 11, during the growth of a surface film of acid potassium germanate on a *P*-type germanium electrode, anodically polarized in an aqueous solution of potassium hydroxide and a conducting salt [13]. Solid line: measured values. After 10 minutes, at a thickness of 210 nm, the curve reaches its starting point and then repeats itself, all thickness values now being larger by 210 nm. Thick dashed lines: calculated $\Delta(\psi)$ dependence for a given index of refraction; only real values of n_1 need to be considered. Thin dashed lines: curves for constant film thickness. Angle of incidence 50°, wavelength of the light 565 nm.

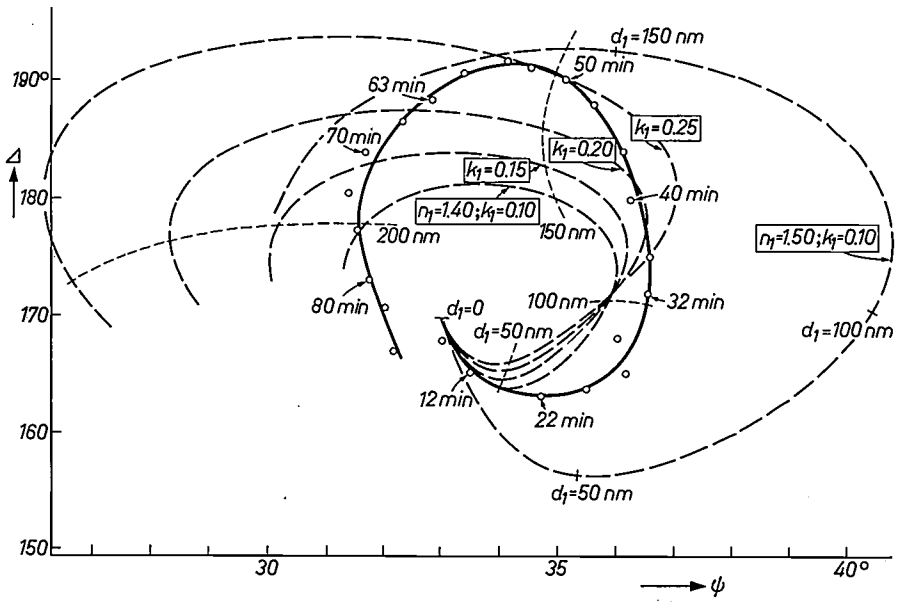


Fig. 17. The same experiment as in fig. 16, but now with film growth in an aqueous solution of caesium hydroxide and a conducting salt. Solid line: measured value. Thick dashed lines: calculated dependence of $\Delta(\psi)$ for a given index of refraction; since we have now an absorbing medium, complex values $n_1 + jk_1$ have to be considered. Curves are shown for $n_1 = 1.40$ with different values k_1, and for $n_1 = 1.50$, $k_1 = 0.10$. Thin dashed lines: curves for constant film thickness.

Results of an application of ellipsometry in some electrochemical investigations on semiconductor surfaces [13] carried out at the Philips laboratories in Hamburg are illustrated in *fig. 16*. It had been found that during anodic polarization of *P*-type germanium electrodes in alkaline solutions, there was a growth of film on the electrodes in a certain range of potential. This growth could be followed by ellipsometric measurements. The values of Δ and ψ measured during the film growth are plotted in fig. 16 (solid line). Each thick dashed curve represents the calculated changes of Δ and ψ with the thickness of the film for one value of the index of refraction n_1 of the film; this calculation was carried out on a computer. Since the substrate (germanium) is optically absorbing, the value of Δ for the case of no film ($d_1 = 0$) differs from 180°. The lines of equal film thickness are also drawn (thin dashed curves). It can be seen from the solid line that the measured behaviour of the film during the first 10 minutes can be described with an index of refraction of $n_1 = 1.40$ at the start, which then increases with increasing film thickness. This change may perhaps be caused by a variation in the film composition or by a change in its structure (porosity). Beyond a thickness of 70 nm, which is reached after ten minutes, the value of the refractive index is $n_1 = 1.50 \pm 0.01$. At a thickness of 210 nm, the curve

reaches its starting point and proceeds from there in approximately the same way. When the thickness had reached 320 nm, the measurements were stopped since the further growth of the film was too inhomogeneous.

Apart from the kinetic data which may be obtained from this curve the fact that the measured curve is closed leads to the conclusion that the film is non-absorbing. But the value of the refractive index determined by this measurement does not allow a determination of the chemical composition of the film since this constant is not specific enough. This value is however of some use in comparative studies, as may be seen from a comparison of fig. 16 and *fig. 17*. This plot of Δ against ψ, as measured during anodic polarization in an aqueous solution of caesium hydroxide (solid line), does not form a closed curve but it is similar to those calculated with complex refractive indices of the film (heavy dashed lines). In these calculated curves the complex index of refraction has been used: $\mathbf{n}_1 = n_1 - jk_1$, where n_1 is the real part and k_1 the absorption coefficient. From the different results shown in figs. 16 and 17 the conclusion can be drawn that to a certain extent the alkali atoms present in the electrolyte determine the chemical character of the films produced. It could be shown by measurements of the infra-red absorption by IRS that the films consisted of acid alkali germanate, whose refractive index in the visible spectral region depends on the kind of alkali present in this compound.

Summary. Measurements of infra-red absorption by internal reflection spectroscopy (IRS) permit the determination of the chemical composition of thin films down to thicknesses of less than 1 nanometre. Electronic processes at the surface of solids such as changes of the carrier density in the space-charge region of semiconductors can also be measured by this method. The ellipsometric determination of film thicknesses is a very sensitive method which can readily be used for measuring compact films of several atomic layers; it can even be used for measuring sub-monolayers consisting of a two-dimensional arrangement of atoms or molecules with only some of the possible sites occupied. The method can therefore be used for measurements of adsorption processes and also for determinations of film growth due to chemical reactions between solids and electrolytes or gases. Several forms of both techniques and a number of applications are discussed.

The frosting of glass

The purpose of frosting the bulbs of incandescent lamps is to increase the light-emitting surface and thus to reduce the luminance. Frosting takes place in two operations.

In the first operation a "frosting liquid" is used which consists of a saturated solution of $NH_4F.HF$ in 30% HF with more $NH_4F.HF$ suspended in it. This liquid is sprayed on to the glass bulb for a few seconds at a temperature of 40 °C. In the reaction of the liquid with the glass (an etching process), one of the reaction products is $(NH_4)_2SiF_6$ (ammonium silicofluoride), which is not readily soluble in this environment and crystallizes out on the glass. At a place which is covered by a crystal the glass wall is more or less protected; the etching of the surrounding glass continues and fresh $(NH_4)_2SiF_6$ is formed, causing further growth of the existing crystal. After a few seconds the growing crystals touch each other, but the join is so poor that the etching continues along the lines where they touch, and deep, sharp grooves are formed. A simplified diagram of a cross-section of the bulb after the first operation is given in *fig. 1*. The cross-hatched area represents the deposited layer of $(NH_4)_2SiF_6$.

Under mechanical stress, the bulb would easily crack at these deep grooves. The breaking strength of the bulb in this state is therefore very poor. This is reme-

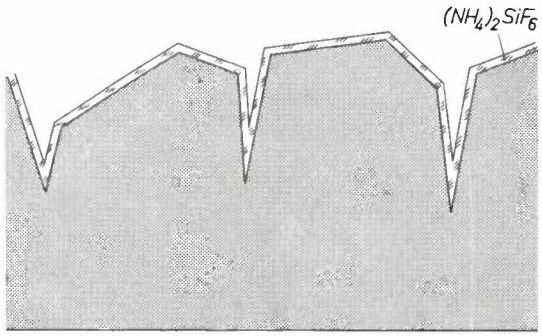


Fig. 1

reaches its starting point and proceeds from there in approximately the same way. When the thickness had reached 320 nm, the measurements were stopped since the further growth of the film was too inhomogeneous.

Apart from the kinetic data which may be obtained from this curve the fact that the measured curve is closed leads to the conclusion that the film is non-absorbing. But the value of the refractive index determined by this measurement does not allow a determination of the chemical composition of the film since this constant is not specific enough. This value is however of some use in comparative studies, as may be seen from a comparison of fig. 16 and *fig. 17*. This plot of Δ against ψ, as measured during anodic polarization in an aqueous solution of caesium hydroxide (solid line), does not form a closed curve but it is similar to those calculated with complex refractive indices of the film (heavy dashed lines). In these calculated curves the complex index of refraction has been used: $\mathbf{n}_1 = n_1 - jk_1$, where n_1 is the real part and k_1 the absorption coefficient. From the different results shown in figs. 16 and 17 the conclusion can be drawn that to a certain extent the alkali atoms present in the electrolyte determine the chemical character of the films produced. It could be shown by measurements of the infra-red absorption by IRS that the films consisted of acid alkali germanate, whose refractive index in the visible spectral region depends on the kind of alkali present in this compound.

Summary. Measurements of infra-red absorption by internal reflection spectroscopy (IRS) permit the determination of the chemical composition of thin films down to thicknesses of less than 1 nanometre. Electronic processes at the surface of solids such as changes of the carrier density in the space-charge region of semiconductors can also be measured by this method. The ellipsometric determination of film thicknesses is a very sensitive method which can readily be used for measuring compact films of several atomic layers; it can even be used for measuring sub-monolayers consisting of a two-dimensional arrangement of atoms or molecules with only some of the possible sites occupied. The method can therefore be used for measurements of adsorption processes and also for determinations of film growth due to chemical reactions between solids and electrolytes or gases. Several forms of both techniques and a number of applications are discussed.

The frosting of glass

The purpose of frosting the bulbs of incandescent lamps is to increase the light-emitting surface and thus to reduce the luminance. Frosting takes place in two operations.

In the first operation a "frosting liquid" is used which consists of a saturated solution of $NH_4F.HF$ in 30% HF with more $NH_4F.HF$ suspended in it. This liquid is sprayed on to the glass bulb for a few seconds at a temperature of 40 °C. In the reaction of the liquid with the glass (an etching process), one of the reaction products is $(NH_4)_2SiF_6$ (ammonium silicofluoride), which is not readily soluble in this environment and crystallizes out on the glass. At a place which is covered by a crystal the glass wall is more or less protected; the etching of the surrounding glass continues and fresh $(NH_4)_2SiF_6$ is formed, causing further growth of the existing crystal. After a few seconds the growing crystals touch each other, but the join is so poor that the etching continues along the lines where they touch, and deep, sharp grooves are formed. A simplified diagram of a cross-section of the bulb after the first operation is given in *fig. 1*. The cross-hatched area represents the deposited layer of $(NH_4)_2SiF_6$.

Under mechanical stress, the bulb would easily crack at these deep grooves. The breaking strength of the bulb in this state is therefore very poor. This is reme-

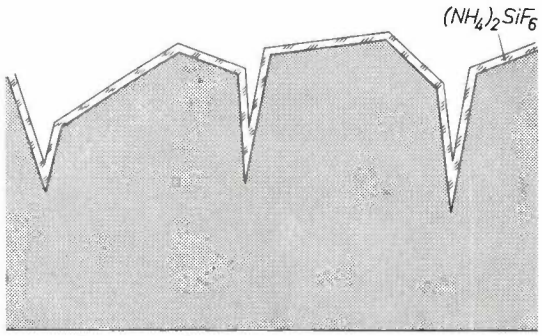


Fig. 1

died by the second operation, which rounds off the grooves. In this second operation the bulb wall is washed with 10-15% HF at 80 °C. This dissolves the layer of $(NH_4)_2SiF_6$ and attacks the glass. Here again, the extent to which the glass is attacked depends upon the screening provided by the silicofluoride crystals. The second process starts in the grooves, where the reaction lasted longest during the first operation, and spreads out on a more or less spherical front until, after a few seconds, all the crystals have been dissolved. This is illustrated diagrammatically in *fig. 2*. The lines numbered *1* to *4* denote the boundaries of the glass surface at several stages of the treatment with HF. The grooves become rounded, giving the glass much greater mechanical strength.

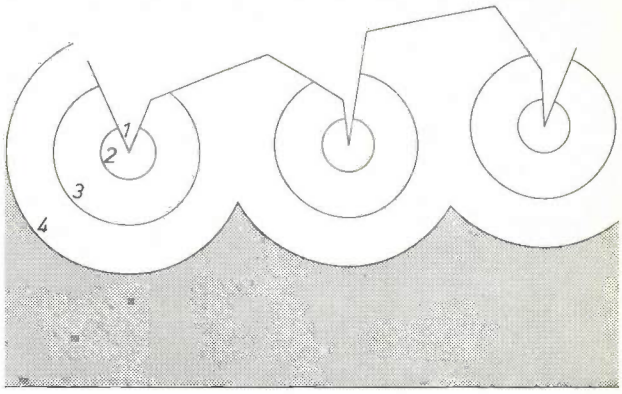


Fig. 2

It seemed interesting to follow the course of the frosting process in order to be able to investigate the cause of irregularities sometimes encountered in frosting. The course of the process can be analysed by interrupting the process at certain moments and taking photographs of the glass surface with the electron microscope. This makes it necessary, however, to delay the processes, which normally take place very rapidly.

When an attempt is made to slow down the *first* operation by modifying concentrations and temperatures, the structure of the surface becomes quite different from the structure obtained after the normal process. Because of this, it was not possible to record intermediate stages in this process by interrupting the operation. All that was obtained was the final phase, a photograph of which is shown in *fig. 3*. (For this photograph the ammonium silicofluoride crystals were removed, since we were primarily interested in the state of the glass surface.) The sharp grooves corresponding to line *1* in fig. 2 can clearly be seen.

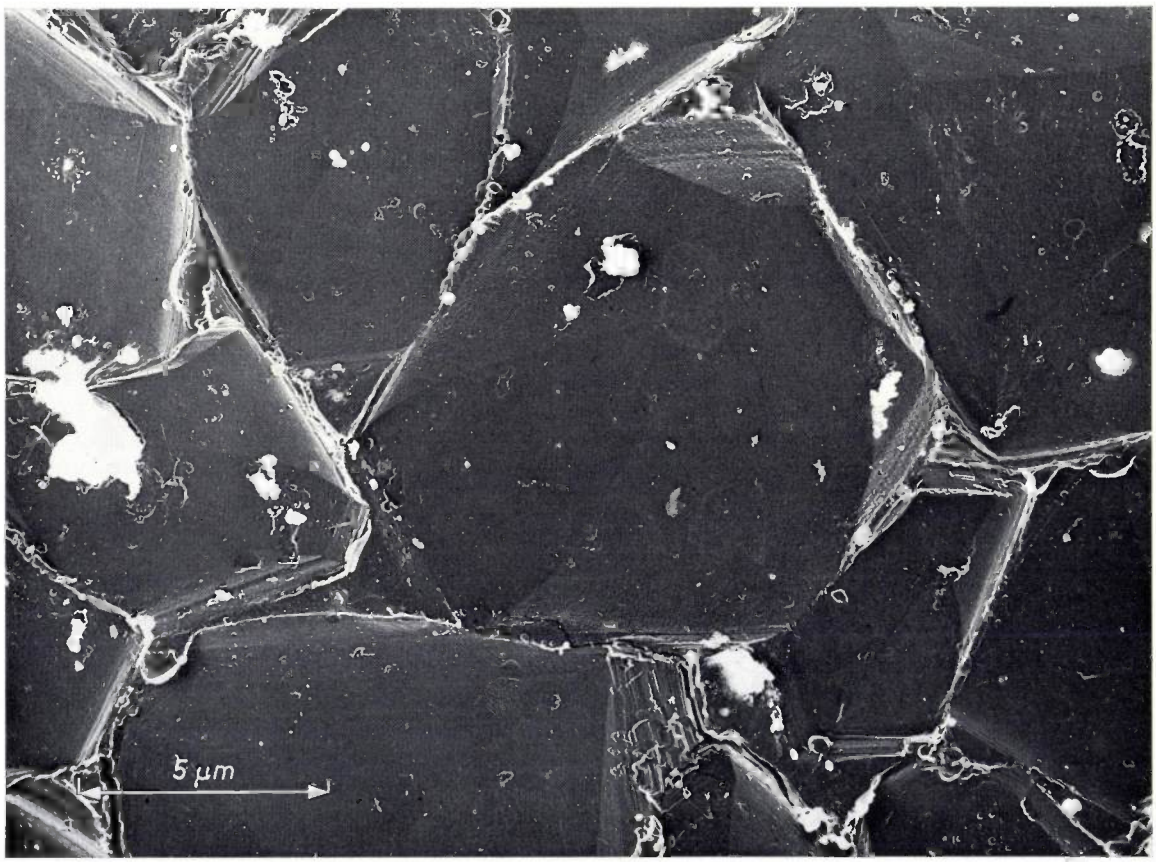


Fig. 3

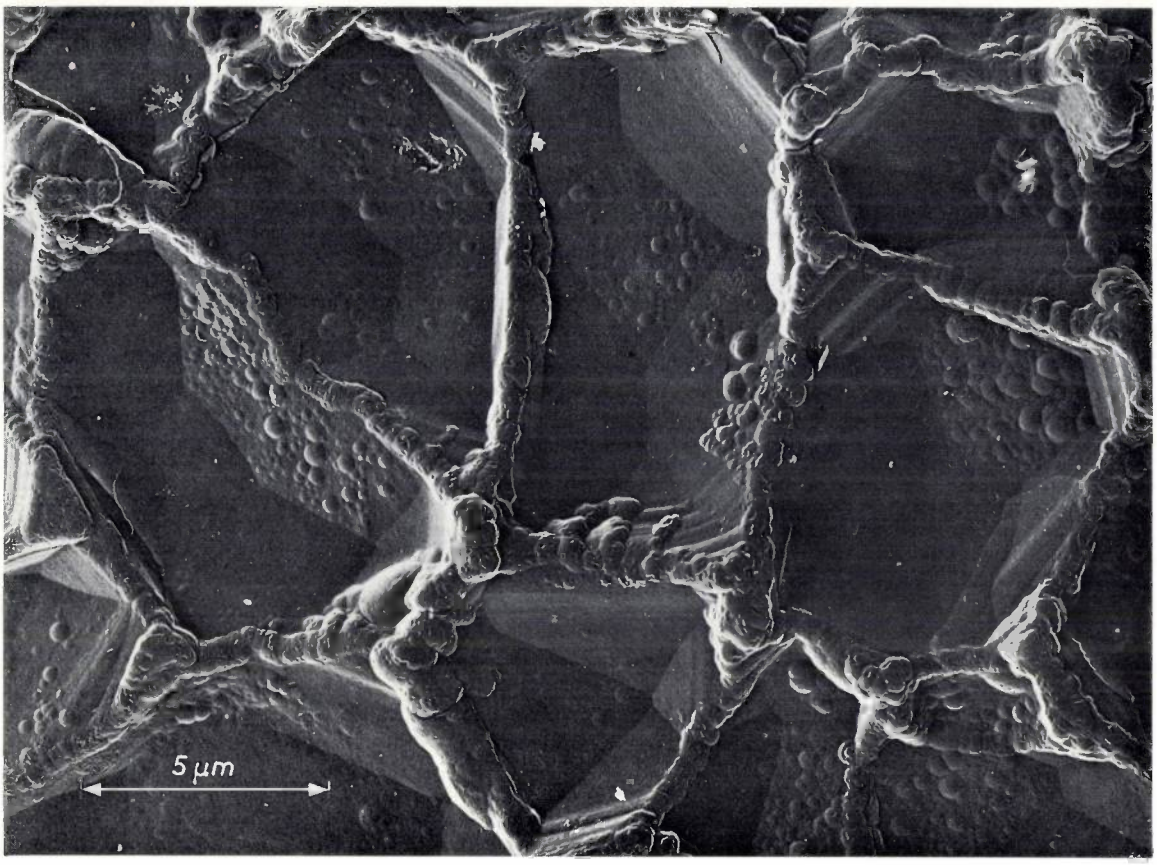


Fig. 4*a*

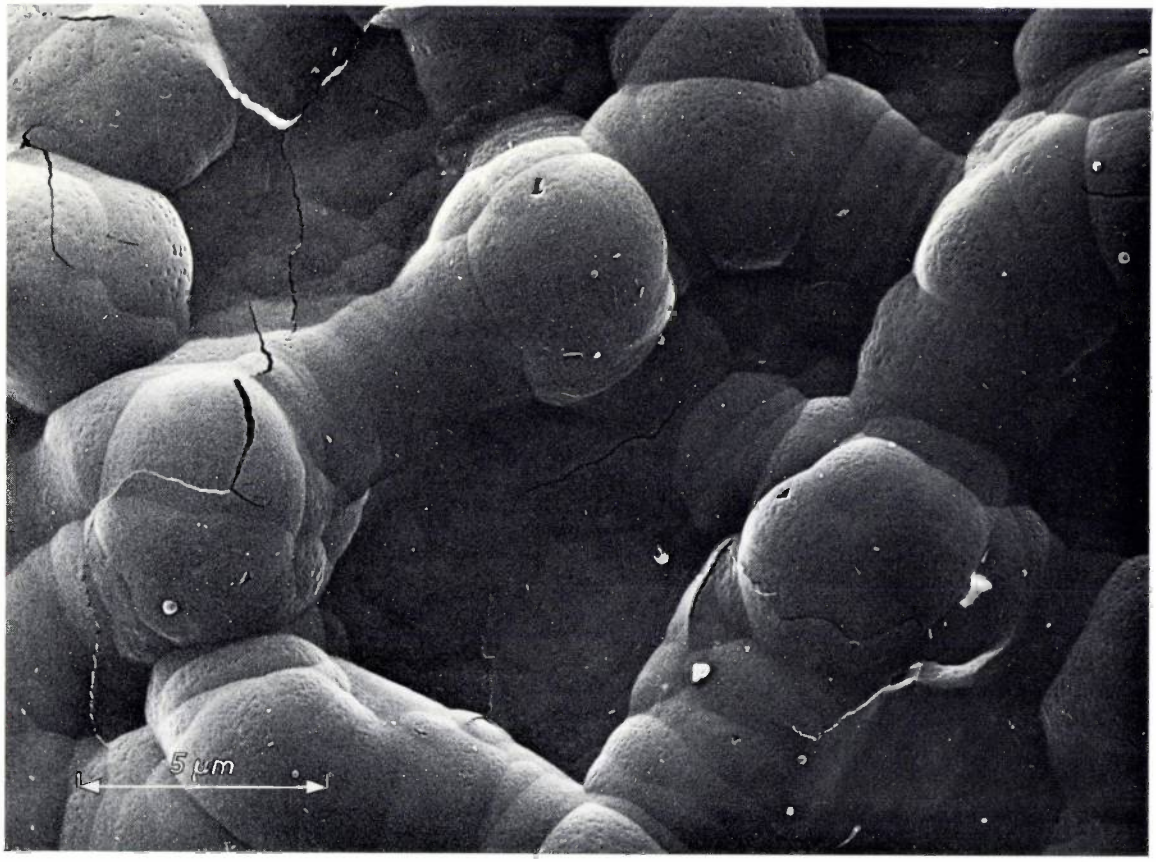


Fig. 4*b*

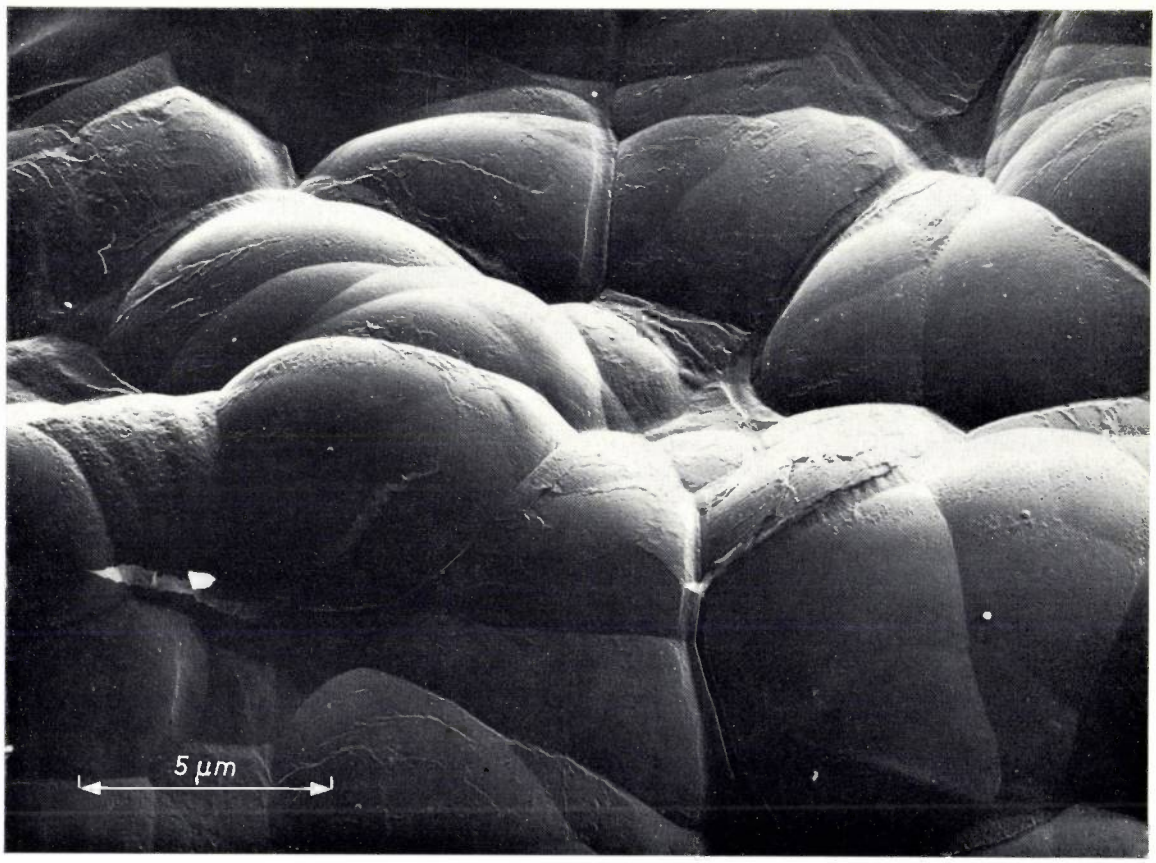


Fig. 4*c*

The *second* operation can be slowed down to take about ten minutes, which enables the process to be broken up into various stages. Several of the stages are shown in *fig. 4*. For fig. 4*a* the treatment was interrupted after half a minute; the surface corresponds to line *2* of fig. 2. The grooves are already beginning to be rounded off. After three minutes' treatment the picture in fig. 4*b* is obtained (line *3* in fig. 2). Fig. 4*c* is a photomicrograph of the frosted glass at the end of the operation. The rounding off is complete and the glass wall now consists of a continuous pattern of tiny, concave "lenses" which scatter the light uniformly on all sides — which is the desired effect of frosting.

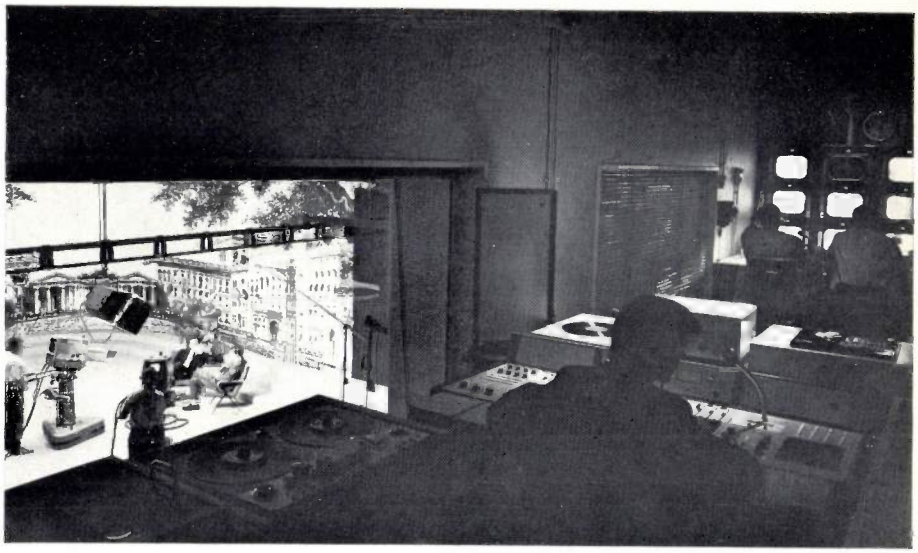
The replicas were made from a piece of the bulb that had been washed in running water and then dried. A piece of "Technovit" [*], a thermoplastic material, was pressed on to the glass, and the glass and plastic were heated. After cooling, an impression was thus obtained of the relief surface of the glass. Next, to improve the contrast, a layer of platinum was deposited on the plastic by vacuum-evaporation at a small angle, and a carbon layer was then deposited. After dissolving the plastic, small pieces of the carbon layer with the platinum adhering to them were transferred to specimen grids.

The electron photomicrographs were made by J. Th. Gruijters with a Philips EM 100 electron microscope.

H. B. Haanstra
S. C. Rademaker

[*] Registered trademark of Kulzer & Co. GmbH, Bad Homburg, West Germany.

H. B. Haanstra is with Philips Research Laboratories, and Ir. S. C. Rademaker with the Philips Glass Development Centre, Eindhoven.



Colour television studio

Developments in the field of colour television at the Philips Research Laboratories in Eindhoven made it necessary to have facilities for operating the equipment under practical conditions. In 1964 a television studio was therefore built adjoining the laboratory buildings and fully equipped on up-to-date lines. A large number of colour television programmes of widely different types have been produced here during the last three years. The colour photograph shows a programme rehearsal in progress.

Most of these experimental programmes were produced by the Dutch Television Corporation and the Broadcasting Associations (and in a later stage by several German broadcasting organizations as well), who thus turned to good account the opportunities provided here for gaining experience in colour work.

The studio has extensive lighting facilities, including a number of movable overhead spotlights that can be controlled from a central control desk. The equipment includes a number of colour cameras, 16 mm and 35 mm film scanners for colour-film transmissions, etc. The photograph above shows the producer's control room with audio and video desks. The control-room window gives the producer a good view of the action in the studio. Other rooms include the room which houses the electronic equipment, dressing rooms, property stores, scenery workshop and canteen.

At first the experimental transmissions were made with the NTSC system. The later transmissions have been made with the PAL system which has since been adopted in the Netherlands and a number of other European countries for their regular transmissions. The pictures were transmitted by Philips' own transmitter. Colour television receivers were installed in many private homes in and around Eindhoven for evaluating the results.

Motional feedback with loudspeakers

J. A. Klaassen and S. H. de Koning

The use of negative feedback to reduce linear and non-linear distortion has long been standard practice in amplifier techniques. In audio equipment, however, the loudspeaker is an important source of distortion, and the distortion which it contributes remains unaffected by negative feedback in the electronic circuit. It is possible, however, to include the loudspeaker in the negative feedback loop. One method of doing this, which makes use of an acceleration transducer, has been investigated theoretically and experimentally by the authors.

Problems with small loudspeaker cabinets

Distortion in the electronic circuits of an audio system — microphone, amplifier, loudspeaker — has traditionally been dealt with by means of negative feedback. However, this feedback has no effect on the distortion produced outside the electronic circuits. This occurs mainly in the loudspeaker, in the conversion of an electrical current into a mechanical force and of a mechanical force into the movement of a diaphragm and the surrounding air. This is true both for distortion of the distribution of intensity over the frequency spectrum (linear distortion) and for the production of higher harmonics (non-linear distortion).

The diaphragm in an electrodynamic loudspeaker usually takes the form of a cone in a spring suspension. Mounted near the apex of the cone there is a cylindrical coil (the "speech coil"), which can move in a radial magnetic field. A varying current through the coil generates a varying force which sets the coil and cone in motion. In this process linear distortion may occur since the spring-mass system has its own resonant frequency. Non-linear distortion can occur because the resilience of the cone suspension system does not vary completely linearly with the displacement of the cone. Non-uniformity of the magnetic field can also cause non-linear distortion.

The air near the cone opposes the motion and thus constitutes a load on the cone. This load has a reactive part, the *radiation reactance* X_r, and a resistive part, *the radiation resistance* R_r. The latter is a measure of the power P which is radiated in the form of sound at a particular cone vibration velocity $\dot{x}$ ($P = \dot{x}^2 R_r$), where R_r depends on the dimensions of the cone and on the frequency (see *fig. 1*). If the air is prevented from moving from the front to the back of the cone, or vice versa, R_r is proportional at low frequencies to the square of the frequency. At high frequencies, however, R_r is virtually constant. This means that much greater velocities, and therefore much greater displacements of the cone, are needed to radiate a given power at low frequencies than at high frequencies. At low frequencies, therefore, the risk of non-linear distortion is greater.

Although, with many kinds of sound, the energy contained in the lower-frequency components is smaller than the energy in the middle range, there are sounds in which a considerable contribution is made at the lowest frequencies. To reproduce such sounds well the equipment must be able to deliver as much sound power at the low frequencies as at the high ones.

At low frequencies the radiated power decreases sharply if sound waves from the rear of the cone are able to interfere with waves radiated from the front. In practice an effective separation can only be achieved by mounting the loudspeaker in the wall of a cabinet which is otherwise completely closed. If, for the sake of compactness, we make the cabinet small, the enclosed air behaves like a spring, which has the effect of reducing the displacement of the cone — again to the detriment of the radiation at low frequencies — and increasing the resonant frequency. The designer generally attempts to keep the resonant frequency so low that it lies at the lower limit of the frequency range, but if a small loudspeaker cabinet is used the resonant frequency may well be within this range, giving severe linear distortion. Loudspeaker resonance is also a great nuisance in the reproduction of "step-function" sounds or sudden

Ir. J. A. Klaassen is with Philips Research Laboratories, Eindhoven; Ir. S. H. de Koning, formerly with Philips Research Laboratories, is now with the Philips Electro-Acoustics Division (ELA), Eindhoven.

transients. Furthermore, the resilience of the cushion of air behind the cone is not proportional to the cone displacement, and this again introduces non-linear distortion.

In most practical situations — even in high-fidelity equipment — linear and non-linear distortion can be kept within bounds by conventional means. It is interesting to note, however, that if the very best sound reproduction is required or if there is not much room for the loudspeaker, the sound reproduction can

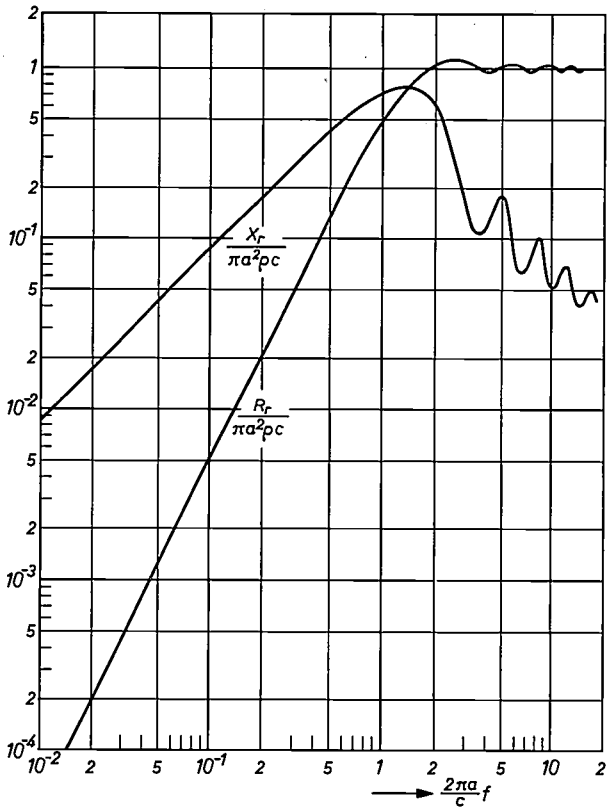


Fig. 1. The radiation reactance X_r and the radiation resistance R_r of a loudspeaker as a function of frequency f (on relative scales). a radius of the cone. ϱ air density. c velocity of sound in air. At high frequencies the radiation resistance per unit area approaches ϱc, the specific acoustic impedance of air. $X_r/2\pi f$ is known as the radiation mass.

be improved by extending the negative feedback we spoke of earlier to include the loudspeaker, and not limiting it to the electronic circuits.

This means that the voltage fed back to the input of the amplifier is related to the radiated sound rather than to the current or voltage supplied to the loudspeaker. This form of feedback, known as *motional feedback*, is the subject of this article. After looking at the theory we shall describe two systems in which use is made of motional feedback. Our measurements with these systems have shown that motional feedback does in fact give the improvement expected.

Motional feedback

History

Motional feedback requires the generation of a signal that is related to the radiated sound. This suggests in the first place that a microphone should be used. This would hardly be practical, however, because the transit time of the sound between loudspeaker and microphone introduces a considerable phase shift which is highly frequency-dependent.

At low frequencies (below about 500 Hz) a voltage related to the radiated sound can be derived from the movements of the coil. This is possible since the cone vibrates as a rigid body in this frequency range. (The surface of the cone is in "isophase motion". This is not the case at high frequencies.)

The movement of the speech coil can be converted into a signal voltage in various ways. A voltage may be derived from the *displacement* of the coil (e.g. by a capacitive method), from its *velocity* (e.g. by an inductive method), or from its *acceleration* (using inertial forces). For brevity, we shall use the terms displacement, velocity and acceleration feedback in this article. Two of these methods are significantly different in principle from the third: a stationary reference point is always needed for determining the position and velocity of a moving point, but not for determining an acceleration. We shall return to this point later.

A motional feedback system sometimes used today employs *velocity feedback*, the voltage used as a measure of the velocity of the speech coil being the e.m.f. induced in the coil by its motion in the magnetic field [1]. This voltage has of course to be separated from the voltage fed from the amplifier to the coil. Several methods are illustrated in *fig. 2*. In series with the loudspeaker in all these circuit diagrams there is an impedance Z_e whose phase angle is equal to that of the electrical impedance of the loudspeaker when the movement of the coil is prevented. Since the loudspeaker is not blocked in this way in reality, its impedance is dependent on the signal frequency. The voltage across Z_e can be subtracted from the loudspeaker voltage, for instance by means of an extra transformer (fig. 2*a*), an extra winding on the output transformer in the amplifier (fig. 2*b*), or with the aid of a bridge circuit (fig. 2*c*). The difference between these voltages is then fed back to the amplifier input.

A feedback method of this kind can also be regarded as a method of making the output impedance of the amplifier equal and opposite to the impedance of the loudspeaker in the blocked state. The amplifier is then

[1] See W. Holle, Gegenkopplung an Lautsprechern, Funk-Technik 7, 490-492, 1952.

said to have a *negative output impedance* [2]. In such an arrangement the amplifier functions as a voltage source which is loaded purely and simply by that part of the loudspeaker impedance which is due to the motion of the speech coil (motional impedance).

A difficulty in the methods referred to above is that the resistance of the speech coil is temperature-dependent. Another drawback is the non-uniformity of the magnetic field, as a result of which the voltage obtained

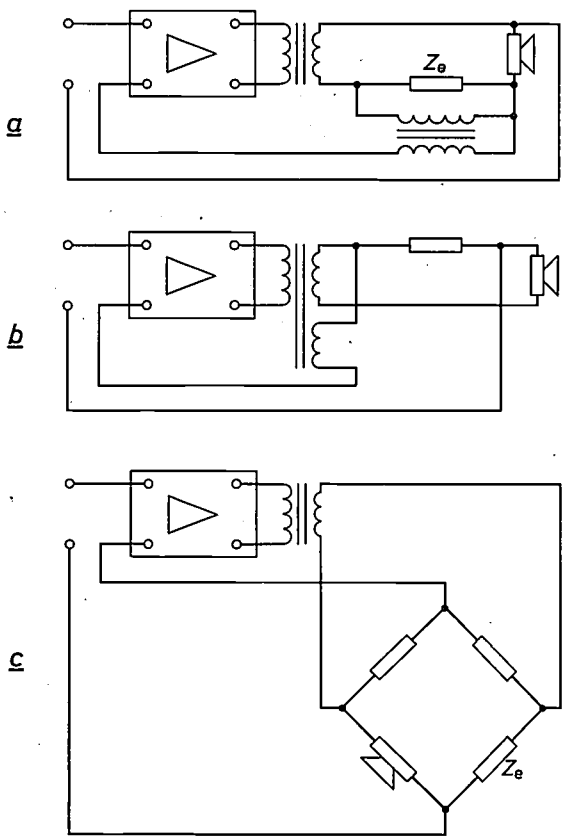


Fig. 2. Some methods of motional feedback, using the voltage induced in the speech coil itself by its movement in the magnetic field. Connected in series with the loudspeaker is an impedance Z_e, whose phase angle is equal to that of the impedance of the loudspeaker in the blocked state. The difference between the voltages across the loudspeaker and across Z_e are obtained by means of a separate transformer (*a*), with an additional winding on the output transformer (*b*), or with a bridge circuit (*c*). This difference voltage is added in the correct phase to the input signal.

does not give a true picture of the velocity of the coil. For these reasons circuits of this kind are not often used nowadays.

In another method an extra coil is wound on to the former of the speech coil and the voltage induced in it is used as a measure of the velocity of the speech coil. This extra coil must be arranged in such a way that the mutual inductance between it and the speech coil is very low. It is also most important that there should be an absolutely stationary reference point. This is usually not the case because of vibrations from the loudspeaker itself and because of the radiated sound, which can excite mechanical resonances.

Displacement feedback by means of capacitance variations also has the disadvantage that a stationary reference point is required. Moreover, in this case non-axial movements of the speech coil also affect the voltage obtained.

The methods given above can in fact be used if the aim is to *measure* the movements of the speech coil, and account is taken of the errors introduced. For negative feedback, however, the signals obtained in this way have not been found suitable, because of the drawbacks mentioned earlier.

The various difficulties can be avoided if the feedback is effected by means of a device which measures the *acceleration* of the speech coil. The system described in this article makes use of this method. As the introduction of motional feedback at higher frequencies involves considerable problems, due to the difficulty of keeping the system stable, we have limited the frequency range so that the feedback operates only at low frequencies. Since, as we have shown, the distortion is most severe at low frequencies, this in itself gives a substantial improvement.

The voltage obtained from an acceleration transducer can be integrated once or twice, thus giving, without any of the drawbacks described above, a voltage which is proportional to the velocity or the displacement. A single transducer can therefore be used to give acceleration, velocity or displacement feedback, or any combinations of these.

Theoretical considerations

At low frequencies the coil and cone of a loudspeaker can be treated as a single spring-mass system. Its resonant frequency will be referred to briefly as the resonant frequency of the loudspeaker. The motion of this spring-mass system is described in the following equation:

$$m\ddot{x} + r\dot{x} + sx = F, \quad \ldots \quad (1)$$

where F is the force acting on the coil, x the displace-

[2] See R. L. Tanner, Improving loudspeaker response with motional feedback, Electronics **24**, No. 3, 142 etc., 1951; R. E. Werner, Loudspeakers and negative impedances, IRE Trans. AU-6, 83-89, 1958; H. W. Holdaway, Design of velocity-feedback transducer systems for stable low-frequency behavior, IEEE Trans. AU-11, 155-173, 1963.

[3] The resonant frequency is understood here to be the frequency at which the mechanical impedance of the loudspeaker is resistive.

[4] The amplitude characteristic is the curve representing the amplitude of the sound pressure, measured along the axis of the loudspeaker in a free sound field, as a function of frequency, for constant voltage across the loudspeaker terminals.

ment from the position of equilibrium, m the mass, r the frictional resistance and s the spring constant. The resonant frequency of such a system [3] is given by:

$$f_0 = \frac{\omega_0}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{s}{m}}, \quad \ldots \quad (2)$$

and the Q (quality factor) by:

$$Q_0 = \omega_0 \frac{m}{r} = \frac{\sqrt{ms}}{r}. \quad \ldots \quad (3)$$

When acceleration feedback is applied the current through the speech coil, and hence the force F proportional to it, are reduced by a term proportional to $\ddot{x}$. If the proportionality factor is indicated by m', the equation of motion can then be written:

$$m\ddot{x} + r\dot{x} + sx = F - m'\ddot{x}$$

or:

$$(m + m')\ddot{x} + r\dot{x} + sx = F. \quad \ldots \quad (4)$$

The *mass* is thus effectively increased. As a result of this the resonant frequency is lower and the Q higher. Similarly it can be demonstrated that velocity feedback effectively increases the *damping term*; the resonant frequency remains unchanged, but the Q is reduced. Displacement feedback effectively increases the *spring constant* s, thereby increasing the resonant frequency and also the Q.

If the three feedback systems are combined we can choose the resonant frequency of the loudspeaker and the Q independently of one another, so that we can control the shape of the amplitude characteristic [4] at low frequencies.

The resonant frequency and the Q of a system using motional feedback can be calculated with the aid of an equivalent circuit. *Fig. 3* illustrates this with the diagram of an output amplifier represented by a voltage source of e.m.f. ge (g voltage gain, e input voltage) and an internal impedance R_i. The loudspeaker is represented by an electrical analogue in the form of a parallel circuit, in which the capacitance is the analogue of m, the inductance is the analogue of $1/s$ and the resistance the analogue of $1/r$ (see equation 1). The radiation mass is assumed to be included in m and the radiation resistance in r. The force F acting on the speech coil is now the analogue of the current supplied to the circuit. Since this force is equal to Bli (B magnetic flux density, l length of wire, i current), the transition from electrical to mechanical quantities can be indicated in the circuit by an ideal transformer of turns ratio Bl:1. The output circuit of the amplifier also includes the resistance R_s of the speech coil. (Since this equivalent circuit is only used at low frequencies, we can neglect the inductance of the speech coil.)

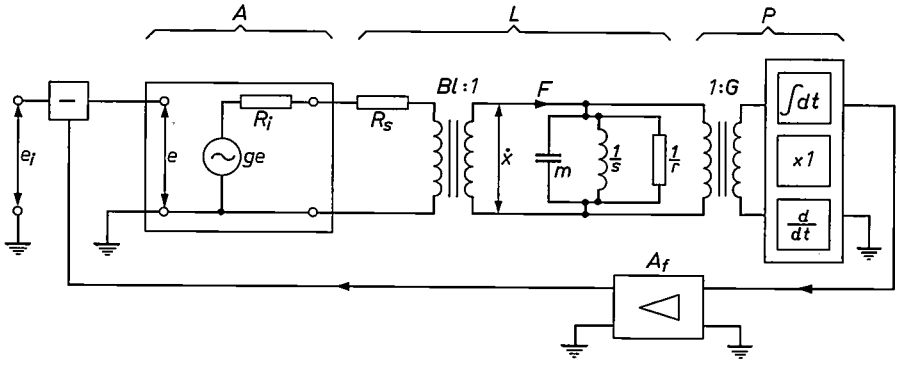


Fig. 3. Equivalent circuit diagram of a system consisting of an output amplifier A with loudspeaker, using motional feedback. The loudspeaker L is replaced by an electrical analogue in the form of a parallel circuit connected by means of two ideal transformers to the amplifier and the transducer P. The voltage across the circuit is now the analogue of the velocity $\dot{x}$ of the speech coil and cone. Depending on its type the transducer delivers a signal proportional either to $\int \dot{x}dt = x$, to $\dot{x}$ or to $d\dot{x}/dt = \ddot{x}$.

The voltage appearing across the circuit is now seen to be the analogue of the velocity $\dot{x}$ of the speech coil. The equivalent circuit of the transducer contains another ideal transformer of turns ratio 1 : G (G being a measure of the sensitivity). The transducer delivers a signal which is proportional either to $\int \dot{x}dt = x$, to $\dot{x}$ or to $d\dot{x}/dt = \ddot{x}$, depending on whether it is a displacement, velocity or acceleration transducer. (Accordingly, the turns ratio G takes different dimensions.) The input impedance of the transducer is considered to be infinite, and its mechanical characteristics are assumed to be included in the quantities m, s and r of the loudspeaker. The output signal from the transducer is amplified A_f times and fed back to the amplifier input.

With the aid of fig. 3 we can examine the effect of the feedback on the resonant frequency and the Q of the system. If we represent the resonant frequency and the Q without feedback by $\omega_0/2\pi$ and Q_0 respectively, we arrive at the results summarized in *Table I*. This table gives a quantitative representation of the effect of motional feedback on the resonant frequency and the Q.

Table I. Resonant frequency and Q of a loudspeaker with motional feedback.

	Acceleration feedback	Velocity feedback	Displacement feedback
Resonant frequency	$\frac{\omega_0/2\pi}{\sqrt{1 + \frac{gA_fGBl}{m(R_i + R_s)}}}$	$\frac{\omega_0}{2\pi}$	$\frac{\omega_0}{2\pi}\sqrt{1 + \frac{gA_fGBl}{s(R_i + R_s)}}$
Q	$Q_0\sqrt{1 + \frac{gA_fGBl}{m(R_i + R_s)}}$	$\frac{Q_0}{1 + \frac{gA_fGBl}{B^2l^2 + r(R_i + R_s)}}$	$Q_0\sqrt{1 + \frac{gA_fGBl}{s(R_i + R_s)}}$

Choice of type of feedback

Since the sound pressure generated at low frequencies is proportional to the acceleration of the cone, a flat amplitude characteristic in this range can be obtained if the *acceleration* of the cone is made independent of the frequency. This can be done by means of acceleration feedback. A drawback of this method, however, is that the resonance becomes more noticeable. Whilst the feedback lowers the resonant frequency, the Q increases, so that there is a peak in the amplitude characteristic. This situation can be improved by the simultaneous application of velocity feedback, which, as we have seen, reduces the Q.

An attractive feature of acceleration feedback is that the feedback factor increases with rising frequency (since, with a constant displacement amplitude, the acceleration increases with frequency). Because of this the higher harmonics generated by non-linear distortion are more strongly suppressed than with displacement or velocity feedback.

The transducer

As was stated earlier, we have used an acceleration transducer. This consisted primarily of a disc of piezoelectric material, fixed on one side to the coil former of the loudspeaker. When this disc is set in vibration, inertial forces are set up in it which produce a voltage between the end faces of the disc, and this voltage is a measure of the acceleration of the coil former. The mass must be kept down to avoid affecting the reproduction at higher frequencies.

For a given mass the highest signal power is obtained with a piezoelectric element whose longest dimension lies in the direction of acceleration. The impedance may then be so high, however, that an amplifier connected to the element would be upset by leakage currents. The arrangement would also be very susceptible to interfering voltages. Although an optimum configuration can be found with a stack of thin piezoelectric elements connected in parallel and stacked in the direction of acceleration, simplicity dictated the use of one thin element, of not too low a capacitance, with a loading mass on the side turned away from the coil. In the examples discussed below the capacitance was approximately 600 and 3000 pF.

In the first experiments the piezoelectric element was attached to the edge of the coil former with adhesive. It was then found that non-axial forces also occurred in the coil, giving rise to unwanted output signals. The mounting of the transducer was later changed so as to prevent those forces from affecting the piezoelectric element.

Another possible source of unwanted voltages in the piezoelectric element is the variation in air pressure to which it is subjected during its movements. This can be remedied by enclosing the element in a completely sealed housing. If this is made of metal it can provide electrical screening at the same time.

The efficiency of a loudspeaker

Uniform reproduction from a normal loudspeaker can only be obtained in a frequency range that lies *above* the resonant frequency. Below this frequency the efficiency quickly deteriorates. This is illustrated by curve a in *fig. 4*, representing the acoustic power radiated at constant input voltage. This curve relates to a loudspeaker with a Q of 1. We see that at a frequency equal to half the resonant frequency f_0 the radiated power has dropped by 12 dB (a factor of 16) compared with that at f_0. To obtain at the frequency $\frac{1}{2}f_0$ an acoustic power that is equal to that at frequencies higher than f_0, the amplifier must thus be capable of delivering an electrical power 16 times greater than is needed at higher frequencies.

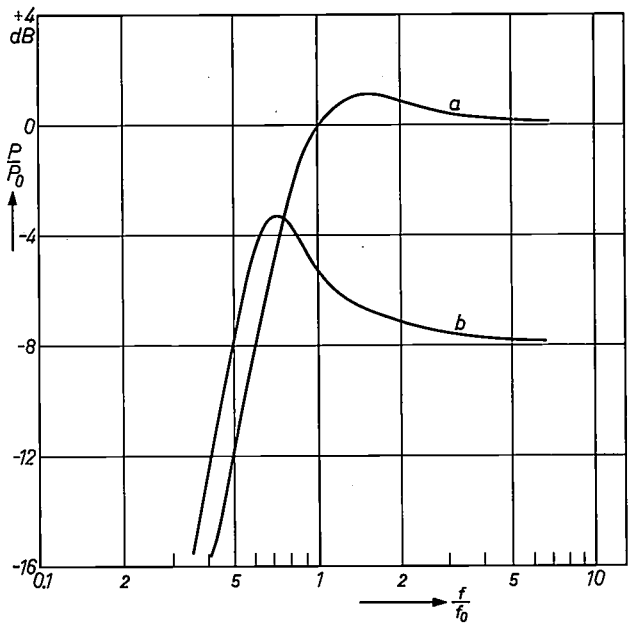


Fig. 4. The power P radiated by a loudspeaker, at constant input voltage, as a function of frequency f, both on relative scales. P_0 power at the resonant frequency f_0. The curve a is for $Q = 1$. In this case the power at the frequency $\frac{1}{2}f_0$ is 12 dB lower than that at f_0 and above. Increasing the mass of cone and speech coil by a factor of 2.5 (curve b) makes the power at $\frac{1}{2}f_0$ equal to that at higher frequencies. The Q is then $1/2.5 = 1.58$.

A more uniform distribution of the required power is possible if the moving mass of the loudspeaker is increased. Although this reduces the radiated power for frequencies above f_0, it increases the power for lower

frequencies. In fig. 4 curve *b* shows the acoustic power for the same loudspeaker as for curve *a*, but with the moving mass now increased by a factor of 2.5. The Q has now been increased by a factor of $\sqrt{2.5} = 1.58$, and the acoustic power at higher frequencies has dropped by 8 dB, while the power at $f = \frac{1}{2}f_0$ has increased by 4 dB, so that these two powers are now at the same level. Thus, if this loudspeaker is to be used down to a frequency of $\frac{1}{2}f_0$, the electrical power needs to be increased not by 12 dB but only by 8 dB. The resonant frequency has now been shifted to $f_0/\sqrt{2.5} = 0.63f_0$. It should be noted that this method of obtaining a more uniform efficiency can only be used for low-frequency loudspeakers. Increasing the moving mass in full-range loudspeakers generally gives poorer reproduction at the higher frequencies.

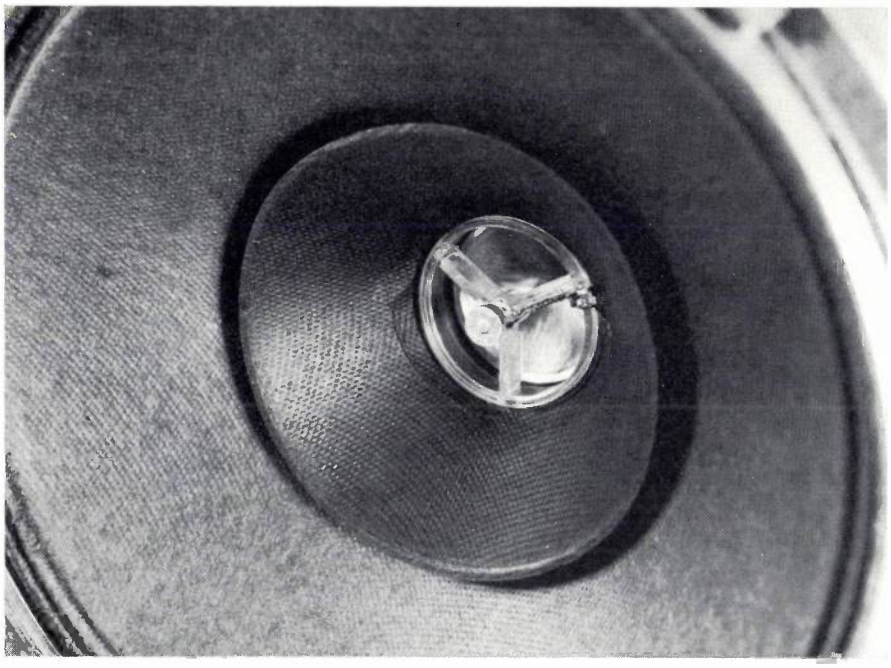


Fig. 5

Fig. 6

Fig. 5. Method of attaching the acceleration transducer to the full-range loudspeaker, type 9710 AM.

Fig. 6. Method of attaching the acceleration transducer to the bass loudspeaker, type 9710 A.

Two loudspeakers for motional feedback

We have applied motional feedback on the principle described above in experimental circuits using two loudspeakers. One of them was a bass loudspeaker (type 9710 A), the other a full-range loudspeaker (type 9710 AM) [5]. Both loudspeakers were mounted in a closed cabinet whose internal volume was 10 litres. The part of the cone inside the coil was cut away to simplify the attachment of the acceleration transducer. This part of the cone does not affect the radiation of sound, but serves only for damping the spring-mass system. This damping is not required when motional feedback is used.

As we noted earlier, the transducers contain a disc of piezoelectric material [6], loaded with a certain mass. The transducer for the full-range loudspeaker (*fig. 5*) is made as light as possible in order to minimize its effect on the reproduction at the higher frequencies. For the bass loudspeaker (*fig. 6*) a larger piezoelectric disc and a larger loading mass were used. The effect of this is twofold: it improves the efficiency at low frequencies as demonstrated above, and increases the

[5] The difference between the two loudspeakers is that type 9710 A has a single cone while type 9710 AM has a double cone.

[6] Philips type designation PXE-5. This material was chosen in consultation with C. M. van der Burgt of this laboratory.

sensitivity of the transducer. *Table II* lists the main characteristics of the two transducers.

The total mass in this table relates to the complete transducer (piezoelectric element, loading mass and base); the output voltages relate to an acceleration of 126 m/s². This acceleration is reached when the displacement of the coil, at a frequency of 40 Hz, is so great that the coil only just stays within the magnetic field. If the loudspeaker has a flat frequency response, the acceleration is the same at all frequencies; the output voltages shown thus correspond to the displacement limit when the lowest frequency to be reproduced is 40 Hz.

Table II. Transducer data.

Loud-speaker	piezoelectric element				loading mass	total mass	output voltage
	dia.	thickness	cap.	mass			
9710 AM	5 mm	0.5 mm	620 pF	76 mg	300 mg	2 g	26 mV
9710 A	16 mm	1.1 mm	2880 pF	1.69 g	29 g	37 g	500 mV

The circuit

Fig. 7 shows a block diagram of the circuit we used for applying motional feedback to a type 9710 AM loudspeaker connected to an output amplifier *A* (output power 10 W). The acceleration transducer *P* drives a cathode follower *K*, which is placed in the loudspeaker cabinet so that the connecting leads can be short. The valve used is a nuvistor (type 7586). The cathode follower is followed by an amplifier *A'*, whose output signal is an electrical representation of the acceleration of the speech coil. *Fig. 8* shows the variation of this signal voltage in amplitude and phase with varying frequency of the (constant) input signal. The feedback loop here is open and the curve thus corresponds to the Nyquist diagram of the feedback system. It is known that an amplifier is stable only if the curve does not enclose the point —1. In the figure the position of this point is indicated for the case where the acceleration feedback has been increased until the resonant frequency — 97 Hz without

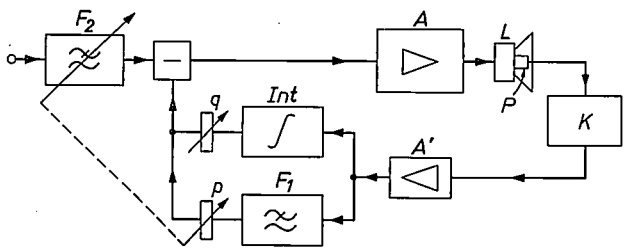


Fig. 7. Block diagram of a circuit for applying motional feedback to a full-range loudspeaker. *A* output amplifier (output power 10 W). *L* loudspeaker. *P* acceleration transducer. *K* cathode follower. *A'* amplifier. F_1 and F_2 low-pass filters. *Int* integrator. *p* and *q* variable attenuators.

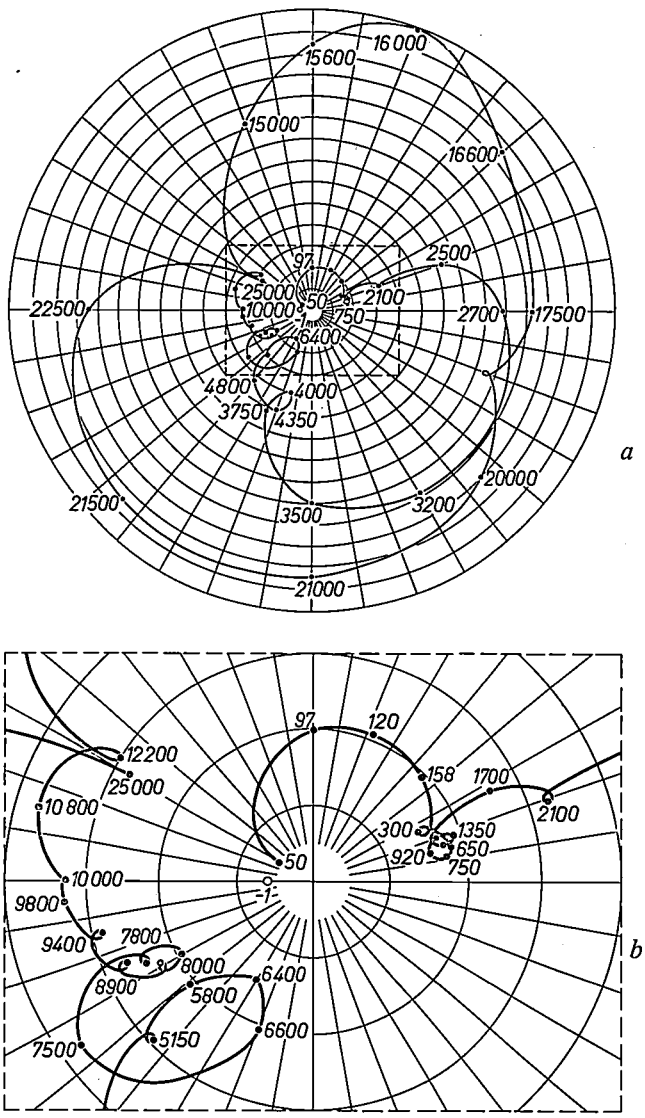


Fig. 8. *a*) Nyquist diagram for a feedback system like that of fig. 7 when the filter F_1 and the integrator *Int* are not included in the circuit. The feedback is made so large that the resonant frequency — 97 Hz without feedback — is shifted to 42 Hz. The figures along the curve indicate the frequency in Hz. Since the curve encloses the point —1, this system is not stable.
b) A magnified view of the region bounded by the dashed line in (*a*).

feedback [7] — is shifted to 42 Hz. Since, as the figure shows, the curve does enclose the point —1, the system will not be stable as it stands. Stability can be achieved by incorporating a low-pass filter F_1 in the feedback circuit, as shown in fig. 7. As we have seen, the application of motional feedback in this form is only worth while at the lower frequencies, so there is no objection to the use of a low-pass filter in the feedback circuits. The cut-off frequency of the filter is 250 Hz and the attenuation above cut-off is 12 dB per octave. The filter is followed by a variable attenuator p, used for adjusting the amount of acceleration feedback. The Nyquist diagram resulting from the introduction of

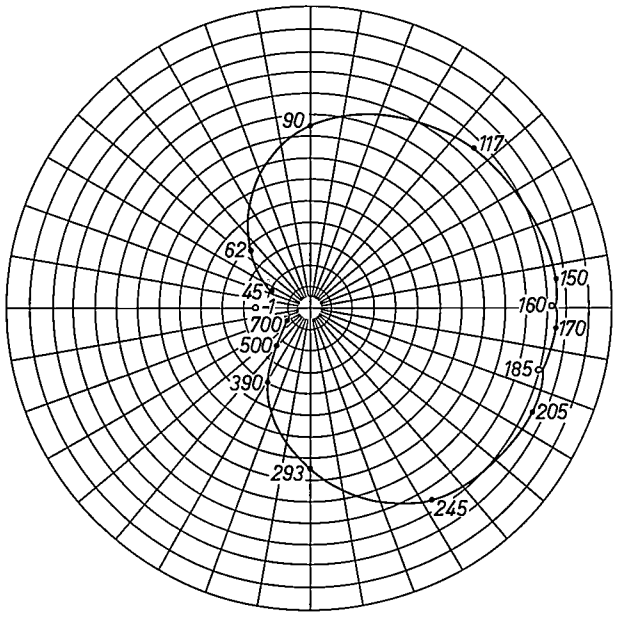


Fig. 9. The introduction of the low-pass filter F_1 makes the circuit in fig. 7 stable, as can be seen from the fact that the Nyquist diagram no longer includes the point —1.

the filter F_1 is shown in *fig. 9*. It can be seen that the circuit is now stable. The fact that there is feedback at low frequencies and not at high frequencies causes a difference in level, which is compensated for by the low-pass correction filter F_2 (fig. 7). The extent to which this filter attenuates at high frequencies can be adjusted, the adjustment being coupled with that of the attenuator p.

To prevent the resonance effect from having too great an effect on the amplitude characteristic because of the acceleration feedback (which, as we have seen, increases the Q), velocity feedback is also used. A voltage proportional to the velocity of the speech coil is derived from the acceleration signal by means of the integrator *Int*. This voltage is added to the acceleration signal via a variable attenuator q.

Fig. 10 shows the block diagram of the circuit used for applying motional feedback to the bass loudspeaker 9710 A. (The upper-frequency range is reproduced by means of a separate amplifier and loudspeaker.) Basically, the circuit is the same as that shown in fig. 7.

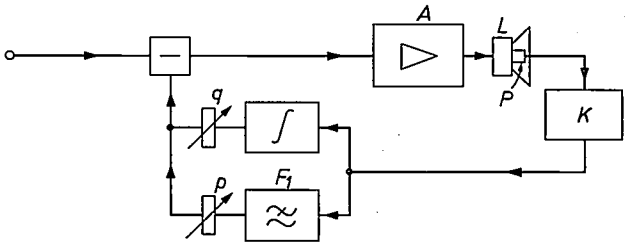


Fig. 10. Block diagram of a circuit in which motional feedback is applied to a low-frequency loudspeaker, type 9710 A. The letters have the same significance as in fig. 7.

However, since the sensitivity of the transducer used with this loudspeaker is greater, no amplifier is needed here after the cathode follower. In addition, a low-pass filter with a higher cut-off is used because, with the bass loudspeaker, the accelerations at high frequencies are lower than for the full-range loudspeaker. This frequency has therefore been fixed at 600 Hz, which means that the feedback is effective over a wider frequency range. Finally, if the motional feedback is only applied to the bass loudspeaker, there is no need for a correction filter like that shown in fig. 7 for equalizing the level at high and low frequencies.

Results of measurements

Amplitude characteristics

In *fig. 11* curve *a* is the amplitude characteristic of the full-range loudspeaker 9710 AM mounted in a closed cabinet whose internal volume is 10 litres. The resonant frequency is 97 Hz. Curve *b* shows the characteristic obtained when the maximum permissible acceleration feedback for stability was applied with the circuit of fig. 7 (attenuator q at zero). The resonant frequency is now at 42 Hz. Curve *c* shows the characteristic obtained when velocity feedback is applied as well to reduce the effect of the resonance. If we compare this with *a*, we see clearly the improvement in the reproduction of the lower frequencies obtained by the application of motional feedback.

[7] The mechanical resonance of the loudspeaker can be seen to be at this frequency since the phase angle of the acceleration is here 90°. The phase angle of the velocity is then zero.

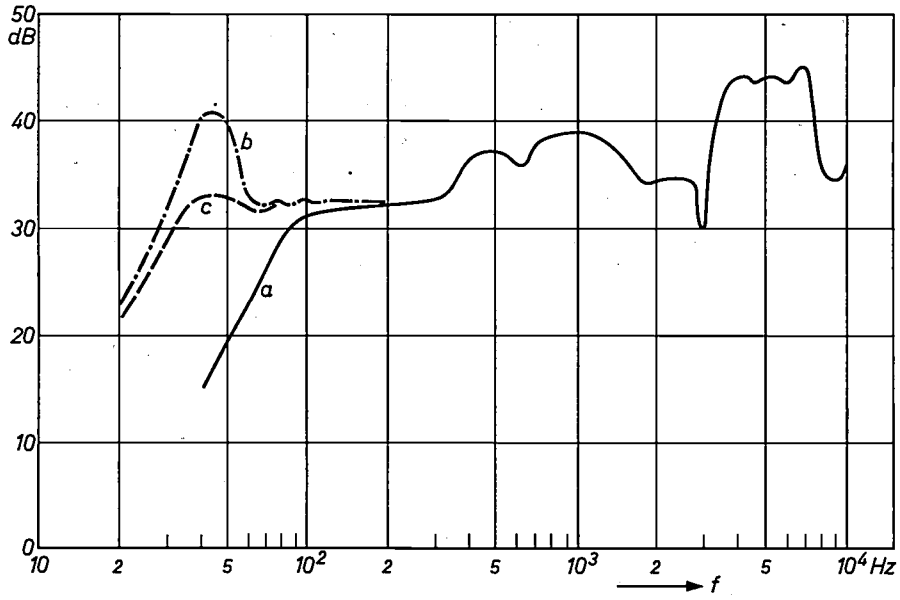


Fig. 11. Amplitude characteristics measured on the system of fig. 7, the loudspeaker being mounted in a completely closed cabinet with a volume of 10 litres, *a*) without motional feedback, *b*) with acceleration feedback, *c*) with both acceleration and velocity feedback.

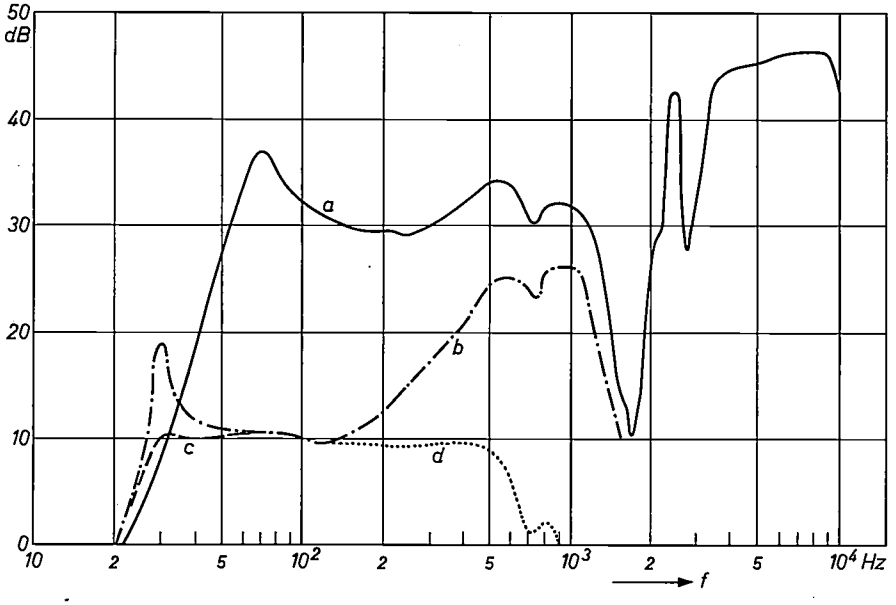


Fig. 12. Amplitude characteristics measured on the system of fig. 10, *a*) without motional feedback, *b*) with acceleration feedback, *c*) with both acceleration and velocity feedback, *d*) with the addition of a filter for separating high and low frequencies.

Fig. 12 shows the effect of acceleration feedback on the bass loudspeaker 9710 A, also mounted in a closed 10-litre cabinet. Here again, curve *a* relates to the case where no feedback is used, and curve *b* to the application of acceleration feedback within the limits of stability. We see that the resonant frequency of 70 Hz is shifted to 30 Hz. Curve *c* is the result of applying velocity feedback as well as acceleration feedback. Using an appropriate filter to separate the high and low frequency channels, we obtain the amplitude characteristic *d*, which is flat to within ± 1 dB from 23 to 500 Hz.

Non-linear distortion

The reduction of non-linear distortion that can be brought about by motional feedback depends, like the feedback factor, upon the frequency. We have already seen that the distortion caused by the loudspeaker

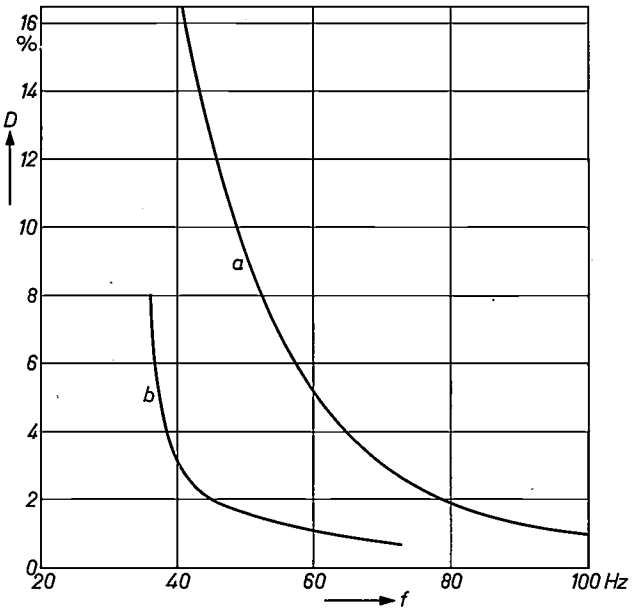


Fig. 13. Non-linear distortion D as a function of frequency f for loudspeaker 9710 AM. The radiated power is kept the same at all frequencies; the electrical power supplied to the loudspeaker is 4.5 W at 40 Hz. Curve a is obtained with no motional feedback, and curve b is obtained when acceleration and velocity feedback are combined in such a way as to give the amplitude characteristic c of fig. 11.

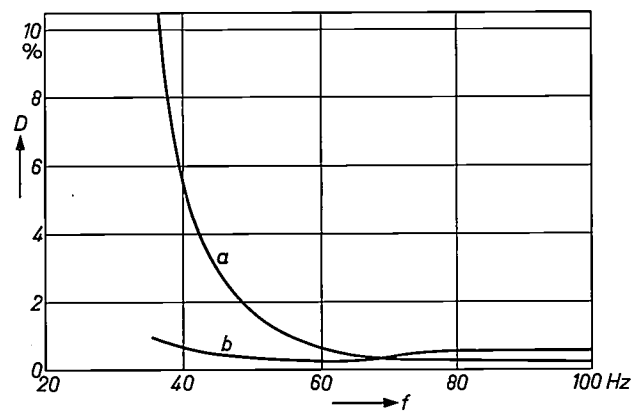


Fig. 14. Non-linear distortion D as a function of frequency f for the bass loudspeaker 9710 A under the same conditions as for fig. 14, *a*) without motional feedback, *b*) with motional feedback adjusted to obtain the amplitude characteristic d of fig. 12.

mechanism is greatest in the reproductions at low frequencies; motional feedback here is therefore of special importance. Figs. 13 and 14 give some results of measurements. *Fig. 13* shows the percentage distortion as a function of frequency, for the 9710 AM loudspeaker and at constant radiated power. The electrical power supplied to the loudspeaker is 4.5 W at 40 Hz. Curve a shows the distortion without motional feedback (but using electrical feedback in the amplifier), and curve b gives the result of using the motional feedback that produces the amplitude characteristic in fig. 11*c*. It can be seen from this figure that at 40 Hz the distortion has decreased from 17% to 3%. *Fig. 14* gives the corresponding curves for the bass loudspeaker 9710 A. Curve b here corresponds to the amplitude characteristic d in fig. 12.

These results show that the method of motional feedback which we have described can in fact yield substantial improvements in sound reproduction, and that it will be of great assistance in achieving good reproduction in a very wide frequency range with highly compact equipment.

Summary. One of the main causes of distortion in an electroacoustical system consisting of an amplifier and a loudspeaker is to be found in the mechanism of the loudspeaker, particularly if the loudspeaker is mounted in a small, closed cabinet. This distortion is not affected by the electronic negative feedback conventionally used for reducing distortion. With motional feedback, the loudspeaker is included in the feedback loop. This article discusses various forms of motional feedback. A method specially developed for feedback at low audio frequencies has been tested experimentally. In this method an acceleration transducer containing a piezoelectric element is fixed to the moving coil of a normal loudspeaker; this gives a feedback signal proportional to the acceleration of coil and cone. A signal proportional to the velocity is also obtained by integration and a certain fraction of this is added to the first feedback signal. In this way an amplitude characteristic can be obtained which is flat down to below the mechanical resonant frequency of the loudspeaker, and the non-linear distortion in this frequency range is substantially reduced. The sound pressure and the non-linear distortion have been measured as a function of frequency for two loudspeakers with motional feedback, one being a loudspeaker for the bass range and the other a loudspeaker for the full audio range. They showed a substantial improvement in reproduction of the bass.

A compact heat exchanger of high thermal efficiency

G. Vonk

One of the products of the experience gained with regenerators for gas refrigerating machines at Philips Research Laboratories has been the development of an extremely compact heat exchanger of high thermal efficiency. This new heat exchanger is used in the Philips helium liquefier.

Construction, characteristics and application

The production of cold at the temperature of liquid helium (4.2 °K) is usually accomplished by means of the Joule-Thomson effect. Compressed helium gas is cooled with the aid of a number of precooler stages and heat exchangers to below what is called the inversion temperature, and the gas is then allowed to expand through a throttle or expansion valve. In this process it is cooled still further and some of the gas is liquefied. The gas contained in the reservoir above the liquid is sucked in again by the compressor and gives up its cold in the heat exchangers to the helium flowing from the compressor to the expansion valve (*fig. 1*). If liquid helium is drawn off from the reservoir, then fresh helium gas must of course be continuously supplied (at *S* in fig. 1).

The efficiency of such a system for reaching an extremely low temperature depends to a considerable extent on the thermal efficiency of the heat exchangers. The cold loss in the heat exchangers, which has to be compensated by the cold production in the precooler stages and the expansion valve, must be kept as low as possible. The lower the temperature of refrigeration, the greater the power required per watt of refrigeration (the Carnot efficiency decreases with decreasing temperature).

In the research at Philips Research Laboratories into high-efficiency refrigeration at 4 °K using a gas refrigerating machine combined with a Joule-Thomson stage, it was therefore necessary to devise a heat exchanger with a very high thermal efficiency. To keep down the size of the equipment, the heat exchanger had to be very compact, which also keeps to a minimum the cold losses to the surroundings. This investigation led to the type of heat exchanger discussed below, known as the gauze-type heat exchanger.

Ir. G. Vonk, formerly with Philips Research Laboratories, is now with the Philips Product Division Glass, Eindhoven.

The principle used in the design of the new heat exchanger is shown in *fig. 2*. The channels for the two flows are almost completely filled with pieces of thin copper gauze mounted at right angles to the direction of flow, the wires of the mesh passing through the wall dividing one channel from the other. With this construction a large heat-transfer area in a small volume is achieved, and the wall between the channels has a very low thermal resistance, since copper is a

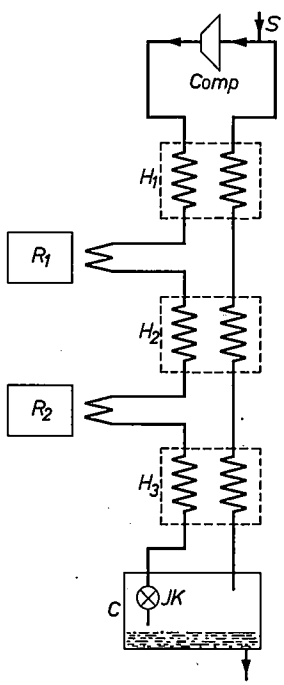


Fig. 1. Illustrating the principle of a helium liquefier or refrigerator at a temperature of 4 °K, making use of the Joule-Thomson effect. *Comp* compressor. R_1 and R_2 precooling stages (e.g. 80 and 15 °K). *JK* expansion valve; as it passes through this valve the helium gas is cooled sufficiently for part of it to become liquid. *C* helium vessel. H_1, H_2 and H_3 heat exchangers; the gas flowing from *C* to the compressor gives up its cold here to the gas flowing from the compressor to the expansion valve.

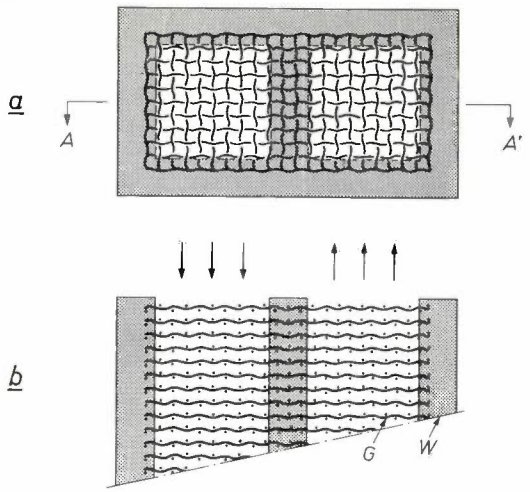


Fig. 2. Principle of the design of a gauze-type heat exchanger. *a*) Transverse cross-section of the channels. *b*) Longitudinal cross-section AA' of the channel. *G* copper gauze. *W* walls.

very good conductor of heat. In addition the heat transfer per unit area of the copper is also relatively high: it is known that the heat transfer to the packing material in a channel, in this case the copper gauze, is greater for a more finely divided material.

The actual construction is somewhat more complicated than fig. 2 suggests. There are in fact many channels, not just two. The two flows are distributed among the channels in such a way that each channel is surrounded by four others carrying the opposite flow (*fig. 3*). This arrangement gives a very intensive thermal contact between the two flows. A further advantage of using a large number of narrow channels instead of two wide ones is that the heat in the wires only has to be transported over a small distance. This also favours a high efficiency.

The new heat exchangers are constructed in the following way. A large number of squares of copper gauze are made, whose wire diameter is say 200 microns. A similar number of squares of paper of the same size

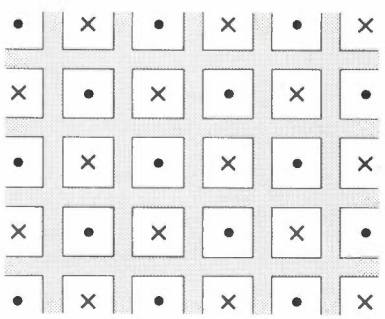


Fig. 3. In the actual heat exchangers there are in fact many channels, not just two as in fig. 2*a*. The two opposing flows (denoted by crosses and dots) are distributed in such a way that each channel is surrounded by *four* others carrying the opposite flow.

are also made: these are impregnated with a cresol resin and an array of holes is punched in them (*fig. 4*). The squares of gauze and paper are then stacked alternately in a jig which is then clamped up and heated to a temperature of about 150 °C. During the heating, an axial pressure is applied. The high temperature softens the resin in the paper, and as a result of the axial pressure, the pieces of paper are pressed into the gauze and stick firmly together. The stack is kept at the high temperature until the thermosetting process of the cresol resin is completed. The result is a coherent structure of paper and gauze containing a large number of parallel channels, as shown in fig. 3.

The compactness achieved by this method is apparent from the heat-transferring area per unit volume, which is as great as 3000 m^2/m^3.

As can be seen in the photograph, the channels at the edges are only half as wide as the inner ones, and the cross-section of those in the corners is only a quarter of that of the inner channels. The reason for this is that these channels are not in contact with four others but only with three (at the edges) or two (at the corners). Since the cross-sections of these channels are made smaller, the copper wires in them do not have to transport more heat than those in the other channels.

A useful feature of the construction is that the successive pieces of gauze are thermally insulated from one another by the impregnated paper, making the longitudinal thermal conduction of the heat exchanger particularly low. The heat flow from the "hot" to the "cold" end of the heat exchanger is therefore small, and thus the cold loss due to this heat flow is also small.

The thickness of the walls between the channels of the heat exchanger shown in fig. 4 is 2.5 mm. At this thickness the thermal resistance of the wall is still sufficiently low and there is absolutely no risk of gas leakage. If the heat exchanger has been properly made, dividing walls of this thickness do not give any leakage of gas even after the heat exchanger has been cycled a thousand times between the temperature of liquid nitrogen and room temperature. In addition, walls 2.5 mm thick can withstand a pressure difference of 200 atmospheres.

The flows are distributed correctly among the channels by means of distribution blocks fitted to the two ends of the heat exchangers: a pattern of holes drilled through the blocks connects the appropriate channels. Fig. 4 shows a transparent model of one of these terminal blocks. The real blocks are made of brass, and are fixed to the heat exchanger by means of an adhesive bond which can withstand extremely low temperatures.

Because of its large number of parallel channels the device is not limited to heat exchange between just *two*

flows. The heat exchanger discussed and illustrated here can in principle be used for simultaneous heat exchange between 49 different flows. The Philips helium liquefier makes use of gauze-type heat exchangers for heat exchange between *three* flows.

Since the heat exchanger was developed for use in a system like that of fig. 1, it was designed with particular attention to heat exchange between gases. This kind of exchanger can however be used equally well for heat exchange between gas and liquid or between liquids.

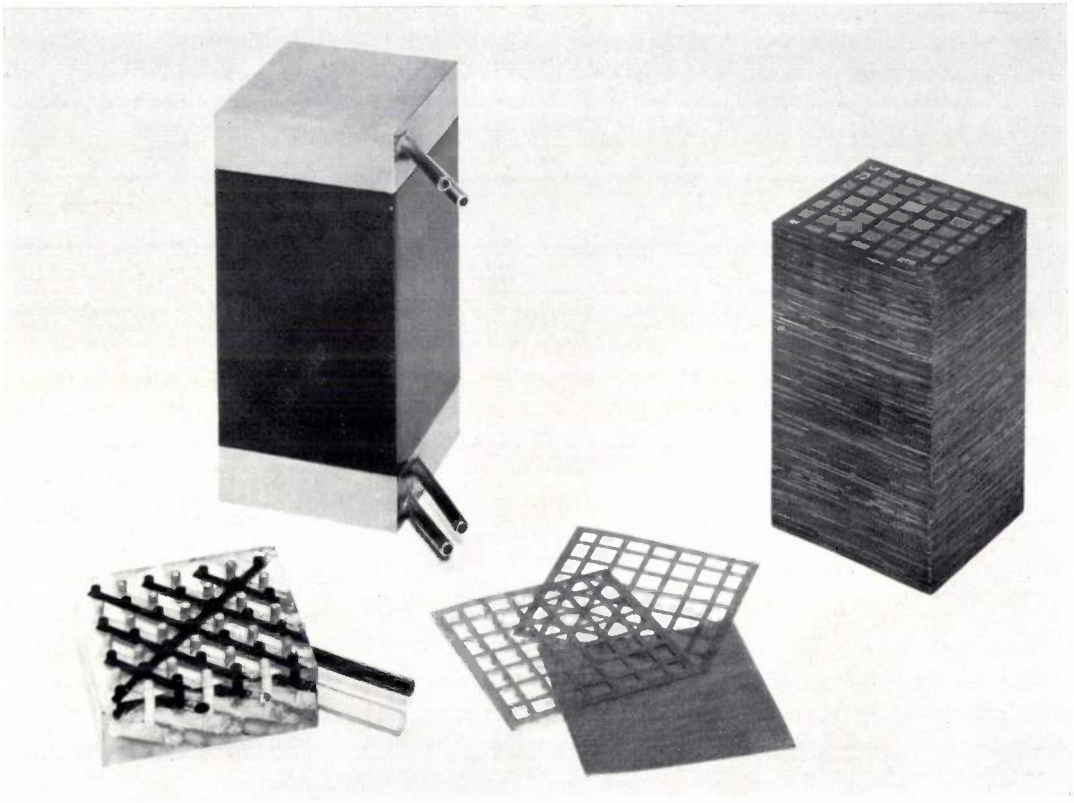


Fig. 4. Complete heat exchanger (at the back on the left) together with a number of components. The heat exchanger consists of a block as shown in figs. 2 and 3 (on the right; height 12 cm, cross-section 8×8 cm), terminated at both ends by a brass block for dividing the two flows between the channels. At the front on the left is a plastic model of a terminal block, and one of the pieces of copper gauze and two of the sheets of impregnated perforated paper from which the stack is built up are shown in the centre of the picture.

Some performance figures

Efficiency

The thermal efficiency η of a heat exchanger is defined as the ratio of the actual quantity of heat transferred to the maximum quantity of heat that can be transferred. We shall now consider the special, simple case of a balanced gauze-type heat exchanger, i.e. a heat exchanger in which the product of mass flow and specific heat is identical for both flows. If we neglect the insulation losses and the longitudinal thermal conduction, and assume that the heat transfer coefficient and the thermal conductivity of copper are not temperature-dependent, then the temperature in such a heat exchanger varies as shown in *fig. 5*. The quantity η is equal to the ratio of the difference between inflow and outflow temperatures — identical for both flows in this case — to the maximum value which this difference could have:

$$\eta = \frac{T_{\mathrm{Hi}} - T_{\mathrm{Hu}}}{T_{\mathrm{Hi}} - T_{\mathrm{Li}}}. \qquad (1)$$

In the above case the relation between η and the design parameters also assumes a simple form:

$$\eta = \frac{\Lambda^*}{\Lambda^* + 1}. \qquad (2)$$

Here Λ^* is a dimensionless quantity for which:

$$\frac{1}{\Lambda^*} = \frac{1}{\varepsilon_{\mathrm{H}}\Lambda_{\mathrm{H}}} + \frac{1}{\varepsilon_{\mathrm{L}}\Lambda_{\mathrm{L}}} + \frac{\dot{W}\delta}{\lambda_{\mathrm{Cu}} S}. \qquad (3)$$

The three terms on the right-hand side may be re-

garded as three reciprocal, dimensionless thermal resistances. The first term relates to the heat transfer from the relatively hot flow to the copper wires (subscripts H), the second to the loss of heat by the wires to the relatively cold flow (subscripts L). The third term relates to the passage of heat through the wall. For each of the channels the quantity Λ is equal to $\alpha A/\dot{W}$, where α is the coefficient of heat transfer between flow and gauze, A is the area of the heat-transfer surface and $\dot{W}$ the heat capacity flow, i.e. the product of mass flow and specific heat. In addition, ε is the "fin efficiency" (see below), δ is the thickness of the dividing wall, λ_{Cu}

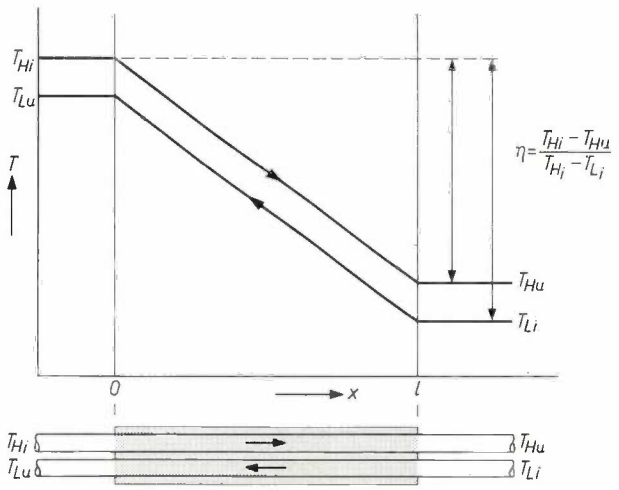


Fig. 5. Variation of temperature T with distance x along the channels (length l) for a balanced heat exchanger with no insulation losses or longitudinal thermal conduction. The temperatures with a subscript H relate to the "hot" flow which is falling in temperature, and those with the subscript L to the "cold" flow which is rising in temperature. The efficiency η is the ratio of the temperature differences indicated by the two arrows.

the thermal conductivity of copper, and S the total cross-sectional area of all the copper wires passing through a dividing wall. For any given heat exchanger the third term on the right-hand side of equation (3) can be directly calculated. The Λ-values have been taken from measurements which have been made on regenerators for gas refrigerating machines at Philips Research Laboratories. It was found that, for a stack of n gauzes not in contact with each other, Λ can be represented by $\Lambda = c_\Lambda \times 2n$, and for gauzes that are in contact with each other by $\Lambda = c_\Lambda \times l/d$, where l is the height of the stack and d the diameter of the copper wires. *Fig. 6* shows how c_Λ depends on the Reynolds number Re, for four values of the packing factor f, when the flowing medium is a gas. An unexpected aspect of the above results is that c_Λ does not depend on any geometric quantities other than f and d.

Now let us look at ε. Since the thermal resistance of the wires differs from zero, heat transport takes place from one channel to another only if there is a tempera-

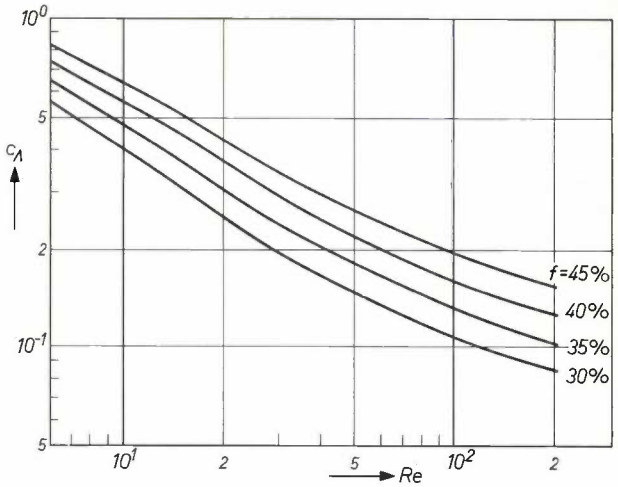


Fig. 6. The dimensionless quantity c_Λ, which is one of the factors that determine the efficiency of the heat exchanger, as a function of the Reynolds number Re, for four values of the packing factor, f, i.e. the part of the volume of the channels taken up by the copper wire. For small Re, $c_\Lambda \propto 1/Re$; with increasing Re the slope of the curve decreases. ($Re = \varrho v d/\mu$, where ϱ is the density, and v the velocity which the gas would have, for the same mass flow, if the channel were completely empty; μ is the dynamic viscosity of the gas.)

ture gradient in the copper wires. This applies not only to the part of the wires contained in the wall, but equally to the parts inside the channels. The temperature gradient in the parts of the wires in the channels can be allowed for in the same way as with cooling fins, by using in the calculations an apparent coefficient of heat transfer α', which is smaller than the true one: $\alpha' = \varepsilon\alpha$. The value of the quantity ε, the "fin efficiency", can be calculated when the form of the fin is known; in our case it is a wire.

In *fig. 7* the calculated and measured values of Λ^* are plotted as a function of the mass flow $\dot{m}$ of the medium for the heat exchanger in fig. 4. The difference

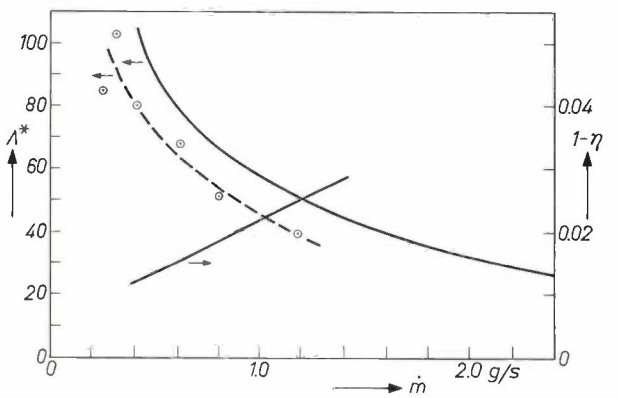


Fig. 7. Variation of the dimensionless quantity Λ^*, which determines the efficiency η, as a function of the mass flow $\dot{m}$ for a heat exchanger like that of fig. 4. The solid line relates to the calculated values; the values derived from measurements differ only very slightly from them (dashed line). The corresponding curve of $1-\eta$ is also shown. The true values of $1-\eta$ are about 10% higher because of the cold loss due to longitudinal conduction of heat.

between the calculated and the measured values is 15-20%. One of the reasons for this discrepancy could be that the wall of the channels is not flat and the flow is not in actual fact distributed completely uniformly over the channels. From the values found for A^* it follows at once that η varies between about 99% and 96% in the region of $\dot{m}$ values investigated. Fig. 7 also shows a curve representing the variation of $1-\eta$ with $\dot{m}$. A curve of this kind shows more clearly how the cold loss increases with increasing $\dot{m}$ than a curve of η plotted against $\dot{m}$.

The measured values of A^* given in fig. 7 were found by deducting from the measured total losses the cold loss due to longitudinal thermal conduction in the heat exchanger — this loss has in fact been disregarded here. As already mentioned, this heat conduction is relatively small: at room temperature the longitudinal thermal conductivity is only 0.5 W/m°C, and it is even less at lower temperatures. In practice — i.e. in the case of fig. 7 at $\dot{m}$ values of 1 to 2 g/s — the cold loss due to longitudinal heat conduction is about 10% of the total cold loss of the heat exchanger. The true values of $1-\eta$ in fig. 7 are therefore of the order of about 10% greater than those shown.

Flow resistance

An important quantity in assessing the usefulness of a heat exchanger is the flow resistance. The expression for the pressure difference Δp across this resistance may be written in the form:

$$\Delta p = C_W \times \tfrac{1}{2}\varrho v^2 l/d. \quad . \quad . \quad . \quad . \quad (4)$$

Here l and d have the same meaning as above; ϱ is the density of the gas, v the velocity which the gas would have for the mass flow if the channel were completely empty, and C_W is a dimensionless quantity depending solely on the Reynolds number and the packing factor (*fig. 8*).

Comparing fig. 8 with fig. 6 we see that with increasing packing factor the quantity C_W increases (and with it the pressure difference Δp), which is a disadvantage. However, c_A (and hence A^* or the efficiency η) increase with the packing factor, which is an advantage. It is therefore necessary to strike a compromise. The proper choice depends on the nature of the application.

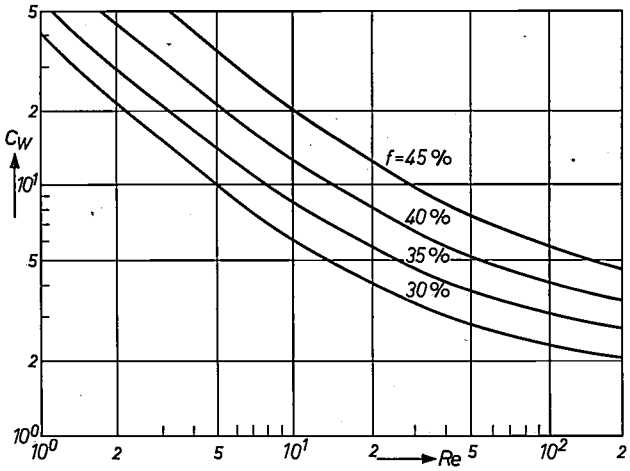


Fig. 8. The resistance factor C_W — see equation (4) — as a function of Re for the same four values of the packing factor f as in fig. 6. For small Re, $C_W \propto 1/Re$.

In an arrangement like that of fig. 1 the total flow resistance is determined mainly by the resistance encountered by the expanded gas: owing to the low pressure the density of this gas is low and the velocity, for the same mass flow, is therefore high. Generally speaking, the total pressure difference can be kept sufficiently low in a heat exchanger with the channel pattern shown, where the flow cross-sections for high-pressure and low-pressure flow are equal. If the pressure of the expanded gas is extremely low — this pressure is related to the temperature at which the cold is required — it is better to increase the flow cross-section for the expanded gas at the expense of that for the compressed gas. This can easily be done in manufacture by punching a different pattern of holes in the impregnated sheets of paper.

Summary. An extremely compact counter-flow heat exchanger has been developed that has a high thermal efficiency and a not unduly high flow resistance for gases. It is made from a stack of alternate rectangular sheets of copper gauze and sheets of paper of the same size, impregnated with cresol resin, in which a pattern of holes has been punched. The stack is heated under pressure, which softens the resin so that the sheets of paper stick to each other through the gauze. The resin then hardens, and a large number of channels are thus obtained, separated by walls penetrated by the wires of the gauze, which gives the walls a low thermal resistance. The flows are divided among the channels by means of terminal blocks in such a way that each channel is surrounded by a number of channels through which the counter-flow passes. The thermal conductivity in the direction of flow is extremely low.

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, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, 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.

G. A. Acket & J. de Groot: Measurements of the current/field strength characteristic of *n*-type gallium arsenide using various high-power microwave techniques.
IEEE Trans. **ED-14**, 505-511, 1967 (No. 9). *E*

G. Arlt & P. Quadflieg: Piezoelectricity in III-V compounds with a phenomenological analysis of the piezoelectric effect.
Phys. Stat. sol. **25**, 323-330, 1968 (No. 1). *A*

F. Th. Backers: Einige Betrachtungen über die PAL-Verzögerungsleitung.
Nachrichtentechn. Z. **20**, 473-477, 1967 (No. 8). *E*

V. Belevitch: On the theory of non-linear resonance in narrow-band circuits.
Philips Res. Repts. **23**, 87-102, 1968 (No. 1). *B*

V. Belevitch & J. Neirynck: Eddy currents in thin surfaces of revolution at low frequencies.
Philips Res. Repts. **23**, 77-86, 1968 (No. 1). *B*

H. J. Bensieck: Second-order approximations for the determination of the resonance frequencies of partially filled resonators.
Philips Res. Repts. **23**, 103-107, 1968 (No. 1). *H*

G. Blasse & A. Bril: Study of energy transfer from Sb^{3+}, Bi^{3+}, Ce^{3+} to Sm^{3+}, Eu^{3+}, Tb^{3+}, Dy^{3+}.
J. chem. Phys. **47**, 1920-1926, 1967 (No. 6). *E*

G. Blasse & A. Bril: Fluorescence of rhodium-activated aluminum oxide.
J. Electrochem. Soc. **114**, 1306-1307, 1967 (No. 12). *E*

H. van den Boom: Simple method of rotating samples in microwave cavities at low temperatures.
Rev. sci. Instr. **38**, 1529-1530, 1967 (No. 10). *E*

K. Bulthuis: Effect of uniaxial stress on silicon and germanium.
Physics Letters **25A**, 512-514, 1967 (No. 7). *E*

K. Bulthuis: Effect of high uniaxial pressure along the main crystallographic axes for silicon and germanium.
Philips Res. Repts. **23**, 25-47, 1968 (No. 1). *E*

K. H. J. Buschow, J. F. Fast, A. M. van Diepen (Natuurkundig Laboratorium der Universiteit van Amsterdam) **& H. W. de Wijn** (idem): Magnetic dilution of rare-earth aluminum cubic Laves phases RAl_2.
Phys. Stat. sol. **24**, 715-720, 1967 (No. 2). *E*

A. Charles-Georges & A. Salmon: Usinage de cavités de pompage pour lasers "solides".
Acta electronica **10**, 315-330, 1966 (No. 3). *L*

A. Cohen & J. 't Hart (Institute for Perception Research, Eindhoven): On the anatomy of intonation.
Lingua **19**, 177-192, 1967 (No. 2).

E. H. P. Cordfunke (Reactor Centrum Nederland) **& A. A. van der Giessen:** Particle properties and sintering behaviour of uranium dioxide.
J. nucl. Mat. **24**, 141-149, 1967 (No. 2). *E*

M. B. Das (Associated Semiconductor Manufacturers Ltd.): Charge-control analysis of m.o.s. and junction-gate field-effect transistors.
Proc. IEE **113**, 1565-1570, 1966 (No. 10).

M. B. Das (Associated Semiconductor Manufacturers Ltd.): Generalised high-frequency network theory of field-effect transistors.
Proc. IEE **114**, 50-59, 1967 (No. 1).

R. Dessert: Lasers "solides" à fonctionnement continu.
Acta electronica **10**, 295-314, 1966 (No. 3). *L*

A. van der Drift, Tine E. Horsman & C. Langereis: Investigations on the preferred orientations in vapour-deposited lead-monoxide layers.
Philips Res. Repts. **23**, 48-61, 1968 (No. 1). *E*

P. P. J. van Engelen & J. C. M. Henning: Electronic structure of a Ti^{3+} centre in $SrTiO_3$.
Physics Letters **25A**, 733-735, 1967 (No. 10). *E*

J. L. Goldstein (Institute for Perception Research, Eindhoven): Auditory spectral filtering and monaural phase perception.
J. Acoust. Soc. Amer. **41**, 458-479, 1967 (No. 2).

N. Hansen: Evaluation of the sticking probability of gases chemisorbed on metal films without measurements of gas pressure.
Suppl. Nuovo Cim. **5**, 389-398, 1967 (No. 2). *A*

G. Klein & H. Hagenbeuk: Accurate triangle-sine converter.
Electronic Engng. **39**, 700-704, 1967 (No. 477). *E*

A. Klopfer: Interactions of electrons with adsorbed water.
Suppl. Nuovo Cim. **5**, 606-611, 1967 (No. 2). *A*

D. J. Kroon & J. M. van Nieuwland: Techniques of propagation at millimetre and submillimetre wavelengths.
Spectroscopic techniques for far infra-red, submillimetre and millimetre waves, ed. D. H. Martin, North-Holland Publ. Co., Amsterdam 1967, p. 309-380. *E*

H. de Lang: In memoriam Dr. P. M. van Alphen.
Ned. T. Natuurk. **33**, 268-270, 1967 (No. 9). *E*

W. Lems: Magnetic properties of electrodeposited FeNi films, II.
Philips Res. Repts. **23**, 62-76, 1968 (No. 1). *E*

F. A. Lootsma: Extrapolation in logarithmic programming.
Philips Res. Repts. **23**, 108-116, 1968 (No. 1). *E*

M. H. van Maaren: Note on the diffusion and solubility of lithium in gallium antimonide.
Phys. Stat. sol. **24**, K 125-128, 1967 (No. 2). *E*

G. Marie, P. Wurtz & R. Le Pape: Etude d'un laser au néodyme déclenché par un obturateur à effet Pockels.
Acta electronica **10**, 249-293, 1966 (No. 3). *L*

A. G. van Nie: Measuring currents down to 10^{-17} A with a low noise level by means of a dynamic capacitor electrometer.
IEEE Int. Conv. Rec. **15**, Part 8, 59-65, 1967. *E*

J. M. van Nieuwland & M. T. Vlaardingerbroek: Cyclotron waves in InSb.
IEEE Trans. **ED-14**, 596-599, 1967 (No. 9). *E*

A. van Oostrom: Modulation of Bayard-Alpert ionization gauges with grid end-caps.
J. sci. Instr. **44**, 927-930, 1967 (No. 11). *E*

C. van Opdorp: The concentration gradient of Zn near a *p-n* junction in III-V compounds.
J. appl. Phys. **38**, 5411-5412, 1967 (No. 13). *E*

P. Penning & D. Polder: Dynamical theory for simultaneous X-ray diffraction, Part I. Theorems concerning the *n*-beam case;
P. Penning: idem, Part II. Application to the three-beam case.
Philips Res. Repts. **23**, 1-11, 12-24, 1968 (No. 1). *E*

J. Polman: Sensitive 4 mm Lecher wire interferometer for electron concentration measurements in low-density plasmas.
Rev. sci. Instr. **38**, 1631-1633, 1967 (No. 11). *E*

M. J. Prescott & J. Basterfield: The observation of dislocations in yttrium gallium garnet by a photoelastic method.
J. Mat. Sci. **2**, 583-588, 1967 (No. 6). *M*

O. Reifenschweiler: Bohrlochuntersuchungen mit Neutronen.
Haus der Technik - Vortragsveröffentlichungen No. 133, 30-42, 1967. *E*

G. Renelt & H. W. Neuhaus: Übertragungseigenschaften des Lesedrahtes einer Ferritkernspeichermatrix.
Nachrichtentechn. Z. **20**, 722-728, 1967 (No. 12). *H*

G. Simon: Calculation of the elastic properties of germanium.
J. Phys. Chem. Solids **28**, 2349-2358, 1967 (No. 12). *A*

G. Simon: Calculations of the elastic moduli and their pressure derivatives for the alkali metals.
J. Phys. Chem. Solids **29**, 63-68, 1968 (No. 1). *A*

J. M. Stevels: Centres de couleur induits par irradiation dans les systèmes vitreux.
Verres et Réfr. **21**, 279-288, 1967 (No. 4). *E*

T. L. Tansley: Forward current injection modulation of photocurrent in *p-n* heterojunctions.
Phys. Stat. sol. **24**, 615-622, 1967 (No. 2). *M*

B. Tuck: Photoluminescence of compensated *p*-type GaAs.
J. Phys. Chem. Solids **28**, 2161-2168, 1967 (No. 11). *M*

K. J. de Vos (Philips Radio, Gramophone and Television Division, Eindhoven): The relationship between microstructure and magnetic properties of alnico alloys.
Thesis, Eindhoven 1966.

J. C. Walling: Characteristic admittance of an array of rectangular conductors.
Electronics Letters **3**, 481-483, 1967 (No. 11). *M*

H. W. de Wijn (Natuurkundig Laboratorium der Universiteit van Amsterdam), **A. M. van Diepen** (idem) & **K. H. J. Buschow:** Nuclear magnetic resonance and magnetic susceptibility of $SmAl_3$.
Phys. Rev. **161**, 253-257, 1967 (No. 2). *E*

G. Zanmarchi: Optical measurements on the antiferromagnetic semiconductor MnTe.
J. Phys. Chem. Solids **28**, 2123-2130, 1967 (No. 11). *E*

PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 6

Principles and design of the automatic single-crystal diffractometer system PAILRED

P. G. Cath and J. Ladell

I. Principles of the diffractometer
II. The automatic control circuits

One of the main applications of X-ray crystallography is the determination of crystal structures, i.e. the arrangement and location of the atoms within the unit cell. The information required is obtained from the intensities of the X-ray reflections from the different crystal planes. Some of the determinations of structure that are made today require the measurement of many thousands of reflections. In the last few years large digital computers have been used for processing these vast amounts of data; but only recently have automatic diffractometer systems become available to collect the data.

One such system, known by the acronym PAILRED, has been developed by the authors. It uses a digitally-controlled diffractometer on which the crystal and the X-ray detector are positioned by means of an analog mechanism for linear tracking.

Part I reviews the theory of X-ray diffraction and describes the operation of the diffractometer. Part II discusses the electronic circuits and the logic configuration of the computer that controls the operation of the system.

I. Principles of the diffractometer

Crystal structures are determined for several reasons. Solid state physicists wish to know how each ion is surrounded by its neighbors; chemists and biochemists are interested in the shape of large molecules. Some methods that are used to solve a crystal structure have been discussed in this journal by Braun and Van Bommel [1].

The solution of a complicated crystal structure requires the determination of a large number of parameters from a much larger number of X-ray reflection intensities. This task can only be performed by a large and fast electronic computer. The number of reflections to be measured may range from about one hundred for a simple structure to several hundred thousand for protein crystals. A single crystal will reflect (or diffract) X-rays when it is properly positioned with respect to the incident X-ray beam. The required orientation of the crystal and the direction of the diffracted beam are determined by the size and shape of the unit cell and the wavelength of the X-rays.

Dr. P. G. Cath and Dr. J. Ladell are with Philips Laboratories, Briarcliff Manor, New York, U.S.A.

[1] P. B. Braun and A. J. van Bommel, X-ray determination of crystal structures, Philips tech. Rev. 22, 126-138, 1960/61.

Until recently, the intensities of the reflections were mostly recorded photographically. This is a fairly rapid technique; but high accuracy is difficult to obtain and much effort is required before the data can be entered into a computer.

A higher accuracy and a digital output can be obtained by using an electronic radiation detector, such as a scintillation counter [2]. This requires the correct positioning of crystal and detector for each fractometer. This alignment must precede the automatic data collection. The complete system is shown in *fig. 1*.

The diffractometer is located on top of the X-ray generator which is shown on the right of the electronic console, which consists of two sections. The section shown on the left in fig. 1 contains the radiation-measuring circuits. The X-ray quanta reflected by the crystal under study are detected by a scintillation

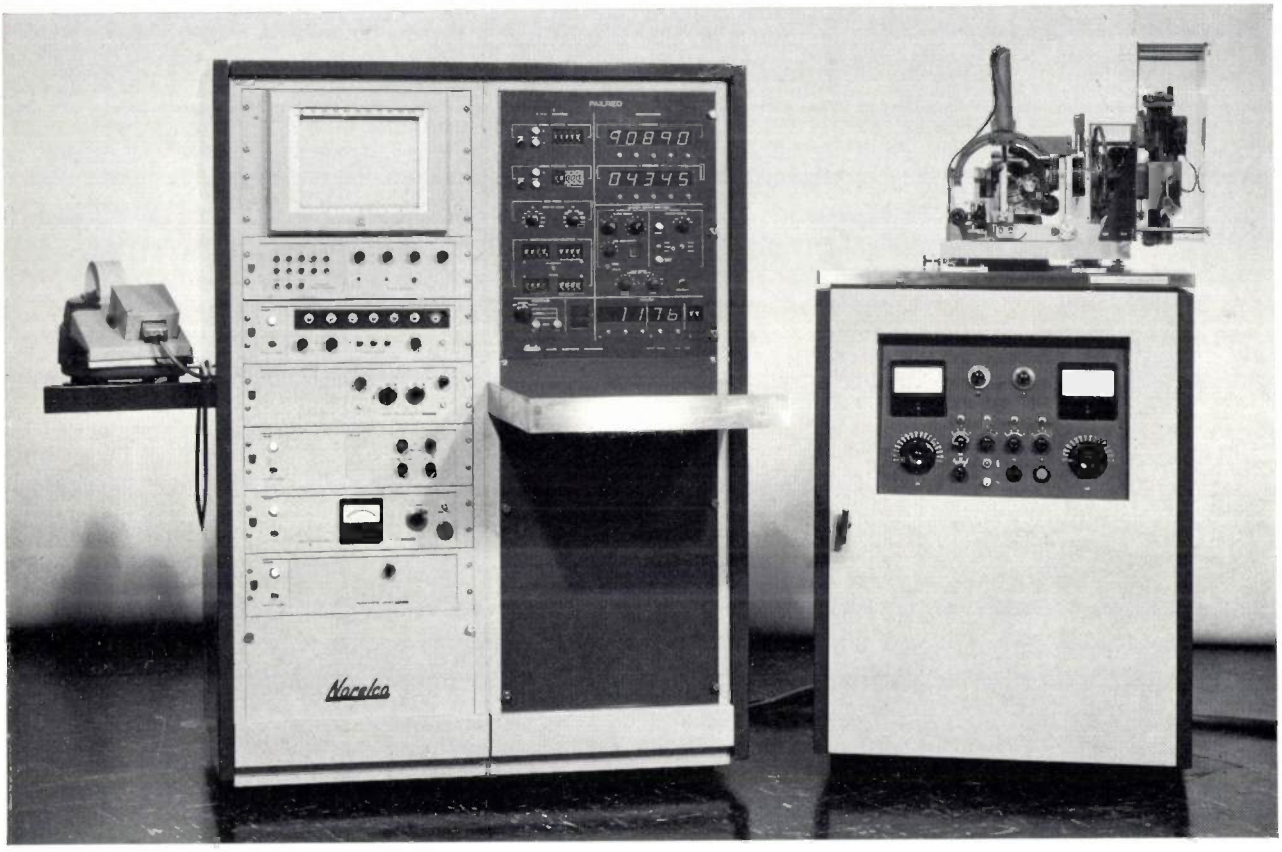


Fig. 1. The heart of the PAILRED automatic single-crystal diffractometer system is the diffractometer, which is located on top of the X-ray generator on the right. The motions of the diffractometer are controlled by automatic control circuits and the X-ray intensities are measured and recorded in the radiation-measuring circuits located in the console on the left.

reflection separately. To do this manually for measuring several thousand reflections would be very time-consuming indeed, to say the least. Thus automation of this data-collection process is almost a necessity.

The single-crystal diffractometer system PAILRED (this is an acronym for Philips Automatic-Indexing Linear Reciprocal-space Exploring Diffractometer) performs this vast data-collecting operation [3]. It can measure several hundred reflections per day automatically, depending upon the experimental parameters. Special features, the manual operations, have been included in the design to facilitate the task of aligning the major axes of the crystal with those of the diffdetector, also mounted on the diffractometer. The output signals from the detector are applied to a single-channel pulse-height analyzer in which X-ray quanta with a wavelength significantly different from the radiation selected for the experiment, are rejected. Quanta with a slightly different wavelength have already been removed by the crystal monochromator to be discussed later on. The combination of crystal monochromator and pulse-height analyzer leads to a low background and hence to a high accuracy. The quanta that pass the pulse-height analyzer are counted in an electronic counter, the scaler. The instantaneous intensity of the detected radiation is also displayed

on a strip-chart recorder which serves as a monitor during automatic data collection. The left-hand side of the console further contains the high-voltage power supply for the scintillation detector, and the printer-control circuit which is used to print the contents of the scaler at the end of each measurement to provide a permanent record. Not shown is a paper-tape punch, or a similar digital recorder, whose output is used to enter the data into an electronic computer for processing.

The automatic sequences are generated in the control unit which also controls the motions of the diffractometer. This unit — a fixed-program special-purpose digital computer — is located in the right-hand side of the electronic console. Part II of this paper will give a detailed description of the operation of the control unit and the manner in which it directs the motions of the diffractometer and synchronizes these with the data-collecting and print-out process. However, before such a description can be given, a basic understanding is required of the operation of the diffractometer and the different operations that must be performed when measuring X-ray reflection intensities.

Bragg's law and reciprocal space

To enable the reader to understand the operation of the PAILRED diffractometer, Bragg's law will be briefly reviewed and the concept of reciprocal space will be introduced [4].

Bragg's law formulates the conditions under which a single crystal can reflect X-rays. As shown in *fig. 2a* a set of parallel planes in a crystal, having a spacing equal to d, will reflect an incident X-ray beam when the following relation exists between the angle Θ, the spacing d, and the X-ray wavelength λ:

$$\lambda = \frac{2d}{n} \sin \Theta, \qquad \ldots\ldots \quad (1)$$

where n is a small integer which denotes the order of the reflection. One can also consider $d' = d/n$ as the spacing of a different set of planes in the crystal and in this case the Bragg equation becomes:

$$\lambda = 2d' \sin \Theta. \qquad \ldots\ldots \quad (2)$$

The angle by which a crystal has to be rotated and the angle at which the detector must be placed to detect the reflection from a given set of planes in the crystal, can also be determined using the concept of *reciprocal space*. Each set of parallel planes with spacing d' is now represented by a vector perpendicular to this set with a length equal to $1/d'$. The Bragg condition for reflection, using this reciprocal vector, is shown in fig. 2*b*. In reciprocal space a sphere (in the figure represented by a circle) is drawn with radius $1/\lambda$ and center C. The incident beam coincides with the line ACO. Point O is the origin of reciprocal space and the center of rotation when the crystal is rotated. In the starting position a reciprocal vector OP is assumed to be perpendicular to the incident beam. The reflection condition is fulfilled when the crystal is rotated to a position such that the end point of this vector lies on the circle (P'). The direction of the diffracted beam is given by the vector CP', as may easily be verified. This construction is called the Ewald construction.

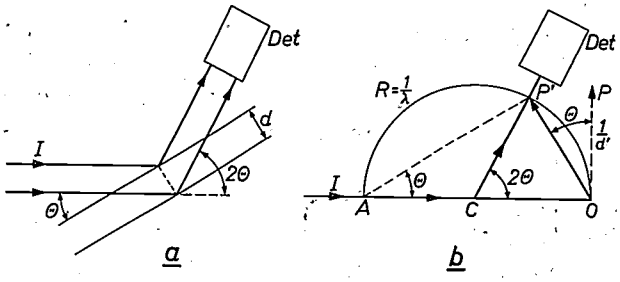


Fig. 2. *a*) The conditions under which a set of parallel planes in a crystal, having a spacing d, can reflect X-rays are governed by Bragg's law. Reflection takes place when the difference in path length between the two rays that are shown equals an integral number of X-ray wavelengths. I incident beam. *Det* detector. *b*) An alternate form of the Bragg condition makes use of the reciprocal vector OP, perpendicular to the set of crystal planes it represents and whose length is equal to $1/d'$. A set of crystal planes will reflect X-rays when the angle between this set and the incident beam I is such that the end point of its reciprocal vector lies on the circle with radius $1/\lambda$.

Many sets of parallel planes exist in a crystal, all positioned at different angles with respect to each other, and having different d' spacings. To select any one of these sets and position it and the detector in such a way that the Bragg condition (eq. 2) is fulfilled, involves a fair amount of computational effort using trigonometric relations. The concept of reciprocal space presents a mathematical aid for finding these angle settings in an orderly and much simpler manner. This arises from the fact that the end points of the reciprocal vectors for all the possible crystal planes form a regular three-dimensional periodic net. This will be shown in the following paragraphs.

[2] J. A. W. van der Does de Bye, The scintillation counter, Philips tech. Rev. **20**, 209-219, 1958/59.

[3] Previous communications on PAILRED are: J. Ladell and K. Lowitzsch, Automatic single crystal diffractometry, I. The kinematic problem, Acta cryst. **13**, 205-215, 1960; J. Ladell, Evolution of the PAILRED system, Norelco Reporter, **12**, 48-56, 1965; J. Ladell and P. G. Cath, A description of the PAILRED system, Norelco Reporter **12**, 63-69, 1965; A. H. Gomes de Mesquita, PAILRED, an automatic single-crystal diffractometer, Philips tech. Rev. **26**, 316-317, 1965.

[4] An alternative, and more general, way of introducing the reciprocal lattice using the Fourier transform theorem has been presented by Braun and Van Bommel [1].

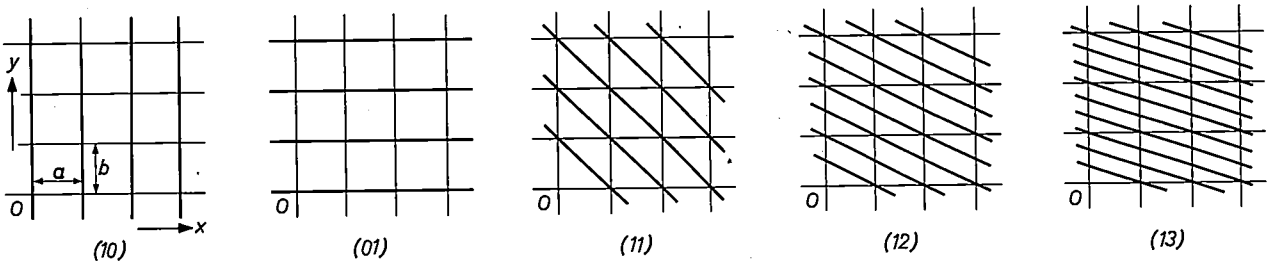


Fig. 3. A two-dimensional crystal is represented as a combination of unit cells (sides a and b) of which only nine are shown. Each set of crystal planes in this repetitive array can be identified by a set of integers h and k, the Miller indices. The five figures show five different sets of planes.

To explain the properties of reciprocal space an imaginary two-dimensional crystal with a rectangular unit cell will be considered as an example. In this case the planes reduce to lines. *Fig. 3* shows a two-dimensional repetitive pattern representing the two-dimensional crystal. The basic building block of this pattern is called the unit cell and its dimensions are a and b.

The different "planes" in the crystal are identified by the Miller indices h,k where h and k are integers. When an index is negative, the minus sign is put above it. The "planes" (hk) are parallel to the "plane" intersecting the x- and y-axis in the points $x = a/h$ and $y = b/k$, and the spacing d' is equal to the distance from the origin to this "plane". In fig. 3 the sets of "planes" (10), (01), (11), (12), and (13) are shown.

How to find the reciprocal vector for an arbitrary set of "planes" is shown in *fig. 4*. The "planes" are parallel to the hypotenuse of the triangle OQR. The sides of this triangle OQ and OR have lengths $x\,(=a/h)$ $y\,(=b/k)$. If a point P is chosen such that its coordinates are $(1/x, 1/y)$, then the vector OP is the reciprocal vector for the set of "planes" parallel to QR and having a spacing $OD = d'$. This can be seen as follows: If OP is indeed the reciprocal vector then it must be perpendicular to the "plane" QR and its length must be equal to $1/d'$. That OP is perpendicular to QR follows from:

$$\tan\alpha = \frac{PA}{OA} = \frac{1/y}{1/x} = \frac{x}{y} = \frac{OQ}{OR},$$

while the length of OP is:

$$OP = \frac{OA}{\cos\alpha} = \frac{1/x}{OD/OQ} = \frac{1/x}{d'/x} = 1/d'.$$

Therefore, OP is the reciprocal vector for the plane QR, and the end point P of the vector for a crystal plane with indices h,k will have the coordinates h/a and k/b.

The reciprocal net of the two-dimensional crystal, containing the end points of all the reciprocal vectors,

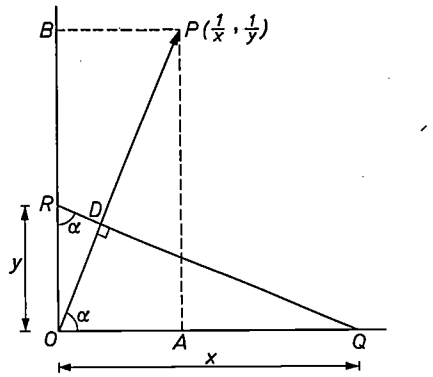


Fig. 4. The vector OP is the reciprocal vector for the set of parallel planes parallel to the line QR and whose spacing d' equals the length OD.

can now be constructed. In this net, the set of (hk) planes is represented by a point $P(hk)$ with coordinates h/a and k/b, and these points form a repetitive rectangular pattern in which the repeat distances are $1/a$ and $1/b$. These distances are called the dimensions of the "reciprocal cell" and are denoted by a^* $(= 1/a)$ and b^* $(= 1/b)$. The reciprocal net is shown in *fig. 5*.

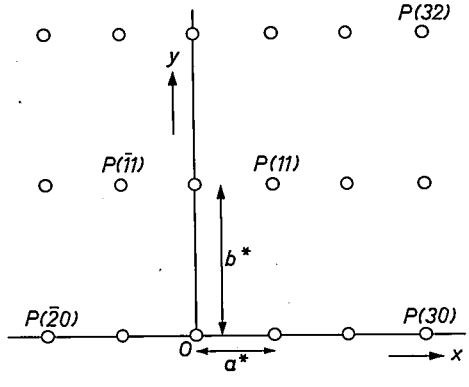


Fig. 5. The end points P of the reciprocal vectors of planes (hk) form a regular repetitive array in what is called reciprocal space. Just as in the case of direct space (fig. 3), the array has a unit cell which is called the reciprocal cell. The x- and y-axis of reciprocal space lie in the direction of the edges a^* and b^* of the reciprocal cell. ($\bar{2}$ denotes -2.)

The reciprocal lattice of a three-dimensional crystal is constructed in a similar way.

Because reciprocal space is a mathematical concept, it is difficult to visualize. It is important to realize, however, that when the crystal is rotated by an angle, e.g. ω, the reciprocal lattice is also rotated by this same angle.

To determine the proper angle settings for the crystal and the detector to measure the X-ray reflection from the crystal plane (*hk*) the Ewald construction of fig. 2*b* can be combined with the concept of reciprocal space in which the points that define the different crystal planes lie in a repetitive pattern. This is shown in *fig. 6*. To bring the plane ($\bar{3}2$) into reflection the end point of the reciprocal vector *OP* of ($\bar{3}2$) must be brought on the Ewald circle *E* with radius $1/\lambda$. To do this the crystal is rotated by an angle ω and the detector is positioned at an angle Υ. In this case $\Upsilon = 2\Theta$, but $\omega \neq \Theta$ because the starting position differs from that of fig. 2*b*.

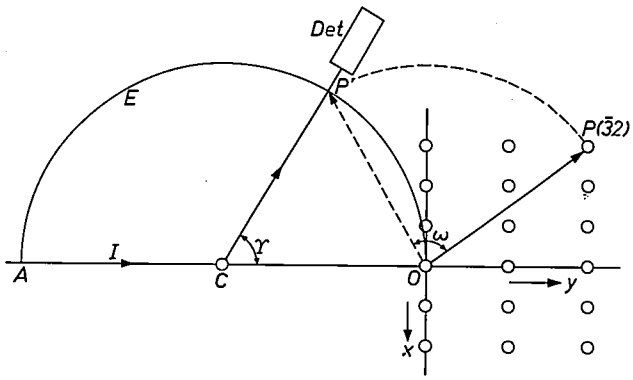


Fig. 6. The Ewald construction. With the aid of this construction one can determine the angles at which the crystal and the detector have to be placed with respect to the incident X-ray beam *AC* in order to bring the plane (*hk*) into reflection. The origin of reciprocal space is at point *O*, diametrically opposite point *A*. The diffracted beam is parallel to *CP'*. *E* Ewald circle.

The linear-tracking mechanism

A linear-tracking mechanism is used on the PAILRED diffractometer to make the angular settings of the crystal and the detector simultaneously [5]. This device is shown schematically in *fig. 7* and a comparison with fig. 6 shows that it performs the Ewald construction. The mechanism consists of an *x,y*-coordinate system which is free to rotate about its origin *O*. The crystal is mechanically coupled to this rotation so that the crystal follows the omega motion of the tracking device. By means of two lead-screws the nut *P* can be moved to the desired *x,y*-coordinates in reciprocal space. This causes a rotation of the coordinate system because the nut *P* is restricted to follow the circumference of the Ewald circle *E* by a link *CP*. This link, in turn, drives a shaft which determines the angular position Υ of the detector.

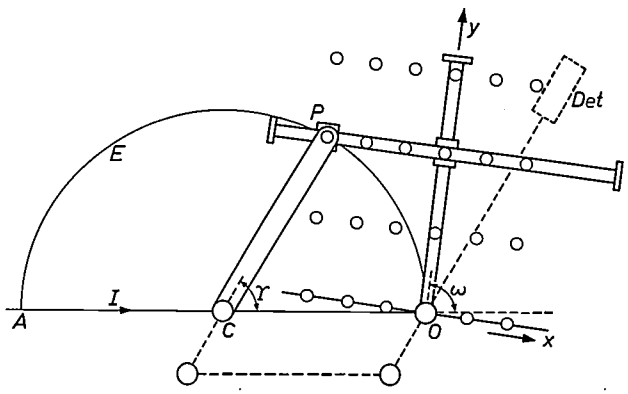


Fig. 7. The Ewald construction can be performed with a mechanical device constituting an *x,y*-coordinate system. The nut *P* slides along the *x*-axis, which in turn can slide in the *y*-direction. The nut can be positioned by means of motors and lead-screws to the desired *x*- and *y*-coordinates in the reciprocal *x, y*-plane. When *OP* corresponds to the reciprocal vector of a given crystal plane (*hk*) this plane is then in the reflecting position.

In order to make full use of the linear-tracking mechanism the reciprocal axes of the crystal must be aligned with the *x*- and *y*-directions of the mechanism on the diffractometer. This alignment is made manually by properly orienting the crystal and in the discussions it will be implicitly assumed that the crystal has been aligned.

Fig. 8 shows a photograph of the actual mechanism. The nut *P* is mounted on the *x*-axis assembly which moves up and down the *y*-axis as a unit. For increased accuracy the whole tracking mechanism is at all times statically balanced. This is achieved with a counterweight *CW* which moves along the *y*-axis at half speed in a direction opposite to the moving *x*-axis. For balance in all positions, the mass of the counterweight is twice that of the *x*-axis assembly. The link *CP* is curved to avoid mechanical interferences.

There are two important differences between the Ewald construction of fig. 7 and the actual situation of fig. 8. Because reciprocal space is a purely mathematical concept, it may be rotated and translated with respect to the incident beam. The only factor of importance is that the angles found in the Ewald construction are accurately repeated on the diffractometer. It has, therefore, no effect that the lines *CO* and *CP* are rotated by 60° with respect to the incident beam and the detector position. It should also be mentioned that figs. 7 and 8 show the linear-tracking system from opposite sides; in fig. 7 the X-ray source is on the left, in fig. 8 it is on the right.

[5] The same principle has been applied by U. W. Arndt and D. C. Phillips, The linear diffractometer, Acta cryst. **14**, 807-818, 1961.

The motors that drive the lead-screws corresponding to the x- and y-coordinates are stepping motors. By electronically counting the number of steps that are made, the central control unit keeps track of the x- and y-coordinates.

With this tracking mechanism the problem of finding the angular settings for the crystal and the detector has now been reduced to finding the coordinates of the desired reflection in the reciprocal plane where they are arranged in a repetitive pattern. Because the length of the link CP is fixed, the radius of the Ewald circle is also fixed on the diffractometer. Consequently, the size of the reciprocal cell on the x,y-coordinate system must be normalized with respect to the wavelength of the incident X-ray beam. The repeat distances in the x,y-plane, therefore, are not a^* and b^*, but $a^*\lambda$ and $b^*\lambda$. These normalized reciprocal-cell dimensions will be called A and B respectively.

The normalized quantities A and B are measured in reciprocal-lattice units (r.l.u.). The radius of the Ewald circle is by definition equal to one r.l.u. Because stepping motors are used on the linear-tracking mechanism the distance A on the diffractometer is divided into discrete steps. Each step of the stepping motors corresponds to 10^{-4} r.l.u. In the discussion of the control unit the quantities A and B will be expressed in steps.

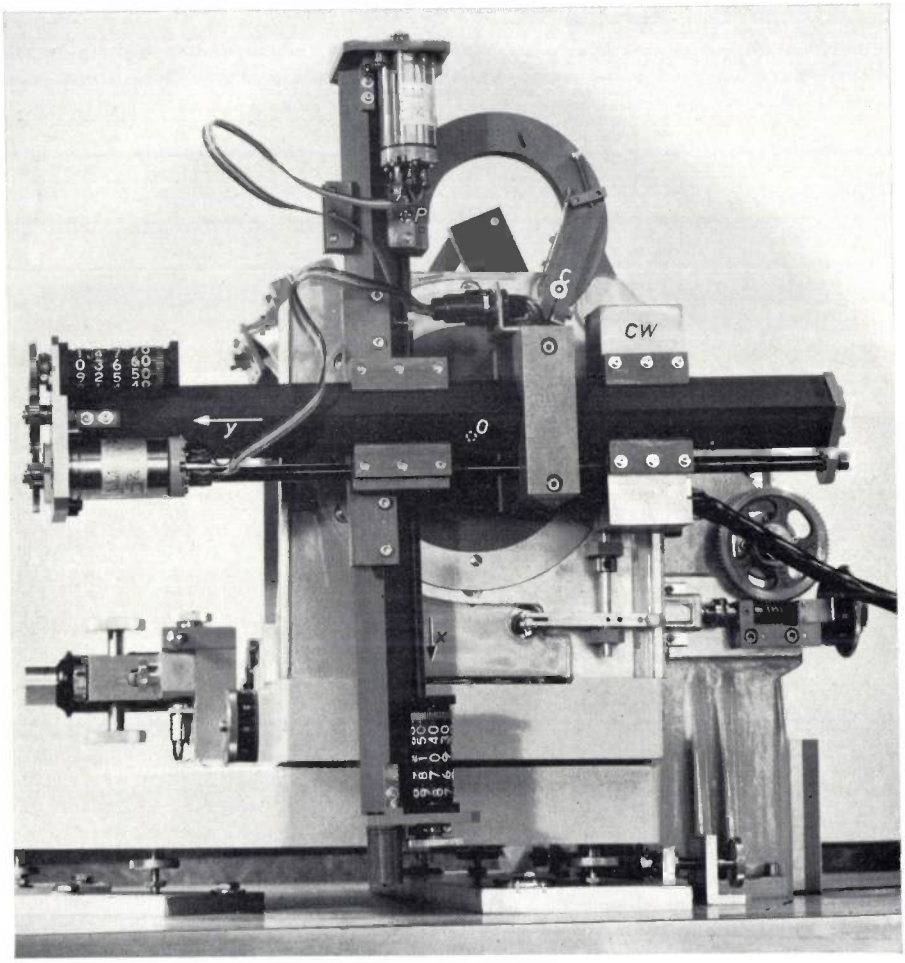


Fig. 8. The linear-tracking mechanism is mounted on the diffractometer. The mechanical Ewald construction has been translated and rotated with respect to the construction shown in figs. 6 and 7. *CW* counterweight. *O* origin. *CP* link.

Once the reciprocal-cell dimensions A and B have been determined, any desired crystal plane (hk) can be brought into reflection by adjusting the linear-tracking mechanism to the proper x,y-coordinates. For instance, in the case of an orthogonal net, the reflection from the (32) plane can be detected by adjusting the x,y-coordinates on the diffractometer such that $x = 3A$ and $y = 2B$. With this method, therefore, it is not necessary to perform complex angle calculations for measuring a given reflection. This fact makes the automation of the data-taking process and the manual operation of the instrument

relatively simple. It also explains why the instrument is called a *linear reciprocal-space exploring diffractometer*. The problem of the three-dimensional reciprocal lattice of a real crystal will be treated when the equi-inclination technique is discussed.

Automatic data collection

To be able to find the structure of a crystal, the crystallographer has to measure the reflectivity of a large number of crystal planes. This operation is performed automatically by the PAILRED system. One by one the different planes are brought into reflection by going from one reflection site on the x,y-coordinate system to the next. At each of these sites the linear-tracking mechanism is stopped and with the detector stationary, the crystal is scanned through a small angle to generate the diffraction profile for each reflection. This part of the measurement will be described in a subsequent section. First, we shall describe the manner in which the tracking device is moved from one reflection site to the next.

Fig. 9 shows an orthogonal reciprocal net similar to fig. 5. From figs. 6 and 7 it can be seen that the length of a reciprocal vector OP determines the angular Y position of the detector. In general, this angle should never exceed a certain maximum to avoid certain mechanical interferences.

There is also a theoretical limitation. When the vector OP is larger than the diameter of the Ewald circle (see fig. 6) it is no longer possible to rotate the crystal so that the point P lies on this circle. In other words, the largest possible detector angle Y is 180°. For the PAILRED system this angle Y should not exceed 150°. Or, stated another way, the largest reciprocal vector that can be measured has a length of approximately 1.9 r.l.u.

The reciprocal plane is therefore limited by a circle. To facilitate automation this limit is sensed with a switch connected to the detector shaft, the counter-high switch. Because of mechanical interference it is also impossible to measure at negative values of y. This limit is sensed by a switch in the linear-tracking system (fig. 8), the y-low switch. Hence, only a semicircular part of the reciprocal plane can be measured. Sometimes it is also necessary to measure the reflections of the other half of this plane. This is done after the crystal has been rotated by 180° with respect to the tracking mechanism.

During automatic data collection reflections are measured in the order in which they occur along lines parallel to the x-axis. After all reflections on one line have been measured, the line below it is scanned. The automation of this process is simplified because of the regular pattern of reflection sites throughout reciprocal space.

Having measured reflection r_1 (fig. 9), the automatic control unit moves to the next position r_2 in the x-direction by delivering a number $A = 10^4\, a^*\lambda$ of stepping pulses to the motor that controls the x-motion. These pulses are counted electronically and after A pulses, the counter is reset to zero and the next reflection r_2 can be measured.

When the end of the line is reached, the switch that defines the semicircle closes. The x-motion is stopped and the number of steps that was taken in the x-direction after the last reflection position is retained in the x-incremental counter. Using a second incremental counter, a distance $B = 10^4\, b^*\lambda$ is measured off in the negative y-direction. After this y-motion has been completed, the x-motion continues in the same direction as before. If during this motion the number A is reached (at r_3), the x-incremental counter is reset and, without measuring a reflection, the x-motion goes on. This process continues until, again, the semicircle is reached. At that moment the x-motor and the x-incremental counter are reversed until the counter reaches the number zero. In this manner it is always possible to find that reflection site (r_3) on the new line which is closest to the semicircle. Starting with r_3, all reflections on this line are now measured consecutively.

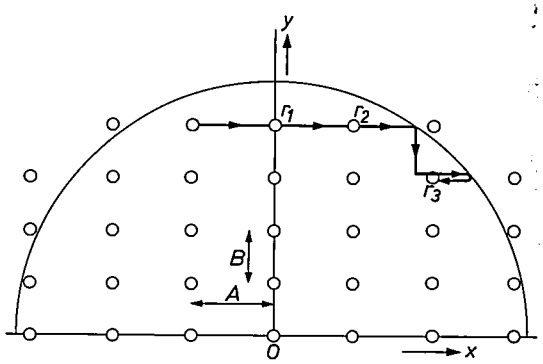


Fig. 9. During automatic data collection the position of the nut (P in fig. 7) of the linear-tracking mechanism is moved from one reflection site to the next. Reflections are measured in the order in which they occur in reciprocal space in a direction parallel to the x-axis. After all reflections on one line have been measured, those on the line below it are measured one by one and so on. The semicircle corresponds to the maximum detector angle.

Because not all crystals have orthogonal cell edges, provisions must be made for crystals whose reciprocal net is oblique. Such a net is shown in *fig. 10*. In this case the y-axis of the instrument does not coincide with a row of reciprocal-lattice points. By defining two additional quantities C and D such that $C + D = A$, this net is uniquely defined. The automatic motion along x from one reflection site to the next is the same

as for orthogonal nets. However, after each y-motion to the next lower line, the reflection sites are displaced in the x-direction by an amount C or D and these constants have to be used to find the outermost reflection.

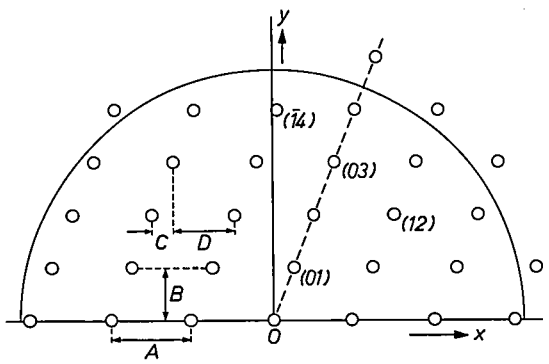


Fig. 10. The reciprocal cell in an *oblique* net can be defined by the parameters A, B, C and D where $C + D = A$. For an orthogonal net only the parameters A and B are required.

One difficulty with the tracking mechanism arises when the point P approaches the origin O of the x,y-plane. An examination of fig. 7 will show that when the point P is allowed to coincide with O, the mechanical x,y-coordinate system is free to rotate about the omega axis. When the distance OP is very small, excessive stresses on the tracking mechanism are necessary to rotate the omega axis. Therefore, protective devices have been incorporated in the diffractometer to ensure that the distance OP does not become too small. This is equivalent to stating that the detector angle Y cannot be less than a certain minimum, which is 4 to 6 degrees for the PAILRED diffractometer. Special circuits have been provided to automatically move around the origin at a safe distance during automatic data collection.

Because the lines are measured in the order of decreasing y, the automatic cycle is started at the maximum value of y. It ends when y becomes negative.

Upper levels and the equi-inclination technique

Having seen that the reciprocal net for a two-dimensional crystal becomes a repetitive two-dimensional pattern, it is evident that the reciprocal lattice for a three-dimensional crystal will be a three-dimensional repetitive pattern consisting of identical two-dimensional nets stacked one above the other to form a three-dimensional array. Up to now only the plane containing the origin of reciprocal space has been discussed. This plane is called the zero level or equatorial level; the other planes of this stack are called upper levels.

The Ewald construction is, in principle, the same as before except that the circle now becomes the (surface of the) Ewald *sphere*. The origin of the three-dimensional reciprocal space is located at the point O where the incident beam leaves the sphere. A third index (l) is added to the Miller indices, and to bring the plane (hkl) into reflection its reciprocal point must be brought on the Ewald sphere and the detector must point from this point to the center of the sphere.

The position of the upper levels of the crystal is shown in *fig. 11a*. The circle represents the Ewald sphere and the Ewald construction of fig. 6 is in a plane perpendicular to the paper. The line AO is the projection of the plane of fig. 6 and rotation around the omega axis OB makes the end points of the reciprocal vectors of the ($hk0$) planes, belonging to the zero level, coincide with the Ewald sphere in this plane.

Fig. 11*a* also shows the reciprocal planes ($hk1$) and ($hk2$) which are parallel to the ($hk0$) plane and perpendicular to the plane of the paper. Point B is the origin of the first level containing the reciprocal points $hk1$. In orthogonal lattices OB is the reciprocal vector of the (001) plane; in oblique lattices all points of the first level are displaced by the same amount with respect to the zero level, and those of the second level by twice this amount, and so on.

The most convenient way to bring the upper levels into reflection is the equi-inclination method. This method has been chosen for automatic measurement on the PAILRED diffractometer. For instance, to bring the planes ($hk1$) into reflection, the upper part of the diffractometer with the omega axis OB is rotated, or inclined, by an angular amount μ so that point B is brought on the Ewald sphere. This rotation results in the configuration shown in fig. 11*b*. Subsequent rotation of the crystal around the omega axis will then bring the reciprocal points in the ($hk1$) plane on the Ewald sphere along a circle perpendicular to the paper and whose diameter is AB. This circle now becomes the circle of the construction shown in fig. 6. The center of the circle is the projection of the center of the Ewald sphere (C) on to the plane AB and the Ewald construction (fig. 6) takes place in this plane. It should be noted that although the radius of the Ewald sphere is $1/\lambda$, the radius of the circle with diameter AB equals $(\cos \mu)/\lambda$. Because the radius of the Ewald circle on the goniometer is mechanically fixed, the repeat distances in the x,y-coordinate system are no longer equal to $a^*\lambda$ and $b^*\lambda$, but are now $A = a^*\lambda/\cos \mu$ and $B = b^*\lambda/\cos \mu$.

The detector must also be positioned to follow the intersection of the ($hk1$) plane with the sphere in such a way that it is directed at the center of the sphere. In the original position of the diffractometer for which

$\mu = 0$ (fig. 11*a*), the detector followed a circle perpendicular to the omega axis of the diffractometer. When the diffractometer is inclined by an amount μ, this becomes the circle whose radius is CD. To bring the detector in the proper position to measure the ($hk1$) reflections it must also be inclined by an amount $\nu = \mu$. The axis of the detector will then lie on the surface of a cone whose apex is the center (C) of the Ewald sphere and which intersects this sphere along a circle with diameter AB and perpendicular to the plane of the paper. Because $\nu = \mu$, this technique is called the equi-inclination technique.

In the design of PAILRED the equi-inclination technique of measuring upper levels has been preferred to others such as the normal-beam technique, in which only ν is changed and μ is kept equal to zero, because the equi-inclination technique allows a greater part of reciprocal space to be measured without re-aligning the crystal.

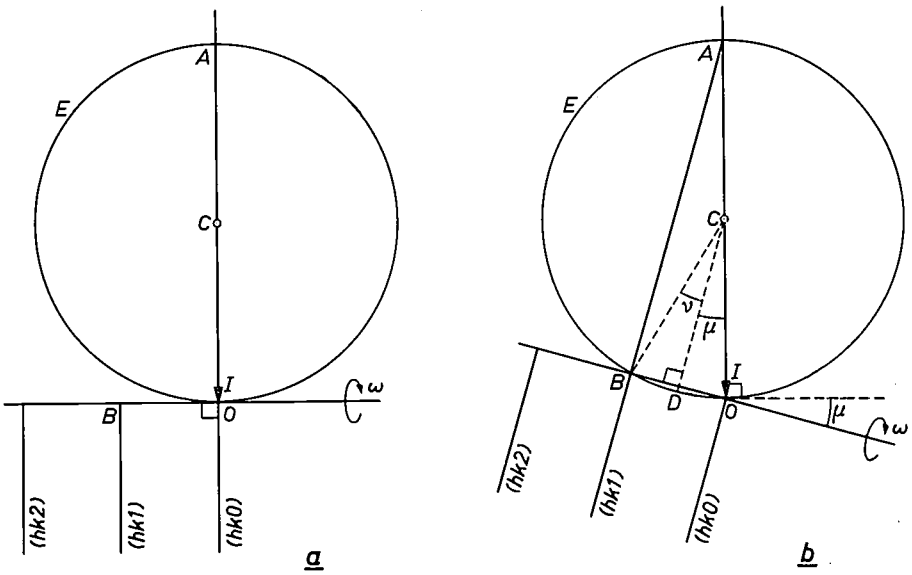


Fig. 11. A three-dimensional reciprocal lattice. *a*) A top view of the Ewald construction shows the position of the diffractometer for measuring the zero level. The circle with diameter AO shown in fig. 6 is the intersection of the Ewald sphere E with a plane perpendicular to the paper. All reciprocal vectors ($hk0$) also lie in this plane and by rotation about the omega axis OB their end points can be brought on to the Ewald sphere. I incident beam.
b) The equi-inclination geometry. To measure the first upper level, the crystal, and therefore its reciprocal net also, is inclined by an angle μ so that the end point B of the vector OB lies on the Ewald sphere E. Point B is the center of rotation of the first level. The detector is also inclined by an angle $\nu = \mu$ and the Ewald construction of fig. 6 again takes place in a plane perpendicular to the paper, but the diameter of the Ewald circle is now AB and the reciprocal-lattice vectors OP are replaced by their projections BP on to this plane. The origin O of reciprocal space is always located diametrically opposite point A.

In *fig. 12* a photograph is shown of the diffractometer in an equi-inclination position. The movable upper base on which the diffractometer is mounted has been rotated forward around the vertical μ-axis with respect to the stationary lower base of the instrument. The detector has been moved along its supporting arc by the same angular amount. The crystal is mounted at the center of the instrument (the point where the ω-axis and μ-axis intersect) on a large Eulerian cradle which provides the necessary degrees of rotational freedom to allow the crystal axes to be aligned with the axes of the instrument. This alignment is necessary to bring the reciprocal net into the proper orientation for the equi-inclination technique.

Omega rotation

With the omega drive the crystal can be rotated about the omega axis with respect to the linear-tracking mechanism. This drive, again, uses a stepping motor so that the amount of rotation can be determined by counting motor pulses electronically. The speed of rotation can be controlled by selecting the repetition rate of the motor pulses. The omega drive is located between the linear-tracking device and the crystal and can be seen in fig. 12.

The effect of this rotation can be understood by studying fig. 6. When the tracking mechanism is stationary, the detector is stationary at point P'. The finite aperture of the detector corresponds to a small region of the surface of the Ewald sphere around P'. Rotation of the crystal about the omega axis will cause the reciprocal net to rotate with respect to the tracking device and the Ewald sphere. A reflection is observed if a reciprocal-lattice point passes the surface of the Ewald sphere within a small spherical cap which is

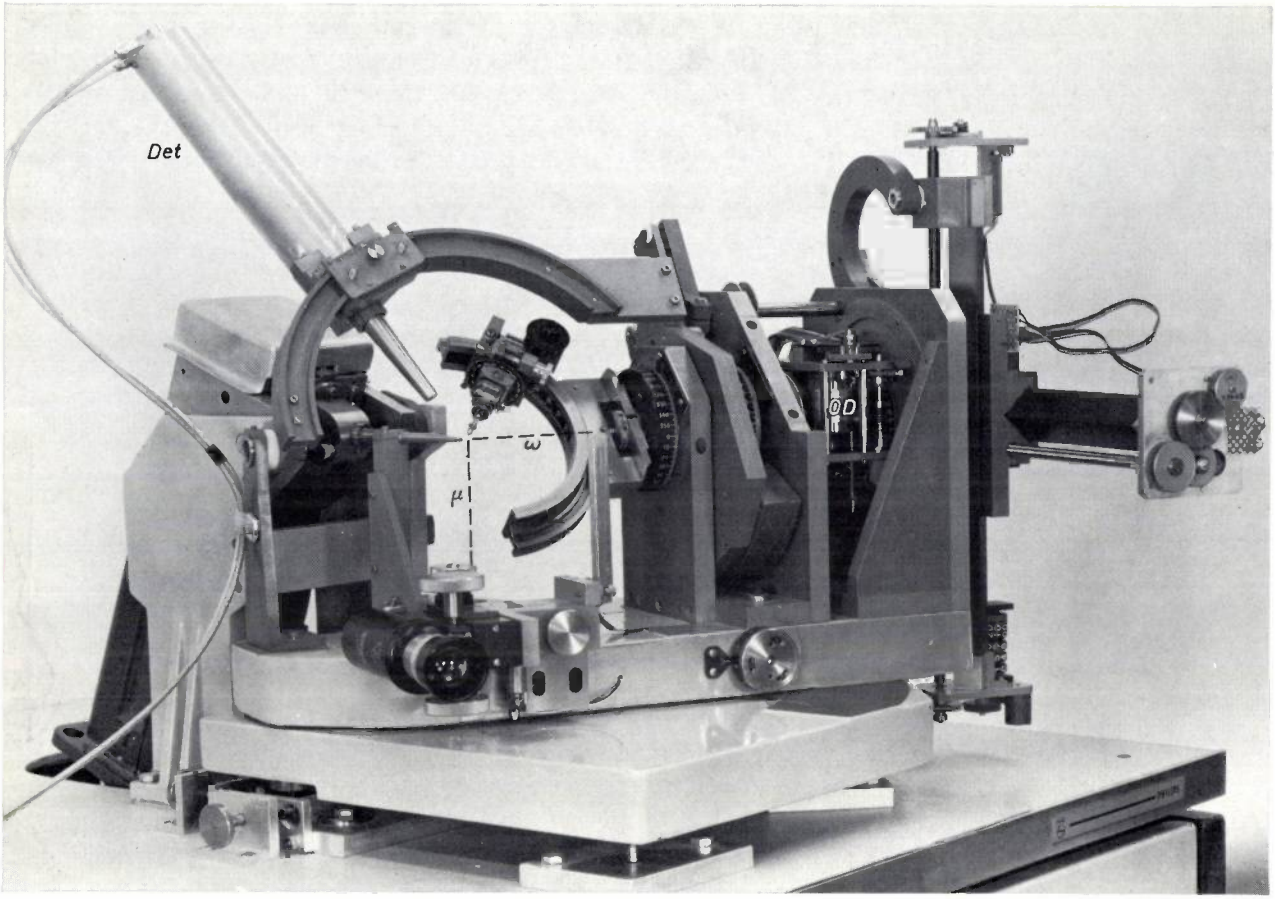


Fig. 12. The diffractometer is shown in the equi-inclination position. The upper base, on which the diffractometer is mounted, has been inclined by an angle μ with respect to the stationary lower base. The detector *Det* has been moved along its supporting arc by an angle $\nu (= \mu)$. *OD* omega drive.

subtended by the detector aperture. Alternatively, one may say that a three-dimensional part of reciprocal space is swept by this cap on the Ewald sphere. The reciprocal-lattice point of the reflection to be measured should traverse the center of the spherical cap.

When the crystal is truly perfect and the incident beam is parallel and ideally monochromatic, i.e., only radiation of one wavelength λ is present, there will be only a very narrow omega region at which a given set of planes can reflect the X-ray beam. This ideal condition, however, never occurs in a practical case. The incident beam, even when crystal monochromatization is used, as in the PAILRED system, will always consist of a narrow band of wavelengths centered about λ. Crystals, also, are never (ideally) perfect. These and other imperfections may be translated into a blurring of the reciprocal lattice points. A small omega rotation, the *omega scan*, is then necessary to make the entire "point" pass the sphere.

Fig. 13 shows a typical plot of intensity versus omega angle that is obtained on scanning through a reflection. In the zero level the width of the peak varies from 0.02 to 0.2 degrees, or more, depending on crystal perfection. During automatic data collection the vicinity of each reflection is probed with an omega scan, where the angular distance from Q to R is usually of the order of 1 to 2 degrees. In upper

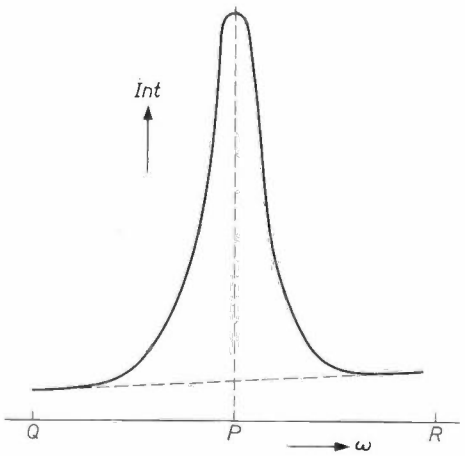


Fig. 13. When the crystal is rotated by a small angle (from Q to R) about the omega axis, the intensity *Int* of the reflected X-ray beam shows a peak at point P which corresponds to the omega position that was determined from the Ewald construction (fig. 6). The integral of the reflection profile, after subtraction of the background radiation, is called the integrated intensity.

levels, especially at high inclination and low values of Y, the profiles may become so broad that the scan range has to be increased. All quanta received during the scan are counted in the scaler. The total count is proportional to the area under the diffraction profile. This "integrated intensity" is used as the measure of reflectivity for that particular set of planes in the crystal [6].

A "background" correction is normally made to achieve a more accurate value for the integrated intensity. The intensity measured at the extremes (Q and R) of the scanning range is assumed to be caused by X-ray scattering other than from the set of crystal planes being measured. With the crystal stationary at Q and R a fixed-time measurement of this background is made and the area under the dashed line in fig. 13 is subtracted from the total area measured when the data are processed. The area above the dashed line is then the corrected integrated intensity.

The sequence of events during automatic data collection can then be summarized as follows. After the linear-tracking mechanism has positioned the x- and y-coordinates of a reflection site, it remains stationary while an omega scan is being made. The omega scan comprises the following operations:

a) The machine indices H, K and L of the plane being measured are printed out; these indices can easily be transformed into the Miller indices h, k and l. The two sets of indices are identical when the c-axis is aligned parallel to the omega axis and the a^*-axis parallel to the x-axis and, moreover, the b^*-axis is either parallel to the y-axis or lying in the positive quadrant of the x,y-plane.

b) The omega drive rotates the crystal from the center position P to one extreme position Q and a fixed-time background measurement is made while the omega drive remains stationary. The result of this measurement is printed out.

c) An omega scan is made at constant velocity from Q to R. During this scan the integrated intensity is measured and subsequently printed out.

d) At position R another fixed-time background measurement is made and the result is printed out.

e) Before the linear-tracking mechanism moves to the next reflection, the omega drive is returned to the center (P) of its range.

Monochromator assembly

Strictly monochromatic X-ray sources of sufficient intensity are not available. All available X-ray tubes yield a spectrum consisting of a number of narrow lines and a generally much weaker continuum. For diffractometry one of the strong lines is selected, usually the Kα line of copper or molybdenum, which is in fact a doublet. The presence of radiation with different wavelengths not only produces a high background, but may cause other errors too. The unwanted lines and part of the continuum may be removed by the use of a proper filter, but even then the tails of the reflection profile are very long. This broadening of reflections affects the ability to resolve and independently measure adjacent reflections.

The most effective means of restricting the band of wavelengths present in the incident beam, the crystal monochromator, is used in PAILRED. Its operation is based on the fact that X-rays reflected by a single crystal must satisfy the Bragg equation (eq. 1). The focus of the X-ray tube and the monochromator crystal are positioned in such a way that radiation of the selected spectral line is reflected towards the crystal under study. Because neither this crystal nor the X-ray focus are infinitely small, there will be a finite spectral band pass, e.g. $\Delta\lambda/\lambda = 0.02$.

Fig. 14 shows the basic configuration of the monochromator assembly. A flat crystal plate, cut parallel to a strongly reflecting plane, is mounted on the X-ray tube housing in such a way that it can be rotated

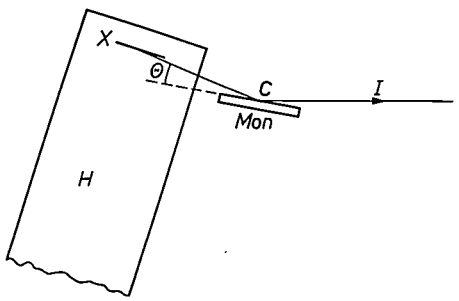


Fig. 14. The incident X-ray beam is monochromatized with the monochromator crystal *Mon* which is attached to the X-ray tube housing H. When this crystal is positioned at the proper angle Θ, its center will reflect X-rays having a certain wavelength as determined by the Bragg equation.The reflected monochromatized beam I can be brought into the horizontal plane by rotating the tube housing and the attached monochromator assembly about the axis C, perpendicular to the paper. X line focus of X-ray tube.

about the axis C (perpendicular to the paper) and therefore with respect to the beam emerging from the X-ray tube. Positioned at the proper angle Θ, the monochromator crystal will reflect radiation of the selected wavelength.

After this position has been found, the X-ray tube and the crystal together can be rotated about the same axis of rotation C to bring the reflected (and monochromatic) beam I into the horizontal plane so that it strikes the crystal which is mounted at the center

[6] The difficulty of the definition of integrated intensity when a crystal monochromator is used has been discussed by J. Ladell and N. Spielberg, Theory of the measurement of integrated intensities obtained with single-crystal counter diffractometers, Acta cryst. **21**, 103-118, 1966.

of the diffractometer.

Fig. 15 shows a photograph of the X-ray tube housing with monochromator.

The use of crystal-monochromatized radiation leads to special requirements for the mechanical stability of the diffractometer and the shape of the crystal to be studied. We shall discuss the case of one single line first. The reflection of a monochromatic beam by a perfect flat monochromator crystal gives rise to a diffracted beam which is very narrow in the vertical direction at the center of the instrument. This can be seen in fig. 14 by drawing lines strictly parallel to the central ray. The width of the beam is equal to the length of the focal line times the sine of the take-off angle, the angle at which the radiation to be used leaves the focus. By varying the take-off angle the width of the beam can be varied between 0.3 and 0.8 mm. Because the intensity of the emitted X-rays may vary along the focus, the beam may be inhomogeneous in the vertical direction at the center of the instrument. The situation is still worse if a doublet is used; but continuum radiation does not do any harm.

The inhomogeneity of the beam causes errors in the intensity ratio of different reflections if the crystal being studied does not have circular symmetry about the omega axis or moves up and down on rotation about this axis. To exclude the latter, PAILRED offers a high stability and a facility for centering the crystal with an error of less than 10 μm in the vertical direction. This stability could be achieved because it depends on the precision of only one set of bearings, those of the omega axis. Experience has shown that reliable intensity measurements can also be made with irregularly shaped crystals, provided they are sufficiently small.

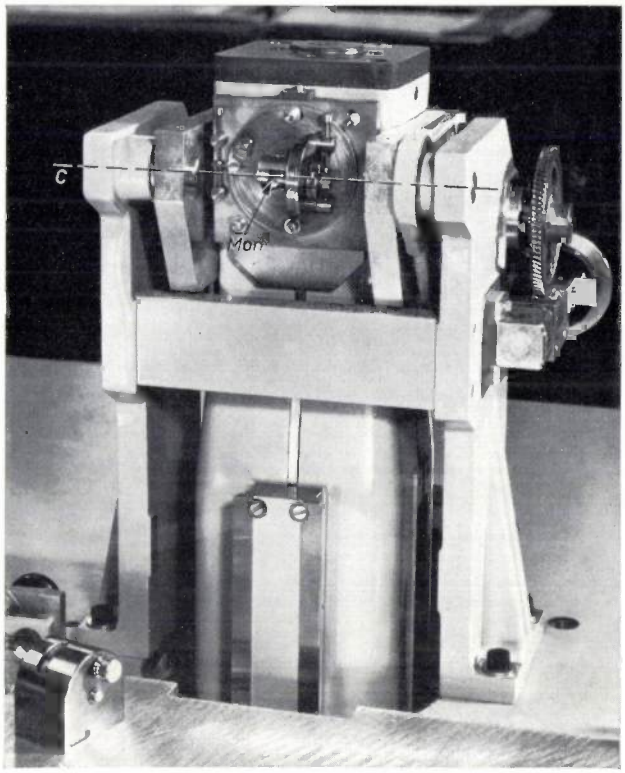


Fig. 15. The X-ray tube housing and the monochromator assembly. The combination can be rotated about the axis *C*. *Mon* monochromator crystal.

II. The automatic control circuits

In the description of the diffractometer it was shown how the positioning of the crystal and the detector to measure the reflection from a certain (*hkl*) plane was reduced to finding the correct *x*,*y*-position in reciprocal space. For the automatic collection of data, the diffractometer moves from one reflection site to the next and at each site an omega scan is performed to measure the integrated intensity from one (*hkl*) plane at a time. The automatic actions of the diffractometer, then, form a fixed program of sequential events. Part II of this paper will describe the principles and some details of the electronic circuits that are used to generate these sequences [7]. In general, the same principles can be applied to any automated sequential fixed-program device.

All control functions are generated using digital techniques and for this reason, one could classify the control unit as a fixed-program special-purpose digital computer. The term fixed program should not be interpreted to mean that experimental parameters cannot be selected to suit the particular problem under investigation. For instance, the parameters of the reciprocal cell *A*, *B*, *C* and *D* (see fig. 10) can be adjusted by means of switches to fit the crystal under investigation. Likewise, the angular range of the omega scan (from *Q* to *R* in fig. 13) and the scanning speed can be selected by means of switches that are located on the control panel.

Although the major part of the control circuits are used to generate the sequences necessary for automatic data collection, some additional circuits are used to make manual or semi-automatic operation of the diffractometer possible. This is especially convenient for the alignment of the crystal axes with the axes of the diffractometer. This function can be regarded as a second fixed program: the manual program.

Main sequences

Before discussing the details of these control circuits, a survey will be given of the main sequences controlling the automatic operation (*fig. 16*). The

[7] P. G. Cath, Operational modes and controls of the PAILRED automatic single-crystal diffractometer, Norelco Reporter **12**, 57-62, 1965.

of the diffractometer.

Fig. 15 shows a photograph of the X-ray tube housing with monochromator.

The use of crystal-monochromatized radiation leads to special requirements for the mechanical stability of the diffractometer and the shape of the crystal to be studied. We shall discuss the case of one single line first. The reflection of a monochromatic beam by a perfect flat monochromator crystal gives rise to a diffracted beam which is very narrow in the vertical direction at the center of the instrument. This can be seen in fig. 14 by drawing lines strictly parallel to the central ray. The width of the beam is equal to the length of the focal line times the sine of the take-off angle, the angle at which the radiation to be used leaves the focus. By varying the take-off angle the width of the beam can be varied between 0.3 and 0.8 mm. Because the intensity of the emitted X-rays may vary along the focus, the beam may be inhomogeneous in the vertical direction at the center of the instrument. The situation is still worse if a doublet is used; but continuum radiation does not do any harm.

The inhomogeneity of the beam causes errors in the intensity ratio of different reflections if the crystal being studied does not have circular symmetry about the omega axis or moves up and down on rotation about this axis. To exclude the latter, PAILRED offers a high stability and a facility for centering the crystal with an error of less than 10 μm in the vertical direction. This stability could be achieved because it depends on the precision of only one set of bearings, those of the omega axis. Experience has shown that reliable intensity measurements can also be made with irregularly shaped crystals, provided they are sufficiently small.

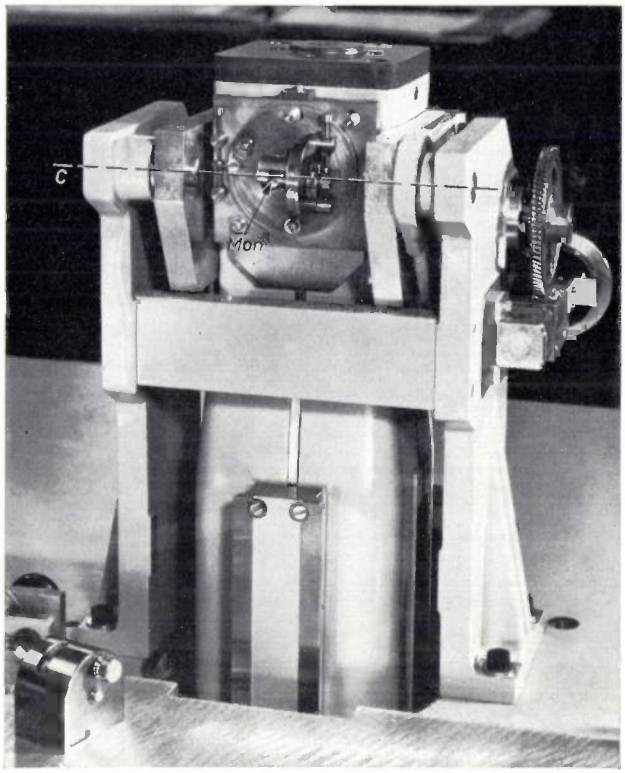


Fig. 15. The X-ray tube housing and the monochromator assembly. The combination can be rotated about the axis *C*. *Mon* monochromator crystal.

II. The automatic control circuits

In the description of the diffractometer it was shown how the positioning of the crystal and the detector to measure the reflection from a certain (*hkl*) plane was reduced to finding the correct *x*,*y*-position in reciprocal space. For the automatic collection of data, the diffractometer moves from one reflection site to the next and at each site an omega scan is performed to measure the integrated intensity from one (*hkl*) plane at a time. The automatic actions of the diffractometer, then, form a fixed program of sequential events. Part II of this paper will describe the principles and some details of the electronic circuits that are used to generate these sequences [7]. In general, the same principles can be applied to any automated sequential fixed-program device.

All control functions are generated using digital techniques and for this reason, one could classify the control unit as a fixed-program special-purpose digital computer. The term fixed program should not be interpreted to mean that experimental parameters cannot be selected to suit the particular problem under investigation. For instance, the parameters of the reciprocal cell *A*, *B*, *C* and *D* (see fig. 10) can be adjusted by means of switches to fit the crystal under investigation. Likewise, the angular range of the omega scan (from *Q* to *R* in fig. 13) and the scanning speed can be selected by means of switches that are located on the control panel.

Although the major part of the control circuits are used to generate the sequences necessary for automatic data collection, some additional circuits are used to make manual or semi-automatic operation of the diffractometer possible. This is especially convenient for the alignment of the crystal axes with the axes of the diffractometer. This function can be regarded as a second fixed program: the manual program.

Main sequences

Before discussing the details of these control circuits, a survey will be given of the main sequences controlling the automatic operation (*fig. 16*). The

[7] P. G. Cath, Operational modes and controls of the PAILRED automatic single-crystal diffractometer, Norelco Reporter **12**, 57-62, 1965.

two main parts are the omega cycle and the x-advance cycle. In the omega cycle, to be discussed in more detail later on, background and intensity of a reflection are measured and the results are printed and punched preceded by the indices H, K and L. The x-advance cycle controls the transfer from one reflection to the next in the way shown in fig. 9. Within one row of the reciprocal lattice only the x-coordinate is changed, each time by presenting A pulses to the x-motor.

When the end of a row is reached the x-advance cycle is interrupted by a signal from the counter-high switch *CH*. This signal will stop the x-motor (or also the y-motor in the case of manual operation) and only allow it to be started in a direction away from the limiting semicircle. The counter-high signal also starts the edge cycle which brings the linear-tracking system to the next row of the reciprocal lattice (fig. 9). On this row the edge cycle then searches for the outermost reflection as discussed above. When the offsets C and D (fig. 10) are needed for oblique lattices, circuitry of the x-advance cycle interacts with the edge-cycle control. Although several reciprocal-lattice points may be passed during this search procedure, the omega cycle is not started until the final point has been found. During automatic operation, the central control unit gives the correct indices H and K to every reflection; this explains the term *automatic-indexing* in the name "PAILRED".

Another interruption of the x-advance cycle occurs in the neighborhood of the origin $x = 0$, $y = 0$, because at this point the position of the linear-tracking system would become indeterminate. This interruption starts when $y < 0.1$ and $|x| = 0.1$ r.l.u. Then y is increased by 0.1 and, after x has been changed by 0.2, decreased again by 0.1.

Basic circuits

Several electronic circuits have to be designed in order to control the three motions (x, y and ω) on the diffractometer and the data-collection and print-out operations. Counting circuits are needed to count the number of steps taken by the stepping motors, as we noted in part I. Frequency dividers are needed to generate accurately controlled frequencies to regulate the motor speeds. Sequence counters are needed to generate the different parts of the automatic sequences. Finally, logic circuits are needed to decide when one

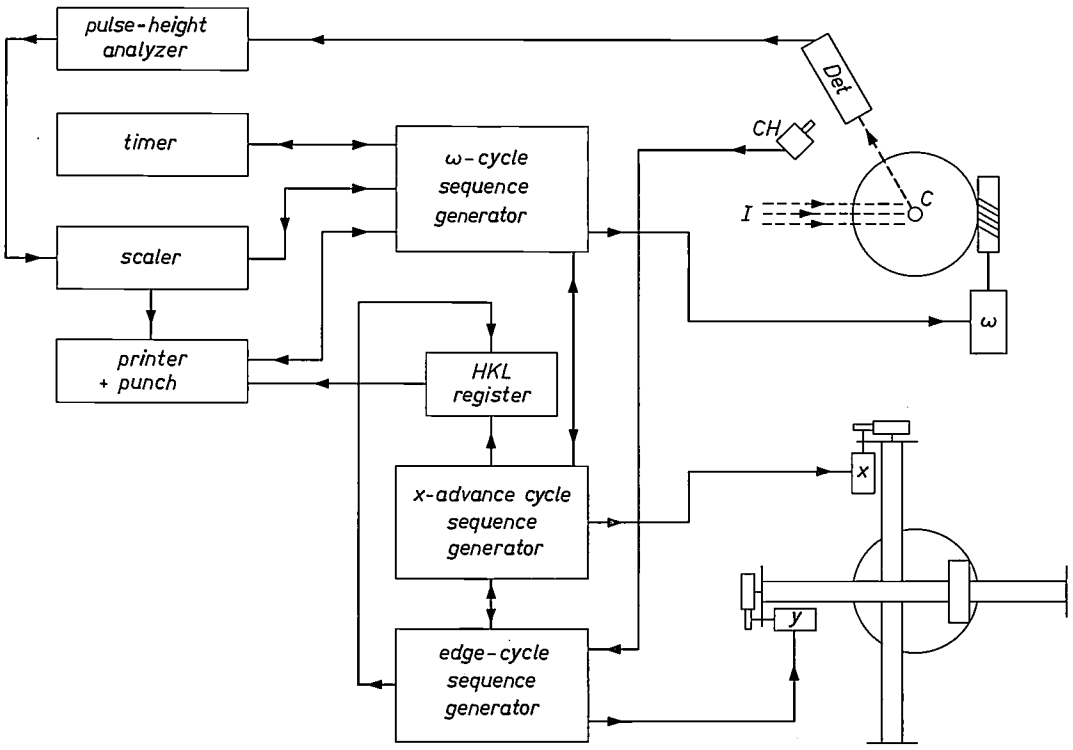


Fig. 16. Simplified diagram of the automatic cycle. On the right the diffractometer is shown twice, from two sides (cf. figs. 8 and 12), on the left the radiation-measuring circuits and the output devices. The omega cycle, the x-advance cycle, and the edge cycle are the main control sequences. In the omega cycle the indices *HKL* of each reflection are printed and punched and the integrated intensity is measured. During this measurement the crystal C is slowly rotated by the motor ω and the quanta of diffracted radiation received by the detector *Det* are counted (*scaler*). Before and after this omega scan, background is measured for a pre-set time (*timer*). After completion of this cycle the x-coordinate is advanced to the next reflection, where the omega cycle is started again. At the end of a line the x-advance cycle is interrupted by a signal from the counter-high switch *CH* activated by the detector. This starts the edge cycle. In this cycle the y-coordinate is decreased by B, the index K by one and the outermost reflection of the next line is found. Then the omega cycle is started again. The automatic cycle ends when y becomes negative.

part of a sequence has been completed and the next one can be started.

All these functions can be created with four basic building blocks (*fig. 17*):

a) A memory circuit which has two stable states or positions, the bistable circuit or flip-flop (fig. 17*a*).
b) A pulse gate which is used to change the state of the memory circuit by transmitting negative pulses when a certain condition is fulfilled (fig. 17*b*). Pulses may be obtained directly from the central pulse generator (clock pulses) or after frequency division (motor pulses).
c) A logic inversion circuit, the inverter amplifier (fig. 17*c*).
d) Two types of logic gates, the AND gate (fig. 17*d*) and the OR gate (fig. 17*e*).

Because these units are well known [8][9], they will not be described in detail. Their function is described in the caption of fig. 17; the symbols are given as they will be used in this article.

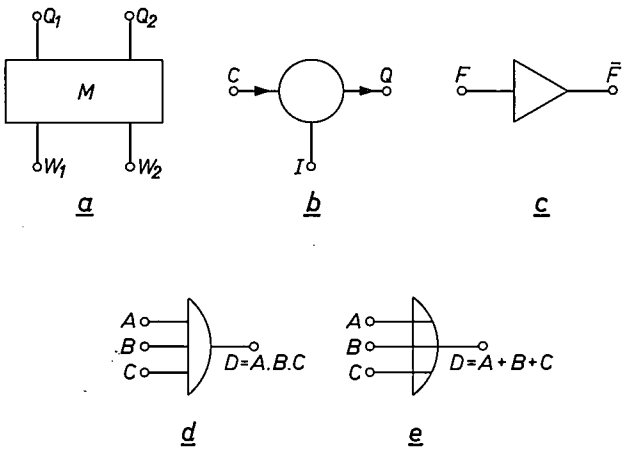


Fig. 17. *a*) Bistable circuit or flip-flop. Of the two outputs Q_1 and Q_2 one is at ground potential ("1") and one at +12 volts ("0"). If Q_1 = "1" and Q_2 = "0" the bistable circuit is said to be in the "1"-state, M = "1" and $\overline{M}$ (not-M) = "0". It can be brought to the "0"-state M = "0" by a negative pulse at input W_1. Because the state of the bistable circuit corresponds to the state of output Q_1, we may say $M = Q_1$ and, similarly, $\overline{M} = Q_2$.
b) Pulse gate. When the control C is at +12 volts (C = "0"), negative pulses at input I will not appear at the output Q, but they will if C is at ground potential (C = "1").
c) The circuit for logic inversion. The state of the output ($\overline{F}$ = not-F) is always the opposite of the state at the input.
d) AND gate. D = "1" when $A = B = C$ = "1", otherwise D = "0".
e) OR gate. D = "0" when $A = B = C$ = "0", D = "1" when at least one of the inputs is in state "1".

A simple combination of these building blocks is the start-stop control. It is desired to start the motion of a stepping motor upon the momentary closure of a push-button switch (*Start*) and to halt this motion when a second push-button switch is closed (*Stop*). Because only momentary closures are used for the switches, a memory element, a bistable circuit, is needed for this control function. The basic circuit for achieving this function is shown in *fig. 18*. As long as the switches are open, the inputs to the pulse gates G_1 and G_2 are at a positive potential and, therefore, the gates are closed and the state of the bistable circuit *SS* remains unchanged because no clock pulses *cp* are transmitted to its inputs W_1 or W_2. Assuming that the bistable circuit is initially in the "0" state, its output Q_2 will be at ground potential and its output Q_1 at +12 volts. In this state pulse gate G_3 is closed and no motor pulses *mp* are applied to the stepping motor *Mot*. When the *Start* button is momentarily depressed, pulse gate G_1 will open and (negative) clock pulses *cp* will be applied to input W_2. This will bring the bistable circuit in the "1" state. Pulse gate G_3 will be opened and motor pulses are applied to the stepping motor as long as the bistable circuit remains in this state. Upon operation of the *Stop* switch, this process is reversed, the bistable circuit is returned to the "0" state and the stepping motor will again stop. More than one clock pulse may be applied to the inputs upon operation of the push-button switches, but, after the first pulse has changed the state of the bistable circuit, subsequent pulses have no further effect on the state of this control element.

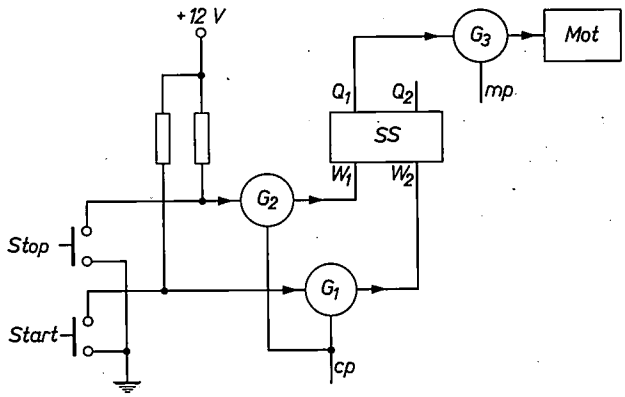
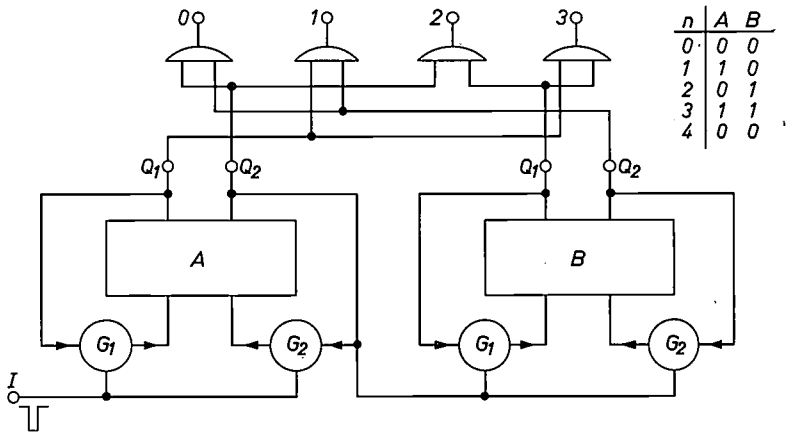


Fig. 18. Motion of a stepping motor *Mot* can be controlled by applying motor pulses *mp*. Pulse gate G_3 is controlled by a bistable circuit *SS* whose state can be changed with *Start* and *Stop* push-button switches. Clock pulses are denoted by *cp*.

Counters

With the basic building blocks of fig. 17 it is possible to generate all the circuit functions that are needed for the automatic sequential control unit. The most widely used function is that of electronic counting. Counters are used to count the number of pulses that are applied to the stepping motors so that the amount of motion can be determined and controlled. Counters are also used to generate the steps of the automatic control sequences.

Fig. 19 shows a counter with two stages A and B. Each stage consists of a bistable circuit and two pulse gates. With the bistable circuit initially in the "0" state, output Q_1 will be at +12 volts and output Q_2 at ground potential. The pulse gates G_1 and G_2 will then be closed and open respectively. A pulse applied to the pulse input I will pass through pulse gate G_2 of stage A and change its state to "1". A second pulse will change it back to "0", etc. With every other input pulse, when the bistable circuit makes the transition from "1" to "0", the voltage at output Q_2 of A will swing from +12 volts to ground. This negative-going transition is used to supply the pulse input signal to the second counting stage B. The table in fig. 19 shows the states of A and B after a sequence of input pulses. Four states are possible and after every fourth input pulse this system of two stages will return to the same position. The four possible states can be sensed with a decoding network consisting of AND gates as shown in fig. 19. Use is made here of the fact that the second output of each bistable circuit presents the inverse of the signal at its first output. Because the circuit can be used to determine the number of input pulses that have occurred after the initial "00" state, it can be used for counting.



n	A	B
0	0	0
1	1	0
2	0	1
3	1	1
4	0	0

Fig. 19. Two bistable circuits A and B are connected as a counting circuit with four possible states. The table shows the states of the two bistable circuits after n input pulses. I pulse input.

Just as the system of two stages has four stable states, a counter consisting of n stages has 2^n stable states. Although such "binary" counters make the most economical use of bistable circuits, it is often more convenient to count using a decimal system. A decimal counter consists of several decades with four bistable circuits each. Of the $2^4 = 16$ possible states only 10 are used and when the 11th state is sensed, using a gate arrangement, the decade is reset to zero and a carry pulse is applied to the next decade [9]. Such decade counters are widely used in the PAILRED system. Some counters are more complicated because they have to be reversible. These counters have two inputs, a counting input to which pulses are applied and an input to control the direction of counting (up or down). They will not be described in detail here.

The AND gates in fig. 19 provide full decoding of all four possible states of the two bistable circuits A and B. To fully decode one decade of a decimal counter consisting of four stages, ten AND gates are needed each having four inputs. For the case where a counter is used to count a specific number of pulses, for instance, a number N equal to the distance between adjacent reflections (the constant A of fig. 9), only the state of the counter corresponding to the number N is of interest. In the case of n decimal digits the number N can be selected by a bank of n switches having 10 positions each. Switches that are used in these applications do not need a separate decoding circuit [9]. They contain an AND gate that, for each position, is connected to the proper combination of the 8 output terminals of the bistable circuits of one decade. The AND gates of the n separate decades are combined to one AND gate, shown symbolically in *fig. 20*. This AND gate will then give an output only when the state of the counter corresponds to the number N that was selected by the bank of switches.

A second function that can be performed by the counter circuit of fig. 19 is that of frequency division.

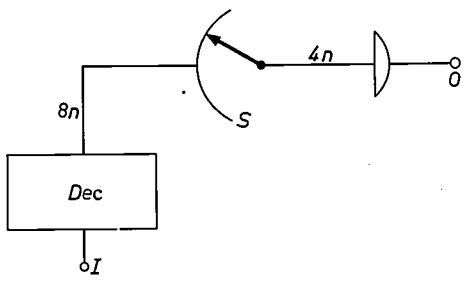


Fig. 20. A symbolic representation for a decade counter *Dec* containing n decades (i.e. $4n$ bistable circuits), together with a sensing circuit that will give an output at O when the counter reaches the state that was selected by a bank of n 10-position switches, S. The AND gate has $4n$ inputs, one from each bistable circuit, and is partially contained in the switches. I is the pulse input.

[8] E. J. van Barneveld, Digital circuit blocks, Philips tech. Rev. **28**, 44-56, 1967 (No. 2).

[9] C. Slofstra, The use of digital circuit blocks in industrial equipment, Philips tech. Rev. **29**, 19-33, 1968 (No. 1). In Slofstra's paper high voltage is "1" and low voltage is "0" (positive logic), whereas in the present paper the opposite convention is used (negative logic).

Each of the AND gates in fig. 19 will have an output once for every four input pulses applied to the pulse input terminal. Therefore, if the input pulses are applied at a frequency of 100 Hz, the pulses appearing at the output of the AND gates will have a frequency of 25 Hz. This type of frequency division is used in the control system to generate the pulse trains that drive the stepping motors and control their speed of rotation. The use of one accurate crystal-controlled oscillator in combination with a set of frequency-divider circuits eliminates the need for a large number of accurate oscillators.

Manual *x*- and *y*-control

The *x*- and *y*-motions of the linear-tracking mechanism on the diffractometer can be controlled manually so that any desired plane (*hk*) can be brought into the reflection position for study and measurement. To make this manual operation convenient, the *x*- and *y*-coordinates can be selected by means of switches called address switches, and the motions will automatically halt when these coordinates are reached. The circuit that is used for this function will now be discussed in detail (*fig. 21*). Two identical circuits of this type are provided to be able to control the *x*- and *y*-motions independently.

The stepping motor *Mot* is located on the linear-tracking mechanism of the diffractometer. The motor is energized by a motor-drive circuit *MD* and pulses *mp* applied to this circuit cause the motor to move by one step for each pulse applied. The motor-drive circuit has a second input, the reverse-control input; a signal *rev* present on this input reverses the direction of the motion. This same reverse signal, and also the motor pulses, are applied to a (reversible) decimal counter *Dec*, consisting of five decades. The direction of counting of this counter, also, is reversed when the reverse-control signal *rev* is present. Thus, the counter at all times contains the coordinate position of the linear-tracking device and, for convenience, it is supplied with a numerical read-out display located on the control panel.

The motion is controlled with a bistable circuit, the start-stop control circuit *SS*, in a manner similar to the control shown in fig. 18. The unmarked pulse inputs to the pulse gates in fig. 21 are connected to a clock signal, which is a continuous train of negative pulses. The motion can be started in the positive or negative direction by operating the *Start*+ and *Start*— push-button switches, provided that the coordinate position is not the same as the position selected by the address switch *AS*. When the position of the coordinate counter reaches, or equals, the address selected with the address switch, the AND gate associated with this switch will generate the address signal *as* which will then return the start-stop control circuit *SS* to the "0" state, thus stopping the motion. The address signal *as*, through the logic inversion,

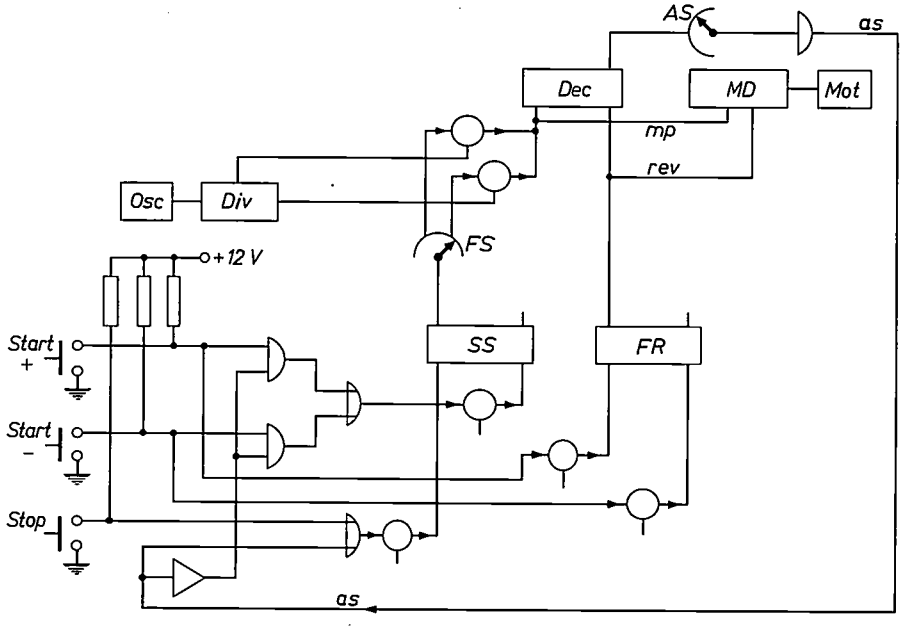


Fig. 21. Manual control of *x*- or *y*-motor *Mot*. Motor pulses *mp* are fed to the stepping motor by the motor drive circuit *MD*. The motor speed is chosen by the speed selector *FS* from an oscillator *Osc* with frequency divider *Div*. The direction of motion is controlled by the reverse signal *rev* from the forward-reverse control *FR*. The decimal counter *Dec* counts the motor pulses. The motor is stopped by an address signal *as* when the address selected by the address switch *AS* is reached. *SS* start-stop control.

will also inhibit the operation of the two *Start* buttons. The motion can also be halted independently by the operation of the *Stop* push-button switch.

Nine different speeds can be selected for the motor movement. By means of the speed selector switch *FS*, the start-stop control signal is applied to one out of nine pulse gates whose pulse inputs are connected to nine different pulse frequencies that are generated in the frequency divider *Div*. The pulse frequency, selected in this manner, determines the speed of the motor.

Sequence control

The sequence of operations that must be performed to measure each reflection and record the data will now be reviewed. In this discussion it will be apparent that the omega-cycle sequence control must interact with many other functions of the electronic system. These other functions will be briefly described as they appear in the sequence.

Step 0. While the linear-tracking device moves to the next reflection site, the omega cycle is inactive and remains in the rest position. Upon arrival at the reflection site the omega cycle is advanced to Step 1.

Step 1. The omega-cycle control first furnishes a print command to print the contents of the *HKL* register which identifies the reflection that is going to be measured. The *HKL* register is part of the sequence control that controls the motion of the linear-tracking mechanism. The information is printed out on paper to provide a permanent record, and also recorded on another medium, such as punched paper tape, which can subsequently serve to enter the data in a large electronic computer for data reduction and processing. At the completion of the print cycle, the print-control circuit gives a signal "print-finished".

Step 2. The omega-cycle control now gives a fast-scan command to the omega-scan control, which is a second sequence control. The omega-cycle control generates the main sequence of events, whereas the omega-scan control is in charge of moving the omega stepping motor and remembering its position. (This omega-scan control will be discussed in detail later on.) The crystal is now rotated about the omega axis by a small amount, for instance 1 degree, so that the detector can measure the background radiation level (position Q in fig. 13).

Step 3. With the crystal stationary at the extreme of the scan range (position Q), a fixed-time background measurement is made. The duration of the measurement is controlled by a timer unit which is reset and started by the omega-cycle control at the beginning of this step. The timer, in turn, starts an electronic counter, which will be called a scaler, which counts the number of X-ray quanta that is received by the scintillation detector.

At the completion of the background measurement, the results are printed out and, again, a print-finished signal is given as the completion signal for this task.

Step 4. The omega-cycle control, next, resets and starts the scaler, and at the same time it furnishes a slow-scan command to the scan-control circuit. The integrated intensity is now measured while the crystal is slowly rotated about the omega axis over a range of 2 degrees (from Q to R in fig. 13).

Step 5. Upon arrival at position R, the scaler is stopped and the measured integrated intensity is printed out.

Step 6. A fixed-time background measurement is made at position R and the results are printed out.

Step 7. The omega-cycle now gives a fast-scanning command to the omega-scan circuit and the crystal is rotated back by 1 degree to its original position (P).

Step 8. Upon completion of the entire omega-cycle sequence, the omega-cycle control is reset to its rest position (Step 0) and the x,y-coordinate system on the diffractometer now moves to the next reflection site.

A discussion of all the details of the omega-cycle control circuit is beyond the scope of this paper. *Fig. 22* will serve to illustrate the basic principles involved. The heart of the control circuit is the sequence generator *Seq* which is nothing more than an electronic counter with full decoding, identical to the circuit shown in fig. 19 except that more stages and more AND gates are needed. The $m + 1$ possible states of this sequence generator will be called S_0, $S_1, \ldots, S_m$, where S_0 is the rest position in which the control is not active.

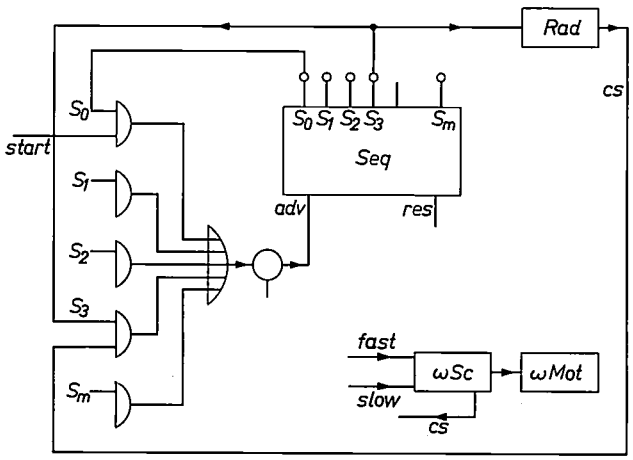


Fig. 22. The omega-cycle control synchronizes the omega ro-ration with the data-collection process. The steps of the sequence are generated by a basic counter circuit, the sequence generator *Seq* with advance input *adv* and reset input *res*. Its outputs are connected to the scan commands *fast* and *slow* of the omega-scan control *ωSc* and to the radiation-measuring circuits *Rad* with timer, scaler, and printer. Completion signals *cs* and the sequence start signal *start* are used to generate advance pulses.

When the sequence generator reaches a certain position, for instance S_3, at which point a certain action must be taken such as a fixed-time background measurement (see Step 3 above), a signal appears on output S_3 which is connected to the control input of the timer circuit, a part of the radiation measuring circuits *Rad.* When the background measurement and the print-out have been completed, the completion signal *cs* from the print-control circuit reaches an AND gate whose second input is connected to S_3. The output signal from the AND gate passes the OR gate and opens the pulse gate so that the sequence generator is advanced to the next position S_4. The use of the AND gate in this manner prevents the sequence generator from advancing when the completion signal is present during other parts of the automatic sequence where a different completion signal is logically needed to signal the end of the control step. Another function of the AND gate arises from the fact that the print-control circuit is not necessarily synchronized with the clock that advances the sequence generator. To ensure that the sequence generator advances, the duration of the "print finished" signal is made long enough. With the use of the AND gate, the advance signal for the sequence generator is removed as soon as it moves out of position S_3 even if the completion signal is still present.

Omega-scan control

The control of the omega rotation is under the direct supervision of the omega-scan control circuit which receives its scan-command signals *fast* or *slow* from the omega-cycle control that was discussed above. Because this control circuit is far less complicated than the omega-cycle control, it will be used as an example for a more detailed discussion of a sequence-control circuit.

When a scan-command signal is given, the omega motor must leave its rest position and move to one extreme of the scan range. If the scan command is then removed, the omega motor will remain at that extreme; if the scan command stays, the motor will reverse and scan in the opposite direction. The back-and-forth scanning motion will continue as long as the scan-command signal is present.

The eight basic states that this control has to recognize are shown in *fig. 23*. The rest state P_0 and state P_4 correspond to the center position of the scanning range (point P in fig. 13). States P_2 and P_6 are the extreme positions of the range corresponding to points Q and R (fig. 13). The angular range of motion from P_0 to P_2, etc., is adjustable by means of a scan-range selector-switch. The actual omega motion takes place during the odd-numbered P states.

Fig. 24 shows the sequence-control circuit that performs this task. Because there are eight basic states, the sequence generator consists of three bistable circuits (A, B and C) which are connected as a standard counting circuit (as in fig. 19) with the exception of the first stage (A). Full decoding of the eight states is done by AND gates in the standard manner (see fig. 19).

Initially the sequence generator is in the zero position ($A = B = C =$ "0") and the same is true for the angle counter *Dec.* The angle counter counts the number of motor pulses, and because each step of the omega stepping motor corresponds to 0.001 degree, thousand pulses must be counted for each degree of omega rotation.

The presence of a slow-scan command *slow* opens one of a set of pulse gates to pass the motor frequency from the frequency divider *Div.* The speed selection is similar to the speed selection of the x- and y-motions as shown in fig. 21. The motor pulses *mp* are applied to the motor-drive circuit *MD* and are counted by the angle counter *Dec.* The first motor pulse sets bistable circuit A to the "1" state after which its pulse gate G_2 is closed by its output Q_2. When the angle counter *Dec* reaches the number preset in the scan-range selector-switch *RS*, it is reset to zero. The reset pulse *res* passes through pulse gate G_1 of stage A and the sequence generator moves to state P_2. The P_2 signal serves as a completion signal to the omega-cycle control. If necessary, the omega-cycle control can then act and remove the scan-command signal.

An interesting timing problem arises from the fact that when the sequence generator moves to state P_2 the scan-command signal is still present. The scan command cannot be removed until the omega-cycle sequence generator moves to the next position, which may take several clock periods. Because the central control clock has a frequency of 10 kHz, the scan command will be present for a few multiples of 100 μs after P_2 has been reached. Since, however, the highest motor frequency is 400 Hz, i.e. 2500 μs is available between motor pulses, the scan command can be removed in plenty of time to prevent the next motor pulse from passing the pulse gate. Consequently,

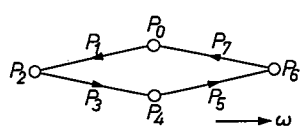


Fig. 23. The eight states that are needed to generate a back-and-forth scanning motion of the crystal on the omega axis. The even-numbered P-states correspond to stationary positions at the center of the range (P_0 and P_4) and its extremes (P_2 and P_6). The actual motion takes place during the odd-numbered P-states.

bistable circuit *A* will remain in the "0" state and the sequence generator will not move to position P_3.

The direction of the omega motion is also determined by the sequence generator and in positions P_1 and P_7 the motor moves in the reverse direction. The reverse control line is also actuated by states P_0 and P_6 because the sequence generator is still in these states when the first motor pulse arrives.

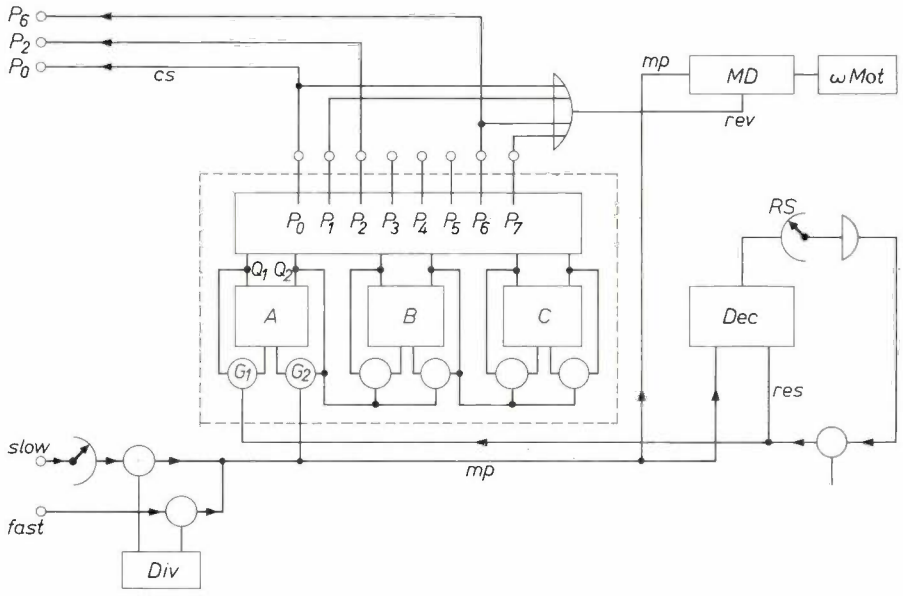


Fig. 24. Omega-scan control. The sequence generator (dotted frame) contains three bistable circuits *A*, *B* and *C* and a decoding network $P_0 \ldots P_7$. Its outputs generate completion signals *cs* and the reverse control *rev* of the omega motor *Mot*. The motor pulses *mp* to the motor-drive circuit *MD* are selected from a frequency divider *Div*. They are admitted by the scan commands *slow* or *fast* from the omega-cycle control. The motor pulses are counted in the angle counter *Dec* connected to the scan-range selector *RS*. At the end of the scan range the counter is reset (*res*) and the sequence generator is advanced.

Control panel

The circuit principles used in the automatic sequence-control unit have been discussed and illustrated with the details of a few representative functions. All these functions are controlled from the control panel which is shown in *fig. 25*. All three stepping motors on the diffractometer (the *x*-, *y*- and omega motors) are controlled from this panel and the experimental parameters for the fixed-program sequential routines are adjusted by means of the switches shown in fig. 25.

The panel is divided into five major areas according to function. The area on the left-hand side contains all switches needed to control the *x*- and *y*-motions. The address switch for the manual *x*-control is located at the top. These controls were discussed in detail in fig. 21. Below it is an identical control for the *y*-motion, followed by switches to independently adjust the *x*- and *y*-speeds. The parameters for the size of the reciprocal cell can be adjusted with the switches *A*, *B*, *C* and *D*. The letters correspond to those used in fig. 10. Of course, *C* and *D* are used only for oblique reciprocal nets.

The function switch is located at the bottom of the panel. The three functional modes of operation are

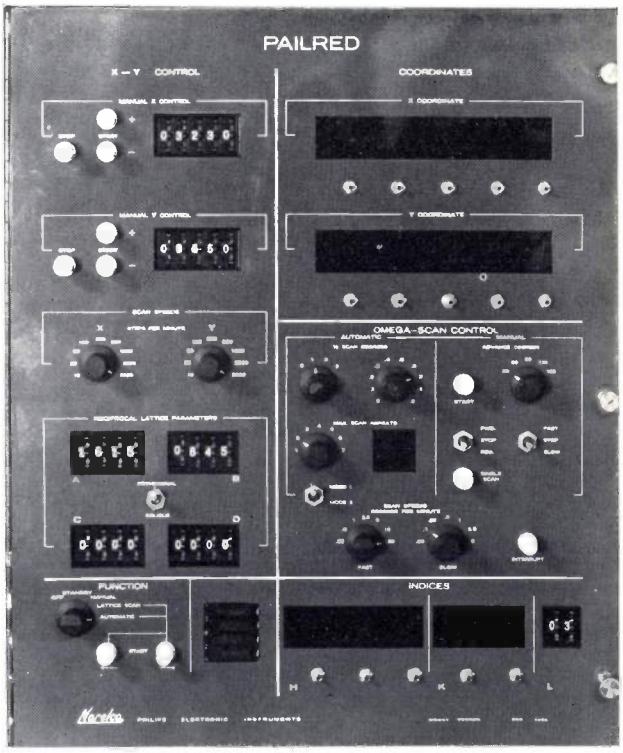


Fig. 25. The experimental parameters of the automatic data-collection process are selected by means of switches located on the control panel. Manual operation of the three motor drives on the diffractometer is also done by means of switches.

Manual, *Lattice-Scan* and *Automatic*. Of these only the lattice-scan has not yet been discussed. This is a semi-automatic mode of operation that is used to assist in the search for reflections during the initial alignment of the axes of the crystal with those of the diffractometer. The lattice-scan searches for reflections by scanning the x, y-space of the diffractometer following a trajectory as indicated in *fig. 26*. The scanning motion is continuous and does not stop at reflection sites because these have still to be determined. When a reflection appears on the strip-chart recorder, its coordinates can be read at the coordinate displays at the same moment, or afterwards from the strip chart.

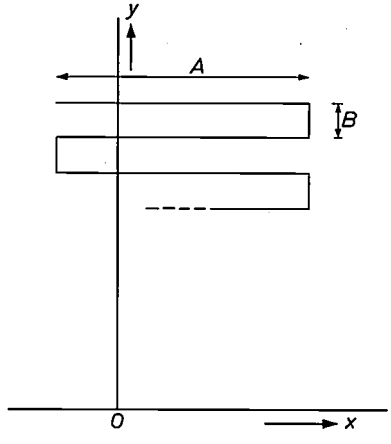


Fig. 26. The lattice scan is a semi-automatic mode of operation. When started positive, x is increased by A, y decreased by B, x decreased by A, etc.

To adjust the distances A and B of fig. 26, the same switches A and B are used which, in the automatic mode of operation, are used for selecting the size of the reciprocal-lattice cell. The lattice scan is a two-dimensional search in reciprocal space which can be extended to the third dimension by using the equi-inclination technique and repeating the two-dimensional scan at different settings of the inclination angle (μ). The lattice scan can also be used to get a quick survey of the reflection intensities in order to determine the symmetry of the crystal or to check the alignment. In this case the constant B is chosen equal to a reciprocal-lattice constant, whereas for search procedures it is chosen much smaller.

The numerical displays that show the contents of the x- and y-coordinate counters (fig. 21) are located at the top of the control panel on the right-hand side (fig. 25). Their only function is to allow the experimenter to conveniently read the x- and y-positions of the linear-tracking mechanism on the diffractometer.

The switches that control the third motor, the omega motor, are located below the coordinate displays. Manual switches are provided to move the omega motor forward or in reverse. Two ranges of speeds are available, slow and fast, with an independent selection switch for either range. The omega motor may be used for accurate determination of the lattice constants. In this procedure the linear-tracking system is not used, the detector is set manually at the proper angle for each reflection, and the crystal is rotated by the omega motor. The angular position at which a reflection occurs can be read with a precision of 0.001° at the scale of the omega drive (fig. 12), or, if desired, at the front panel display of the y-coordinate.

In aligning the crystal it is often desirable to move from one reflection to another symmetrically related to it. A facility is therefore provided for rotating the crystal about the omega axis by preset amounts. Since the only symmetries that can occur in single crystals are 2-, 3-, 4-, or 6-fold, the only angular increments that are needed are 180, 120, 90 and 60 degrees, and they are selected with the *Advance Degrees* control switch. A 30-degree rotation is also provided.

Again, for checking the alignment of the crystal a single complete automatic omega-cycle sequence, including background and integrated-intensity measurements, can be made as a manual function by operating a *Single Scan* push-button switch. The advantage of this scan as a manual operation is that the crystal is returned to its original position. The omega-cycle sequence can be interrupted at any time by pressing the push-button *Interrupt*; then the crystal will also return to the center position.

Switches are also provided to select the range of the omega scan from 0.1 to 3.9 degrees. The scan-range selection was discussed in the description of the omega-scan control (fig. 24) and because the distance from P_0 to P_2 is measured, the numbers above correspond to half the total scan range for the integrated-intensity measurement.

Two additional features of the omega-cycle control have not yet been discussed. Reflections with low intensity can be scanned repeatedly for increasing the accuracy. The maximum number of repeats which can be chosen is seven. To avoid a waste of time when a strong reflection is being measured, the scanning is stopped at one of the extreme positions after a preset count has been reached in the scaler. A preset-count control, which is part of the scaler circuit, interacts with the omega-cycle and the omega-scan controls to accomplish this. The actual number of repeats is printed and punched.

The second feature is the Mode I or Mode II operation of the omega cycle during automatic data collection. These two modes refer to the manner in which the background is measured. In Mode I, the background is measured at both sides of a reflection in the manner

that was discussed above. In Mode II the background is measured and averaged over a trajectory in reciprocal space during the time that the linear-tracking mechanism moves from one reflection site to the next.

Finally, at the bottom right-hand side of the control panel, the contents of the *HKL* register are displayed indicating the reflection that is being measured during automatic data collection. At the start of an automatic cycle the indices H, K and L of the first reflection to be measured are entered manually. Because the L index does not change during an automatic cycle, it is not under the control of the automatic electronic circuits.

Final remarks

An important point in the development of the PAILRED system was accuracy. This led to the use of a crystal monochromator to limit the spectral bandwidth, although the incident beam will then be inhomogeneous at the center of the diffractometer. This does not affect the accuracy if the crystal to be studied is spherical and is centered at the point of intersection of the axes about which it rotates. After the crystal has been aligned, the equi-inclination technique requires only two axes of rotation of the crystal, the omega and the mu axis. With the equatorial technique, where all the reflections are measured in the equatorial plane, the detector needs only one axis of rotation ($Y = 2\Theta$), but the crystal needs three axes ($\omega = \Theta$ and two axes of the Eulerian cradle) to bring an arbitrary plane into the reflecting position. The required high mechanical stability is obtained more easily when the crystal rotates on two axes than when it needs three axes of rotation.

In contrast to the equatorial technique, the equi-inclination technique requires two angles to be set very accurately, the alignment of the crystal and the inclination angle. Therefore, this is done manually and the only sequence to be automated is the measurement of one level. This operation is simplified by the use of the linear-tracking system. A special feature of the PAILRED is that no computer-generated program is necessary before the start of an automatic run.

Several crystal structures have already been determined using the PAILRED. An accurate reinvestigation of the structure of the mineral topaz yielded atomic positions with an accuracy of 0.05 picometer (0.0005 Å) [10].

At the Philips Research Laboratories in Eindhoven, the crystal structures of some steroids have been determined. The average disagreement between observed and calculated intensities was about 8%. With photographic techniques the disagreement is usually two or three times as large. The single-crystal project of the American Crystallographic Association [11] is another example of the accuracy that can be obtained with the PAILRED. In this project one single-crystal specimen of fluorite was measured on seven automatic diffractometers of different types. The resulting intensities were compared with the average and with each other. The PAILRED was one of the instruments for which the average deviation was as low as 3%.

[10] J. Ladell, Redetermination of the crystal structure of topaz: a preliminary account, Norelco Reporter **12**, 34-39 and 47, 1965.

[11] S. C. Abrahams, L. E. Alexander, T. C. Furnas, W. C. Hamilton, J. Ladell, Y. Okaya, R. A. Young and A. Zalkin, American Crystallographic Association single-crystal intensity project report, Acta cryst. **22**, 1-6, 1967 (No. 1).

Summary. The Philips Automatic-Indexing Linear Reciprocal-space Exploring Diffractometer, PAILRED, automatically measures and records the intensities of X-ray reflections, which are used in crystal-structure determinations. Apart from standard components, such as radiation-measuring circuits and the X-ray generator, the PAILRED system consists of a fixed-program control computer for directing the motions of the diffractometer. The diffractometer uses a linear-tracking mechanism, based on the Ewald construction in reciprocal space, to position the crystal and the detector for each reflection separately. Upper levels are measured by the equi-inclination technique. A crystal monochromator is used to reduce background radiation. All the required motions of the diffractometer are obtained by means of stepping motors which are ideally suited to digital control techniques because of the discrete nature of their movement. Electronic counters are used to count and control the number of steps of the motors, and frequency dividers generate accurate frequencies to control the motor speeds. Sequencing circuits generate the different steps of the automatic process and synchronize the motions of the diffractometer with the data collection. Special circuits provide manual or semi-automatic control of the diffractometer for the alignment of the crystal under examination and the determination of its lattice constants. All experimental parameters are adjusted by means of switches at the control panel.

Lubrication with gases

In lubrication with liquids two extreme cases may be distinguished: film lubrication and boundary lubrication. In the first case the running surfaces are completely separated by a liquid which retains all its normal physical properties. The viscosity of the liquid very largely determines the magnitude of the friction. In boundary lubrication the film between the surfaces is no more than a few atoms thick. In this case the chemical properties of the lubricant, which only barely retains the character of a liquid, are of more importance than its viscosity.

The influence of the lubricant in boundary lubrication may be understood as follows. Friction is the result of the binding forces which the atoms in one of the bearing surfaces exert on the atoms in the opposite surface. In the simplest case of two bodies made of the same metal in contact, the interatomic forces are the same as those which act inside the metal. This does not involve the entire surface, for this would mean that two bodies in contact with one another would behave as a single body. However smooth the running surfaces are, they are very uneven in terms of atomic dimensions and only the "peaks" of the two surfaces come into contact. This is where the making and breaking of the bonds between the running surfaces takes place. What is wanted of the lubricant is that, after having been firmly adsorbed by one of the running surfaces, it should only form weak bonds with the opposite running surface. The resistance to a relative displacement of the two surfaces is thus reduced. A familiar example is the reduction of friction obtained by the use of fatty acids and compounds derived from them. Fatty acids have a polar group, by which they are bound to one of the running surfaces, and a non-polar group, which turns towards the other running surface.

Less familiar than lubrication with liquids (and often not recognized as lubrication) is lubrication with gases. Here again, a distinction can be made between film lubrication and boundary lubrication. In connection with research on spiral-groove bearings, an article appeared in this journal some years ago on film lubrication with gases (gas bearings) [1], which had been found of great practical importance. *Boundary lubrication by gases* has so far not been extensively investigated although it is a phenomenon which is in fact encountered daily. We need only recall here the "cold welding" of metals which occurs when two pieces of metal are rubbed together in a high vacuum. The effect of the rubbing is to remove any residual material adsorbed on the surface and to bring the pieces of metal together with clean surfaces. This is sufficient for them to stick to one another. It appears that under normal atmospheric conditions most objects do *not* stick to one another because of the adsorption of atoms or molecules from the atmosphere on the surface. Oxygen is particularly active in this process.

Boundary lubrication by a gas also plays a significant role in the lubricating action of graphite. It has long been known that graphite loses its lubricating properties in a high vacuum. Investigations by Savage have demonstrated that this is due to the absence of water vapour [2]. In equilibrium with water vapour a film of water is adsorbed on the graphite surface, the thickness of the film depending on the water vapour pressure. In the absence of this film of water graphite has no lubricating action.

At the Philips Research Laboratories in Eindhoven we have been looking into the influence of *organic vapours* on the frictional behaviour of metals. The adsorption of organic vapours on metal surfaces is not unknown, but little attention has previously been paid to the application of this adsorption for lubrication.

Under normal atmospheric conditions the lubricating action of organic vapours is usually noticeable only with the *noble metals*. This is because the non-noble metals have a much greater affinity for oxygen, and thus become coated with an oxide film [3].

The fact that the noble metals do not have such an oxide film and therefore provide a low contact resistance, also accounts for their extensive use in electrical contacts. One special group is that of sliding contacts, which have to meet some difficult mechanical requirements. Sliding contacts must at the same time have low contact resistance and low friction and, above all, they must have minimum wear. These requirements are more or less conflicting. It seemed to us that boundary lubrication with organic vapours might

[1] E. A. Muijderman, Philips tech. Rev. **25**, 253, 1963/64.

[2] R. H. Savage, J. appl. Phys. **19**, 1, 1948.

[3] In circumstances where the oxide film is removed and does not quickly re-form (e.g. heavy mechanical loading and low partial pressure of oxygen), organic vapours may well have a lubricating effect on non-noble metals. This has been demonstrated for steel by G. V. Vinogradov, V. V. Arkharova and A. A. Petrov, Wear **4**, 274, 1961; and by R. S. Fein and K. L. Kreuz, ASLE Trans. **8**, 29, 1965; and for copper by G. W. Rowe, Proc. Conf. on lubrication and wear, London, 1957, paper 5.

Strangely enough, the lubrication effect did *not* occur at extremely low partial pressures of oxygen.

[4] See H. W. Hermance and T. F. Egan, Bell Syst. tech. J. **37**, 739, 1958.

offer a good solution to this problem. The adsorbed film would have to be so thin that the electrons would be able to move through it without encountering any significant opposition.

Generally speaking, organic vapours have an *adverse* effect on contacts (e.g. in telephone exchanges), since there is greater wear due to increased sparking and a gradual increase of resistance due to polymer formation on the contact faces [4]. With sliding contacts, however, it appears that the increased sparking is more than offset by the reduction of wear due to the lubricating action of the vapour, and the sliding movement keeps the contact surface free from polymers.

To investigate the effect of organic vapours on the lubrication of sliding contacts, we used an arrangement consisting of a disc which was rotated at a constant speed while a spring-loaded strip was kept pressed against it by a constant force. The whole assembly was enclosed in a bell-jar in which the atmosphere could be accurately controlled. Measurements were made of the frictional force exerted by the disc on the strip. We also measured the contact resistance between disc and strip, to find out for example whether the lubrication was boundary or film lubrication. The resistance will be very high for film lubrication.

Fig. 1 shows the results of an experiment with a gold disc and a gold-plated phosphor-bronze strip. A continuous stream of nitrogen was passed through the bell-jar, with addition of benzene vapour during regular ten-minute periods. The graph shows that at every addition of benzene vapour the coefficient of friction decreased from about 0.6 to about 0.2, whereas the contact resistance showed no significant change. The above experiment also shows that the benzene vapour should be added continuously to keep down the friction: since the lubricating film is extremely thin, it must be continuously replenished.

A very simple solution for this is to place near the running surfaces an organic substance in liquid or solid form, which has a very low vapour pressure and thus gives up its molecules slowly (this is how mothballs work). Anthracene, which is solid at room temperature and has a vapour pressure of only 4×10^{-4} torr, is very useful for this purpose, making it possible to reduce the coefficient of friction of gold on gold, for example, to less than 0.25. When a piece of anthracene was used in experiments on the lubrication of brush-collector contacts in small low-voltage motors the life of the contact was greatly extended.

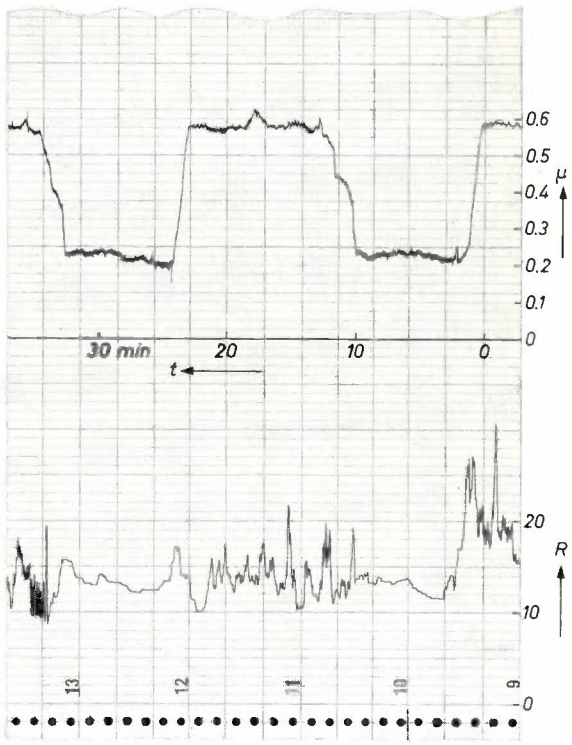
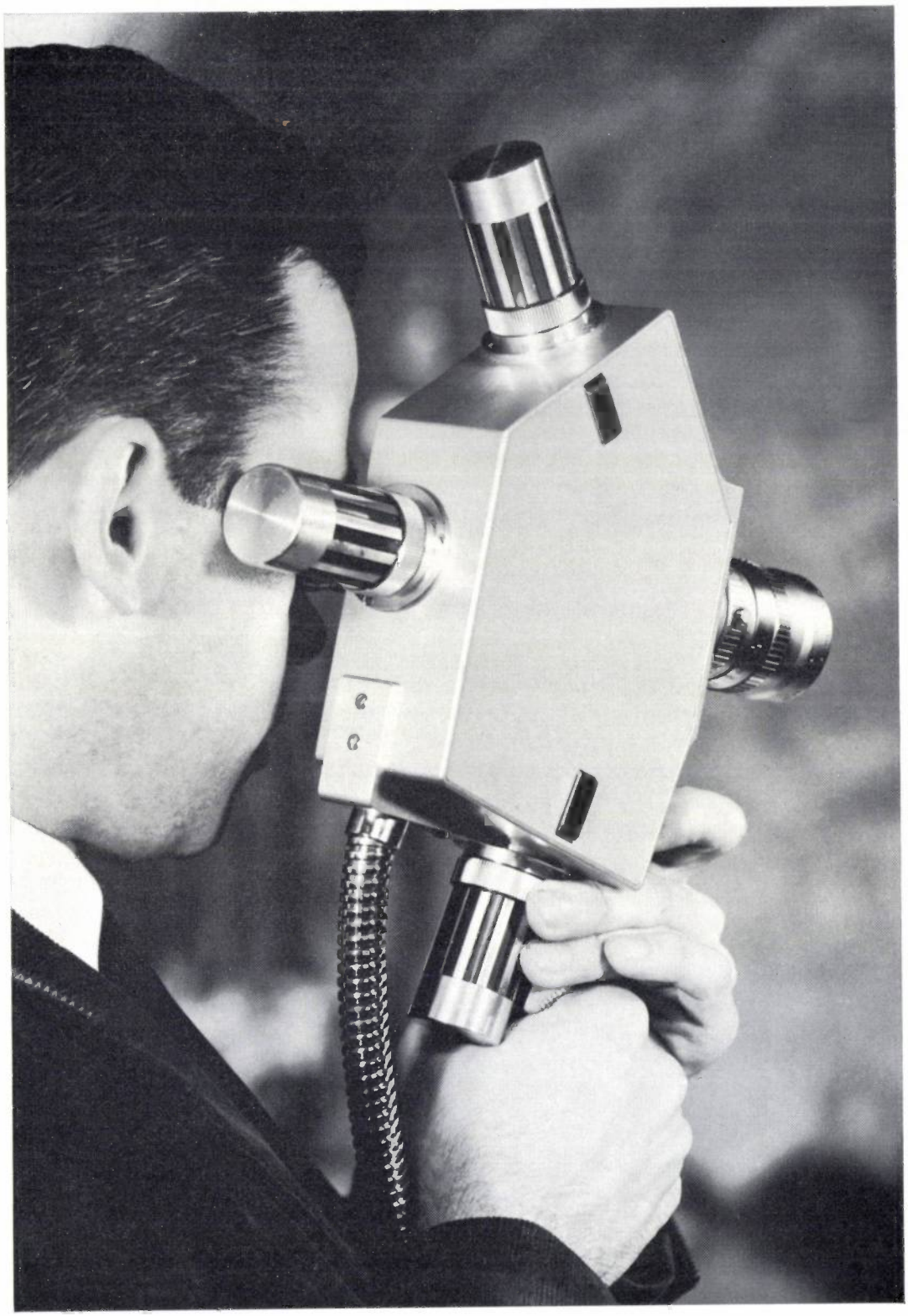


Fig. 1. The coefficient of friction μ of gold on gold and the contact resistance R as a function of time, measured in a stream of nitrogen to which benzene vapour was periodically added. During the addition of benzene vapour the coefficient of friction is reduced from about 0.6 to about 0.2. R is indicated in arbitrary units.

T. J. Berben
C. W. Berghout

T. J. Berben and Dr. Ir. C. W. Berghout are with Philips Research Laboratories, Eindhoven.

Portable colour television camera



A "Plumbicon" camera tube of reduced size (1.6 cm dia., 13 cm length), which gives very sharp pictures in spite of the small size of the image, has recently been developed at Philips Research Laboratories, Eindhoven. Three tubes of this new type are used in the small colour television camera shown, whose weight is only 5 kg (including zoom lens and electronic view finder). This camera was demonstrated to television experts in April 1968.

The tube and the camera will be the subject of a forthcoming article in this journal.

A process for converting quantities to be measured into frequencies

D. Meyer

A new type of amplitude control makes it possible to obtain an essentially linear variation of the frequency or period of a harmonic oscillator as a function of a single circuit parameter. Oscillators controlled in this way can be used to convert the output from a conventional passive transducer into a signal whose frequency variation conveys the information about the measured quantity. This kind of signal is extremely insensitive to interference on the transmission channel.

Introduction

Electrical measurements of mechanical quantities can usually be made to sufficient accuracy with conventional transducers. Difficulties do not occur until the output signals of these transducers have to be transmitted to the indicating instrument along a transmission path which is long and subject to heavy interference — just the kind of conditions likely to be encountered in industrial instrumentation. In conventional processes the quantities to be measured are represented by voltage or current amplitudes or by resistance values. This is known as "amplitude analogue" representation. The difficulties mentioned above arise because amplitude-analogue signals are susceptible to unwanted effects such as: stray voltages (thermoelectric voltages, drift voltages in d.c. voltage amplifiers, induced voltages); variations in gain and frequency characteristics due to changes in the temperature or the supply voltage; and variations in the characteristics of transmission channels.

This kind of interference on the transmission channel has no effect, however, on "frequency analogue" signals, that is, signals whose frequency or period represents the measured quantity. The frequency can be measured very accurately with digital counters by counting the number of periods of oscillation in a reference time interval. The period (the reciprocal of the frequency) can be measured just as accurately by counting the number of pulses of a reference signal of fixed frequency that occur in one or more oscillation periods. This simple facility for digitizing is of itself an advantage of frequency-analogue techniques, since digital representation is often highly desirable for accurate read-out and easier further processing and storage. On the other hand, simple circuits can be used to reconvert the frequency-analogue signals to amplitude-analogue signals, so that the frequency-analogue technique remains compatible with the use of conventional ancillary equipment (recorders, control units, etc.).

Frequency-analogue signals can be derived from a conventional transducer with amplitude-analogue output, by connecting it to a converter whose output frequency (or period) is uniquely and stably related to the output signal of the transducer. This has the advantage that a large number of well-tried and accurate transducers are then available for use. The converter should meet the following requirements:

1) It should be insensitive to changes in external conditions (temperature, operating voltage).
2) It should be small, so that the converter can be placed close to the transducer, and long lines for the amplitude-analogue signal from the transducer are not required.
3) The relation between the quantity to be measured and the output frequency or period should be linear to simplify further signal handling.

We shall now give a description of a converter that works on a new principle which has been developed at the Philips laboratory in Hamburg [1]. A converter working on this principle satisfies the above requirements and is suitable for many kinds of passive amplitude-analogue transducers.

Conversion of resistance variation into variations of frequency or period

For converting amplitude-analogue signals into frequency-analogue signals, an oscillator is used. This can be either a harmonic oscillator or a relaxation oscillator (square-wave or sawtooth output). In passive trans-

Dipl.-Ing. D. Meyer is with the Hamburg Laboratory of Philips Zentrallaboratorium GmbH.

[1] D. Meyer and W. Schott, German Patent application, 9 Feb. 1966; D. Meyer, German Patent application, 22 Nov. 1966.

ducers the amplitude-analogue quantity which has to control the oscillator is a resistance R, a capacitance C or an inductance L.

A relaxation oscillator always contains an active element which undergoes a discontinuous change of state (relaxation) if its input level exceeds (or falls below) a critical threshold. In addition such an oscillator has one or more passive elements, which determine the time intervals between the discontinuous changes of state of the active element (or elements). These oscillators can be controlled easily, and in theory linearly, by voltages, currents or the values of resistances or reactances. A disadvantage, however, is that the frequency is not only affected by the passive elements, but is intrinsically related to the height of the threshold. The threshold is determined by factors which include the characteristics of the active elements. On account of their simplicity, relaxation oscillators are often used for converting amplitude-analogue signals into frequency-analogue signals. However, if great accuracy and low sensitivity to external conditions are required, the circuits become complicated and expensive.

While the relaxation oscillator can only function because of the non-linearity of an active element, a harmonic oscillator operates in theory entirely with linear elements. This type of oscillator has the attractive feature of an oscillation frequency which is completely independent of the amplitude. If the oscillator has an ideal amplifier — which in practice can be closely approximated by an amplifier with heavy negative feedback — the frequency of oscillation is determined entirely by linear passive elements, such as R, C or L. Thus fluctuations in the operating voltage are of no significance, while temperature fluctuations are of importance only to the extent to which they affect the passive, frequency-determining elements. On the other hand there is the objection that if the oscillator is controlled by just a single passive element there will always be a *non-linear* relation between its value and the oscillation frequency f_0 or the period T_0. Thus in LC oscillators, and similarly in RC oscillators and RL oscillators with two resistances and two reactances, the equation for f_0 or T_0 always includes the *square root* of the frequency-controlling element, as can be seen from the example of the LC oscillator:

$$T_0 = 2\pi \sqrt{LC}. \quad \ldots\ldots \quad (1)$$

For oscillators with more than the minimum number of frequency-determining elements, the non-linear functions are rather more complicated.

However, it is possible to make a converter with a linear characteristic in a way which we shall now describe.

A harmonic oscillator consists of a frequency-determining network, with the (complex) transfer function F, and an amplifier whose voltage gain is A (A is real). The amplifier and the network are connected to form a feedback system, i.e. they form a closed circuit. In this circuit stable oscillation is possible only if the loop gain A_k equals unity:

$$A_k = \vec{F} \cdot \vec{A} = 1. \quad \ldots\ldots \quad (2)$$

The frequency-determining network is in effect a voltage divider whose step-down ratio (voltage attenuation) is a function of the frequency in amplitude and phase. There is only a single frequency, f_0, for which the phase shift of this voltage divider is zero. Provided the amplifier gives no additional phase shift, f_0 is the only possible oscillation frequency, since this is the only frequency for which the phase condition of eq. (2) is satisfied. If eq. (2) is also to be satisfied for the amplitudes, the gain A of the amplifier must be exactly equal to the attenuation of the network at the frequency f_0.

But such a structure will not in practice give stable oscillation; the slightest change in any of the parameters — e.g. in the gain or in the frequency-determining elements — will at once affect the loop gain and thus cause an increase or decrease in the oscillation amplitude $\hat{u}_0$. To make the oscillator stable, a control circuit is required which will make the gain dependent on the amplitude. A suitable control circuit can be obtained by using voltage-dependent resistances such as lamps or thermistors in the feedback loop of the amplifier [2].

A well-known example of a harmonic oscillator is the Wien bridge oscillator, whose frequency-determining network consists of two resistances and two reactances (*fig. 1*). The oscillation frequency f_0 is given by [3]:

$$f_0 = \frac{1}{2\pi} \sqrt{\frac{1}{R_1 R_2 C_1 C_2}}. \quad \ldots \quad (3)$$

If the resistances R_1 and R_2 are equal, and also the capacitances C_1 and C_2, then at this frequency the attenuation of the frequency-determining network is 3 times, so that the gain must be $A = 3$.

If a Wien bridge oscillator is to be used for converting "resistance" to "period", then R_1 (say) could be the control resistance which depends on the quantity to be measured. If all the other network elements (R_2, C_1, C_2) remain constant, then we obtain the square-root relation mentioned above between the control resistance and the period of the oscillator:

$$T_0 = 2\pi \sqrt{R_1 R_2 C_1 C_2}. \quad \ldots \quad (4)$$

The attenuation of the network also changes. If R_1 increases, then the impedance of R_1 and C_1 connected in parallel increases, so that the attenuation of the

voltage divider decreases. In conventional oscillators this change is counteracted by the amplitude control circuit, which gives a corresponding and opposite change in the gain.

Our method for keeping the loop gain (eq. 2) constant is essentially different: changes in attenuation are suppressed at source. If R_1 is changed and thus the impedance of R_1 and C_1 in parallel, the attenuation of the voltage divider can be kept constant if the impedance of the series circuit R_2, C_2 is changed by the same amount (*fig. 2*). Taking R_2 as the variable element of the series branch, then the attenuation (at the resulting oscillation frequency) will always be 3 if $R_2 = R_1$. If the gain is kept constant at $A = 3$ and there is a control circuit which adjusts R_2 in such a way that the oscillation amplitude does not vary with time, then the second resistance (R_2) of the frequency-determining network will accurately follow all the changes of the first resistance (R_1). When $R_2 = R_1$, eq. (4) becomes:

$$T_0 = 2\pi R_1 \sqrt{C_1 C_2}, \quad \ldots \quad (5)$$

thus giving the desired *linear* relationship between period and R_1.

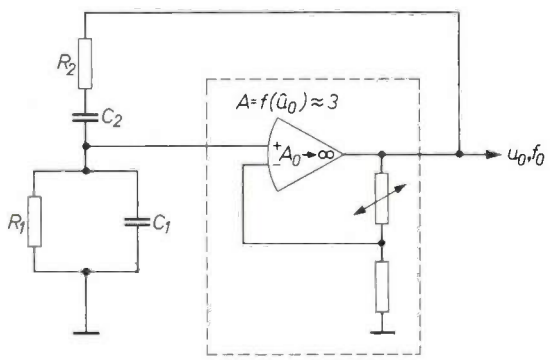


Fig. 1. Wien bridge oscillator, with frequency-determining network consisting of resistors and capacitors. The gain A required for stable oscillation is controlled by amplitude-dependent negative feedback of a difference amplifier.

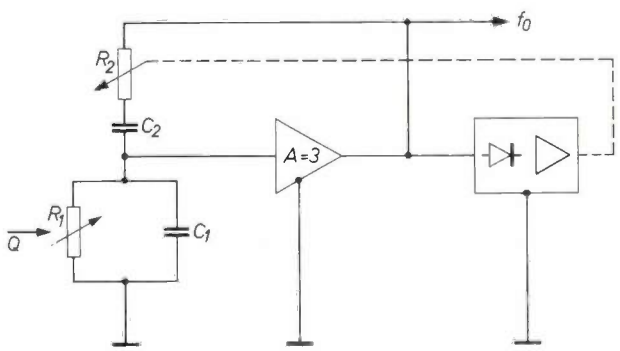


Fig. 2. Wien bridge oscillator for representing a measured quantity Q by a frequency f_0 (or period T_0), with amplitude control via the frequency-determining network.

The control circuit for varying R_2 consists of a rectifier supplying a d.c. voltage proportional to the oscillation amplitude, an amplifier and, as the control element, a resistance which can be varied by the output from the control amplifier. This variable resistance can be a motor-driven potentiometer, a photoconducting cell illuminated by a lamp, or a resistance controlled by a magnetic field.

Although the output from the control amplifier (e.g. the lamp current above) has to take different values and

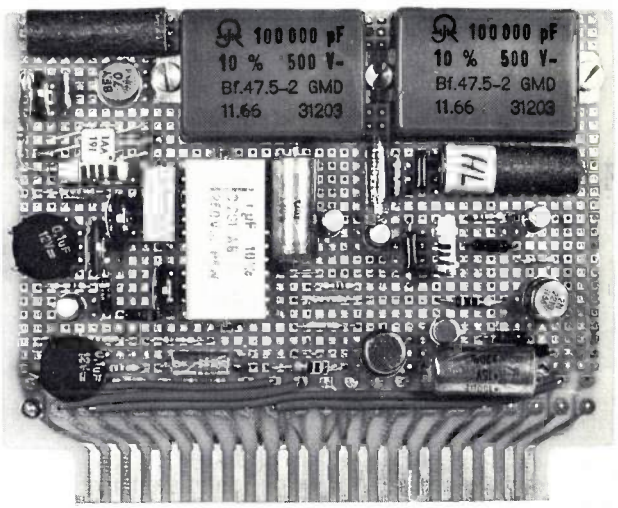


Fig. 3. Experimental version of a resistance-to-period converter based on the circuit shown in Fig. 2. The component HL contains the combination of lamp and photoresistor. The small square component TAA 191 is the control loop amplifier, designed as an integrated circuit. The wiring board is 120×100 mm.

the oscillation amplitude has to follow suit, this has no effect on the oscillation frequency; the frequency of the oscillation is independent of its amplitude. All that is required is that for a given value of R_1 the amplitude should not vary with time. Since the slightest deviation of the resistance R_2 from equality with R_1 gives a change in oscillation amplitude and is thus cancelled out, R_2 is maintained at exactly the same value as R_1 and the relation between T_0 and R_1 is accurately described by eq. (5). (As in circuits with integrating control, there is no need for a residual difference between R_2 and R_1.)

Fig. 3 shows an experimental version of this kind of converter. The three-stage amplifier for the oscillator

[2] Another common practice is to make the gain so high that the oscillation is just limited by the non-linearity present in the amplifier. With this manner of stabilization, however, the frequency is dependent on the amplitude (Balth. van der Pol, Proc. IRE 22, 1051-1086, 1934).

[3] Detailed calculations can be found in: S. Seely, Radio Electronics, McGraw-Hill, New York 1956, Chapters 11-13.

has a gain of about 10 000. With sufficient feedback to reduce the gain to $A = 3$, the amplifier has exceptionally good gain stability, a high input impedance ($R_i \approx 10$ MΩ) and a wide bandwidth (≈ 3 MHz). These features are important for proper operation. An error in the gain would have to be compensated by a change in R_2, which would give an error in the oscillator frequency. The input impedance of the amplifier must be large with respect to the resistance R_1, which determines the frequency, since it is directly in parallel with it. Finally, the wide bandwidth is required to prevent phase shifts from occurring in the amplifier, which could affect the oscillator frequency.

The control circuit consists of a rectifier, a two-stage amplifier, several *RC* chains in the feedback loop for optimizing the control action, and a combination of photoresistor and lamp. The experimental version of the converter handles a range of resistance from 300 Ω to 10 kΩ, corresponding to a variation in the period from about 0.2 to 6 ms. In the range 0.6-6 kΩ the maximum linearity error has the exceptionally low value of ± 0.05% of the measured value (*fig. 4*). The error is greater at the *higher* resistance values because the control element is connected in parallel with the amplifier input resistance, R_i, which remains constant. This explanation is confirmed by the dashed curve shown in fig. 4, which shows how the measured period departs from that given by eq. (5) if the resistance of R_1 and R_i connected in parallel is inserted instead of R_1. At *low* resistance values, corresponding to a high oscillator frequency, there is increasing difficulty due to an error which occurs because of the finite bandwidth of the amplifier.

The temperature error plotted in *fig. 5* (curve *a*) is chiefly due to the temperature drift of the frequency-determining capacitors. This error can be readily compensated if the frequency-determining capacitors in the reference oscillator of the counter for measuring the period are of the same type as those in the resistance-to-period converter, and mounted in the same enclosure to give good thermal coupling (fig. 5, curve *b*; an extra lead is then required between converter and indicating instrument). At the same time this does away with the need for the quartz oscillator usually used as the frequency standard for the counter.

Variations in the operating voltage of ± 10% give errors of less than 0.01% in the output of the converter.

These results show that the resistance-to-period converter we have described is suitable for accurate determination of resistance, and can therefore be used in combination with any transducer which converts the quantity to be measured into a variation of resistance. Transducers of this type include resistance thermometers and various types of sensing elements fitted with potentiometers e.g. for measuring displacement, angle or pressure. If the frequency-determining networks are suitably modified, the same principle may be applied for deriving circuits which will convert resistance variation into frequency variation, or convert a variation in capacitance or inductance into a change of period. We have achieved successful results with circuits such as these, but space forbids a more detailed description here.

The converter is only of limited use for dynamic

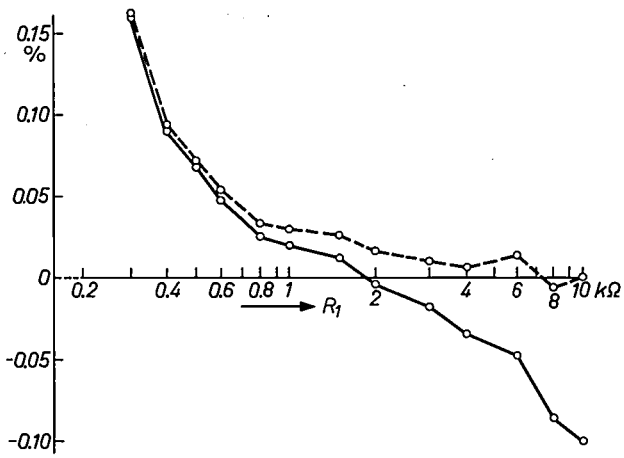


Fig. 4. Linearity error (relative to the measured quantity) in the conversion of resistance (R_1) into period (T_0). The relative deviation from the value of T_0, calculated from eq. (5), is shown. The increasing error as R_1 increases is due to the parallel connection of the constant resistance R_i of the amplifier input impedance. This is confirmed by the dashed curve, which is obtained by substituting the resultant value of R_1 and R_i in parallel instead of R_1 in eq. (5).

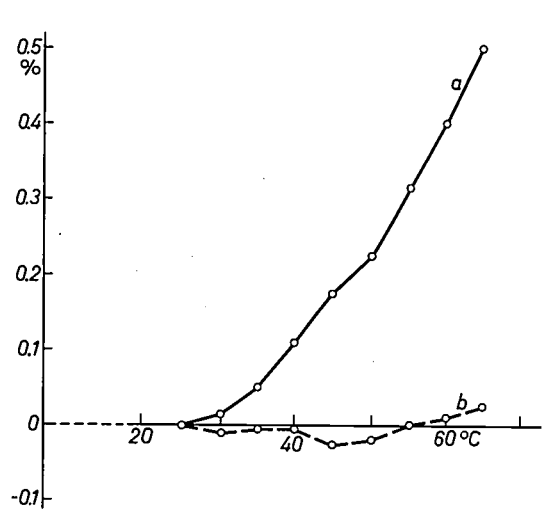


Fig. 5. Temperature error of the resistance-to-period converter: *a*) absolute error, and *b*) compared with the frequency of a reference oscillator whose frequency-determining capacitors were thermally connected to those of the converter by mounting oscillator and converter in the same case.

measurements. The control circuit which is the essential feature of the converter imposes a limit on the rate of change of the measured quantity. In the version described here the limiting frequency is 1 Hz.

"Magnifying" conversion

In some types of measurement the relative change in resistance is only small, either because the measured quantity itself undergoes only a small change or because the sensitivity of the transducer is low. In such cases it is desirable to have a converter which can also "magnify", i.e. convert the relatively small change in resistance into a much greater relative variation of frequency.

We shall now describe the principle of operation of such a converter [4], with particular reference to a device specially designed for use with a strain gauge. A converter of this type is of great interest, since in spite of their low sensitivity (relative change in resistance about 10^{-3} max.) strain gauges are widely used for measuring mechanical strain and all the quantities related to it such as force, pressure, pressure difference (flow), displacement and acceleration. On account of the extremely small output signals (1 to 10 μV minimum) from these transducers, it becomes particularly important to meet the requirement, mentioned earlier, that the converter should be small enough to mount it close to the transducer.

The conversion of the very small relative change in resistance — the unbalancing or "detuning" of a resistance bridge — into a proportional, large relative variation in frequency ($\Delta R/R \rightarrow \Delta f/f$ conversion) takes place in two steps:

1) The detuning of the resistance bridge is converted into a large and proportional variation of conductance.
2) This conductance, which gives a magnified representation of the detuning of the bridge, is made the controlling element of a converter like the one described above (converter from resistance to period, i.e. conductance to frequency), whose output frequency is thus proportional to the detuning of the bridge.

The first step takes place in an electronic compensating circuit. The detuning of a resistance bridge can be compensated for by supplying a compensating current i_k to the bridge at one of the output terminals. This current can be derived from a voltage u_k by means of a constant resistance R_k (*fig. 6a*). The compensating voltage u_k must be measured with respect to the potential of the output terminals of the bridge if it is to be proportional to the resistance detuning. Fig. 6*b* shows a circuit in which u_k can be measured with respect to earth (or in a more general case, with respect to the reference potential of the bridge supply circuit — one of the output terminals of the bridge being connected to this potential). The resulting shift in potential is compensated for by a second resistance $R_k' = R_k$ between the second output terminal of the bridge and earth. Very often all four resistances R_B of a strain-

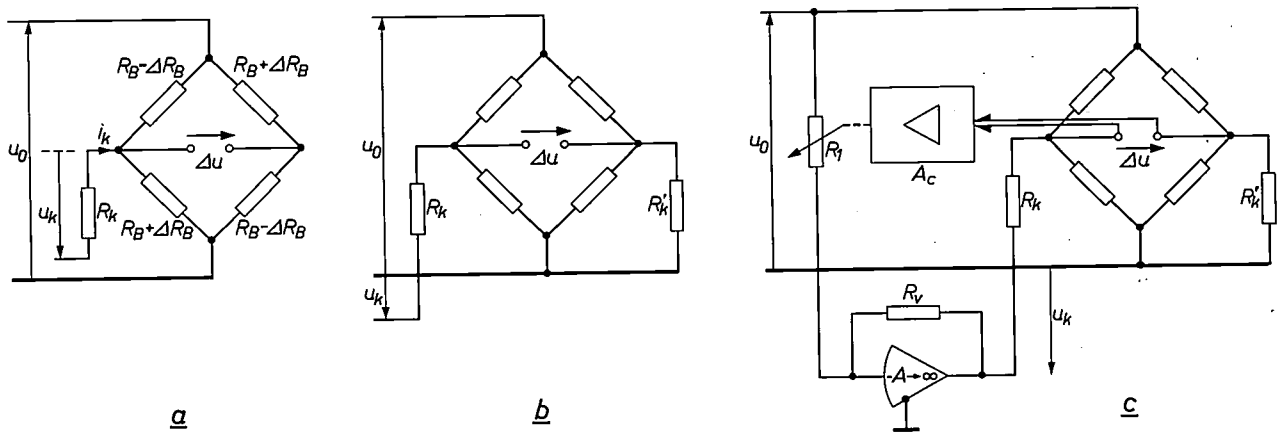


Fig. 6. The use of a self-balancing compensator circuit enables a small relative variation in resistance of a transducer to be represented by a very much greater relative variation in an electronically controlled resistance: *a*) Compensation of the detuning of a resistance bridge by a compensating voltage u_k. *b*) Equivalent circuit, but in which u_k is measured with respect to earth. *c*) Circuit in which the compensating voltage u_k is obtained from the bridge supply voltage u_0 via an operational amplifier, so that its input conductance $1/R_1$ is proportional to the bridge detuning $\Delta R_B/R_B$. A control amplifier A_c automatically adjusts R_1 to make $\Delta u = 0$.

[4] D. Meyer, German Patent application, 25 Apr. 1967.

gauge bridge are affected by the quantity to be measured, two diagonally opposite resistances increasing by an amount ΔR_B, and the other two decreasing by the same amount. The voltage required for compensation of the bridge is then:

$$u_k = -u_0 \frac{2R_k}{R_B} \frac{\Delta R_B}{R_B}. \quad \ldots \quad (6)$$

(This assumes that $\Delta R_B{}^2 \ll R_B{}^2$, which will certainly be the case since the greatest change in resistance that occurs in a strain gauge is still very small.)

The variable compensating voltage can be obtained from the bridge supply voltage u_0 by means of an operational amplifier. The gain of this operational amplifier is completely determined by a fixed feedback resistance R_v and a variable input resistance R_1 (fig. 6c):

$$\frac{u_k}{u_0} = -\frac{R_v}{R_1}. \quad \ldots \quad (7)$$

If the output voltage Δu of the resistance bridge is applied to a control amplifier which adjusts R_1 in such a way that Δu becomes zero, i.e. the bridge is compensated, then from eqs. (6) and (7):

$$\frac{1}{R_1} = 2 \frac{R_k}{R_v R_B} \frac{\Delta R_B}{R_B}. \quad \ldots \quad (8)$$

A linear relation has thus been obtained between the conductance $1/R_1$ and the detuning $\Delta R_B/R_B$ of the resistance bridge. By suitably choosing the resistance values R_k and R_v a very small relative detuning of the bridge (or change in resistance) can be made to correspond to a very much greater relative variation in conductance $1/R_1$. Like the variable resistance R_2 of the resistance-to-period converter described above, the variable resistance R_1 may be a motor-driven potentiometer, a photoresistor illuminated by a lamp, or a resistance with magnetic-field control.

Fig. 7 shows how this compensating circuit should be connected to the resistance-to-period converter so that the conductance $1/R_1$ of the compensating circuit determines the frequency of the oscillator.

Since an operational amplifier only requires an extremely small input voltage, because of its high gain, the lower end of the resistance R_1 may be taken to be earthed. It can therefore be identified with R_1 in fig. 2.

The oscillator output voltage may be used as the supply for the bridge instead of the voltage across R_1, since this is always 3 times as large as the voltage across R_1 because of the constant gain factor $A = 3$. Since the voltage of the bridge then differs from that across R_1, the gain of the operational amplifier must be altered by a corresponding increase in the feedback resistance (now $3R_v$).

With the bridge in balance, i.e. with $\Delta R_B = 0$, it is necessary to have $R_1 = \infty$ (and also $R_2 = \infty$). Since the variable resistances cannot be made infinitely great, the converter cannot operate without detuning of the resistance bridge and thus will not work at zero frequency. A resistance R_{k0} is therefore connected in parallel with one of the bridge resistances to give an initial detuning so that the oscillator will oscillate at a finite frequency when $\Delta R_B = 0$.

An experimental version of the $\Delta R/R \rightarrow \Delta f/f$ converter (*fig. 8*) has been designed for operation with one or more strain gauge bridges operating in parallel. The basic element is the resistance-to-period converter described above. The converter has been designed so that the output frequency varies between 200 and 1200 Hz. The characteristic of the converter (with f_0 in Hz) then becomes:

$$f_0 = 10^6 \frac{\Delta R_B}{R_B} + 200,$$

where

$$\frac{\Delta R_B}{R_B} = 0 \ldots 10^{-3}. \quad \ldots \quad (9a)$$

If the converter is connected to a strain-gauge load cell, then from eq. (9a):

$$f_0 = 10^3 \frac{G}{G_n} + 200, \quad \ldots \quad (9b)$$

where G is the load and G_n is its maximum permitted value.

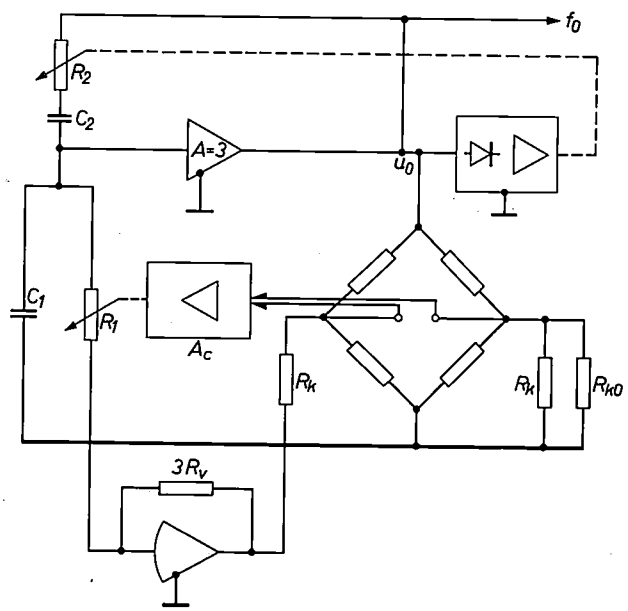
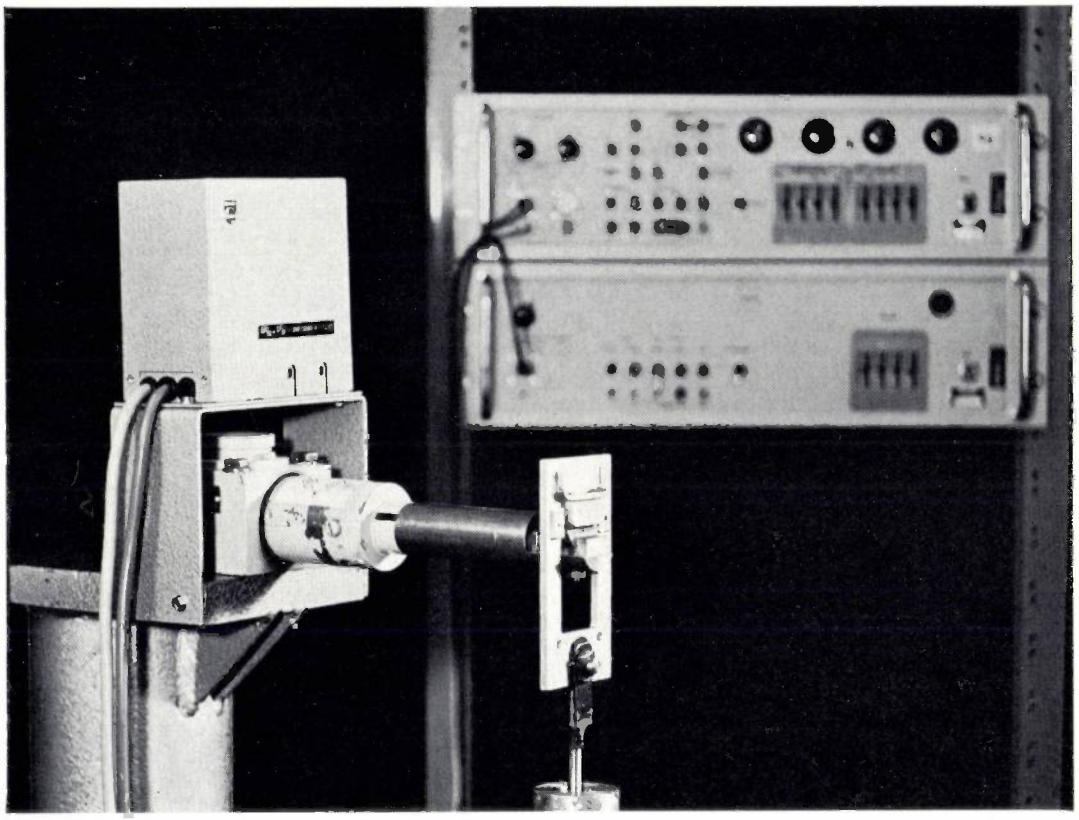
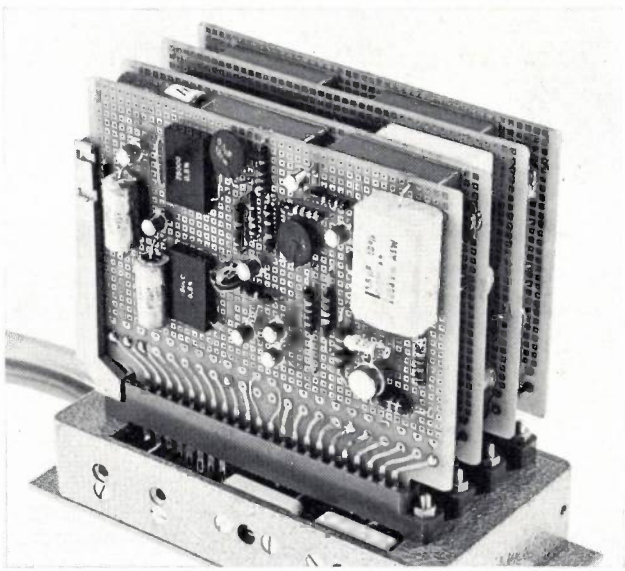


Fig. 7. A small relative change in resistance or bridge detuning $\Delta R_B/R_B$ is converted into a proportional and larger relative frequency variation $\Delta f_0/f_0$ by combining the compensating circuit of fig. 6c with the resistance-to-period converter of fig. 2.



a



b

Fig. 8. *a*) Experimental version of a $\Delta R/R \rightarrow \Delta f/f$ converter like that of fig. 7, connected to a strain-gauge load cell type PR 9228/05 C; the output frequency is measured digitally by means of a pulse counter and a frequency divider (on the right). (These units are shown out-of-focus in the picture to indicate that they can be set up a long way off.)
b) The converter with cover removed.

There is an upper frequency limit because the bridge operates with an a.c. voltage of variable frequency, while the output voltage Δu of the bridge is a very small differential-mode signal which has a common-mode signal of half the bridge supply voltage superimposed upon it. The control amplifier used here consists of an a.c. differential amplifier with a gain of about 100 dB and a rejection factor (ratio of amplification of differential-mode and common-mode components [5]) of about 140 dB in the operating frequency range, a rectifier controlled by the output voltage of the oscillator, and a d.c. voltage amplifier with *RC* chains similar to those in the oscillator control circuit. R_1, like the resistance R_2, is a photoresistor controlled by a lamp which is driven by this control amplifier.

Measurements of the converter with a strain-gauge

[5] See for instance: G. Klein and J. J. Zaalberg van Zelst, several articles on difference amplifiers, Philips tech. Rev. **22**, 345-351, 1960/61; **23**, 142-150 and 173-180, 1961/62; **24**, 275-284, 1962/63.

load cell as a transducer (PR 9228/05 C; see *fig. 9*) showed that the linearity error did not exceed ± 0.01% relative to the maximum load. In the temperature range from 25 to 65 °C the temperature error was confined to a zero drift of —0.03%, again relative to the maximum load, and a change in sensitivity of ± 0.03% (*fig. 10*). A change of ± 10% in the operating voltage introduced an error of less than ± 0.02% of the measured quantity.

The output frequency was measured with a digital pulse counter. The counting interval was determined, with the aid of a digital frequency divider, from the frequency of a reference oscillator mounted inside the case of the converter (see above).

Converters which make use of the principle that has been described can be used not only with the very insensitive strain gauges, they can be used with all kinds of resistive transducers, and also with capacitive and inductive transducers (differential inductance, differential transformer).

It is also possible to design conventional oscillator circuits in which the resistive transducer itself forms an element of a frequency-determining network and in which small changes in resistance lead to large variations in frequency. However, the principle which we have described here has the advantage that it is essentially linear. At the same time the sensitivity to reactive effects, e.g. capacitive detuning of a strain gauge bridge by stray capacitances, is considerably smaller, since the bridge does not form part of the frequency-determining network and the effect of reactive components in the output voltage of the bridge can be reduced by phase-selective rectification. Since the bridge supply is a.c., the conversion is independent of drift voltages and drift currents in the amplifiers.

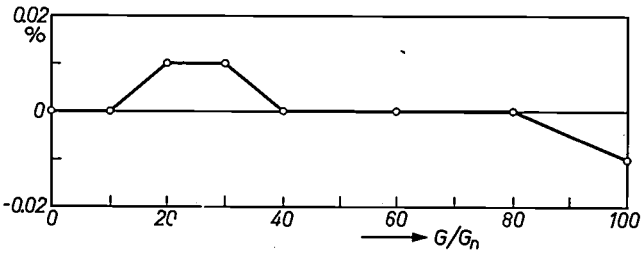


Fig. 9. Linearity error (relative to the maximum load) of a frequency analogue measuring system consisting of the strain-gauge load cell PR 9228/05 C and the $\Delta R/R \rightarrow \Delta f/f$ converter.

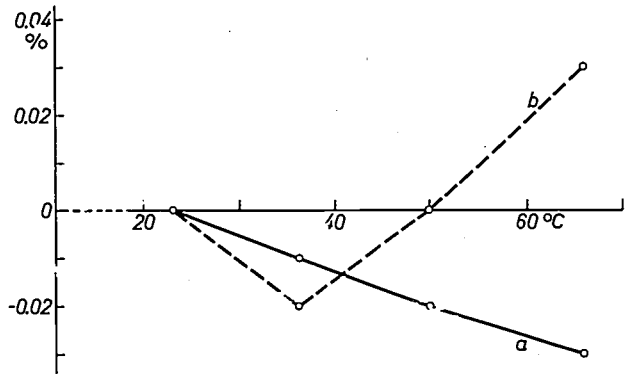


Fig. 10. Temperature drift *a*) of the zero, relative to the maximum load, and *b*) of the sensitivity of the frequency-analogue measuring system.

The converter which we have described in this section has much the same limitations in dynamic measurements as a resistance-to-period converter: the limiting frequency is about 1 Hz.

Summary. Measured intelligence can conveniently be transmitted by means of frequency-analogue signals, in which the information is contained in the frequency. Such signals are insensitive to interference on the transmission channel, and are readily combined with digital techniques. The representation of measurement quantities by frequency-analogue signals requires the use of special converters. These can take the form of oscillators whose frequency is uniquely and stably determined by the output signal of amplitude-analogue passive transducers. Harmonic oscillators are particularly suitable for this kind of application. To simplify further processing, it is desirable to have a *linear* relation between the frequency (or period) and a single circuit parameter, e.g. a resistance value. An essentially linear characteristic can be achieved by applying a new kind of amplitude control: when this method is applied to an *RC* oscillator with two *RC* chains, a change in one of the resistances, controlled by the measured quantity, initiates a change in the other of such a magnitude that the two *RC* chains always vary to exactly the same extent.

For only a small relative change in resistance a large relative variation in frequency can be achieved if the oscillator is connected to a self-balancing electronic compensator. A description is given of a device which makes use of this principle and has been specially designed for use with a strain gauge bridge. Measurements of departure from linearity, temperature drift, and the effects of varying the supply voltage in the experimental versions of these converters indicated errors of 0.01 to 0.05%.

PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 7

The Philips helium liquefier

G. J. Haarhuis

The single-stage gas-refrigerating machine, described in this journal in 1954, provided a simple and efficient means of producing liquid air and nitrogen. With the advent of the two-stage machine, described in 1964, the liquefaction of hydrogen became a simple matter. Another step forward has now been taken with the construction of an efficient helium liquefier based on a pair of two-stage gas-refrigerating machines. This liquefier, the product of close cooperation between a design team in the Industrial Equipment Division and a research team at the Philips Research Laboratories, is simple to operate and can process very impure helium gas.

With the appearance in 1963 of a gas-refrigerating machine which gave high-efficiency refrigeration down to 20 °K, and could even reach 11 or 12 °K [1], the stage was set for the design of a helium liquefier based on a gas-refrigerating machine. The inversion temperature of the Joule-Thomson effect for helium lies between 40 °K and 50 °K, so that it must be possible to achieve the last refrigeration step (down to 4.2 °K) by means of an expansion (or throttling) process, just as in existing helium liquefiers. The strongly increasing demand for liquid helium for scientific and technical applications, and the advantages associated with the use of gas-refrigerating machines, prompted the decision to embark upon the development of a liquefier of this kind. Partly through the invention at Philips Research Laboratories of a new type of heat exchanger with a very high thermal efficiency [2] and the expansion ejector [3], and partly through the application of a new type of compressor equipped with pistons using a rolling diaphragm [4], a helium liquefier has emerged from this development work which possesses a number of very attractive features. The thermodynamic design is by Ir. G. Prast of the Philips Research Laboratories.

The new liquefier system, which is very compact and (once started up) automatic in operation, delivers about 10 litres of liquid helium per hour, and if the helium contains 2% of air it can still operate continuously for 100 hours without requiring cleaning. Even if the percentage of air is much higher the operation of the liquefier remains unimpaired, and an air content of as high as 10% is in fact permissible for a limited time. If the helium used contains less than 2% of air, the period of continuous operation is well over 100 hours. Once the liquefaction process has started, all the operator has to do is to ensure that the liquefier is regularly supplied with helium gas, and to replace full Dewar vessels by empty ones. The efficiency of the system is high: the production of one litre of liquid helium only uses up 2.8 kWh of electrical energy, including the energy required for removing the 2% of air. Unlike other liquefiers, this one does not require a supply of liquid nitrogen for precooling or purification of the gas supply.

Ir. G. J. Haarhuis is with the Philips Industrial Equipment Division (PIT), Eindhoven.

[1] G. Prast, A gas-refrigerating machine for temperatures down to 20 °K and lower, Philips tech. Rev. **26**, 1-11, 1965.

[2] G. Vonk, A compact heat exchanger of high thermal efficiency, Philips tech. Rev. **29**, 158-162, 1968 (No. 5).

[3] J. A. Rietdijk, The expansion ejector, a new cryogenic device, Philips tech. Rev. **28**, 243-245, 1967 (No. 8).

[4] For the rolling diaphragm, see: J. A. Rietdijk, H. C. J. van Beukering, H. H. M. van der Aa and R. J. Meijer, A positive rod or piston seal for large pressure differences, Philips tech. Rev. **26**, 287-296, 1965.

Construction and operation

Fig. 1 shows a simplified diagram of the liquefier. A compressor compresses the helium gas at room temperature from 2.5 bar to 20 bar (gas flow rate 2.5 g/s; 1 bar is about 1 atm). The compressed gas passes the five gauze-type heat exchangers H_1 to H_5 and the precoolers R_1 to R_4 of a pair of two-stage gas-refrigerating machines. In the last heat exchanger a temperature of 7 °K is reached. The helium leaving the system in a liquid state is replaced by helium gas supplied through the tube S_1 at a pressure of 20 bar. This gas flow first passes through all five heat exchangers before joining the main flow (at *Con*). The cold gas expands in the expansion ejector *EE* to 2.5 bar and is cooled down further by the Joule-Thomson effect. A large proportion of this gas flows through the five heat exchangers H_5 to H_1 back to the compressor. On its way it gives up nearly all its cold to the two other flows passing through the heat exchangers. The part of the cold gas that does not flow back undergoes a second expansion in the expansion valve *JK*, this time to a pressure of 1 bar, and a certain fraction liquefies during this expansion. Gas and liquid then flow to vessel *C*, and the gas then returns from *C* to the main flow via the expansion ejector. The whole process described here is continuous.

During the period immediately after the system is switched on, no helium is supplied through S_1. Gas circulates then only in the middle and left-hand tubes shown in H_1 to H_5 in fig. 1. The system then gradually cools down. When the temperature between H_2 and H_3 has dropped to 80 °K, this starting period is considered to be over and the gas supply starts through S_1. The decrease of pressure in the system as a result of the cooling during the starting period is compensated by the supply of pure helium through S_2.

In the actual system the vessel *C* is a Dewar flask in which the liquid helium produced is collected and stored. The components inside the chain-dotted rectangle are contained in another Dewar flask, the cryostat. This is in communication with *C* by means of a specially constructed connecting system with coaxial feed tubes.

Special features

The chief feature that distinguishes the diagram of fig. 1 from that of any other liquefier is that it includes the expansion ejector *EE* (*fig. 2*). This works both as an expansion valve and as a pump of the water-jet type:

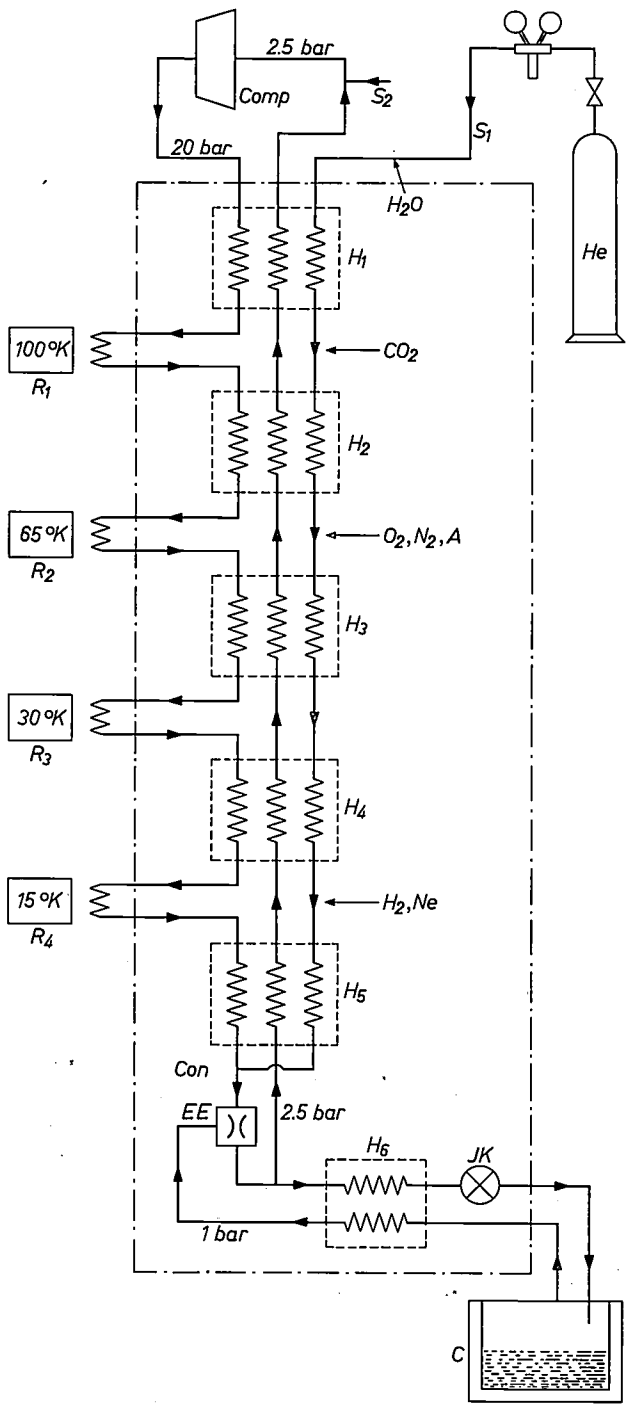


Fig. 1. Simplified diagram of the Philips helium liquefier. *Comp* compressor. R_1 to R_4 four precooling stages obtained from two two-stage gas-refrigerating machines. H_1 to H_5 gauze-type heat exchangers. *EE* expansion ejector. *JK* expansion valve. H_6 heat exchanger. *C* vessel for collecting liquid helium. Fresh gas is regularly supplied to the system through tube S_1; it passes through the heat exchangers H_1 to H_5 and then joins the main stream at *Con*. Impurities in the gas supply are eliminated at the places denoted by the relevant chemical symbol. The part inside the chain-dotted line is contained in a special Dewar vessel.

it draws the helium gas at 1 bar from the vessel C and raises it to a pressure of 2.5 bar. This is at the same time the suction pressure of the compressor, which can therefore be very much smaller than if it had to draw the gas from C. This is important for efficiency and also saves space.

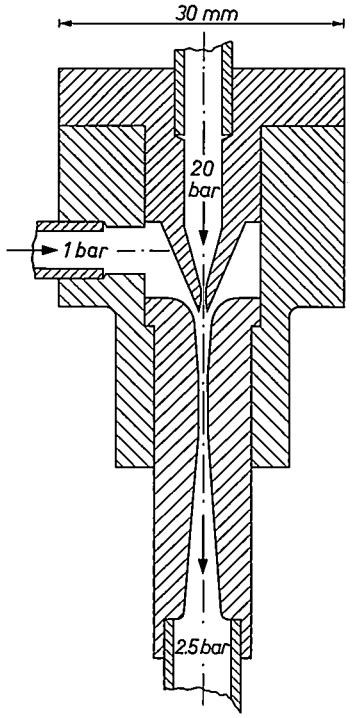


Fig. 2. Cross-section of the expansion ejector. The helium gas flowing from above (20 bar) expands to 2.5 bar, giving cooling through the Joule-Thomson effect. At the same time, through the opening on the left, helium gas of 1 bar is drawn in from the collector vessel (C in fig. 1) and again compressed to 2.5 bar.

The gas flowing from the ejector EE to the expansion valve JK is cooled in the heat exchanger H_6 by the gas flowing back from the vessel C to the ejector. It therefore arrives at JK with a temperature of only 5.2 °K, and because of this low temperature no less than 60 to 70% of the gas is liquefied during the expansion at JK.

The pair of two-stage gas-refrigerating machines with which the liquefier is equipped work at different temperatures and thus drive *four* precooling stages. Two machines have to be used since the output yield with one machine of the conventional type (A20) is less than the amount considered desirable. The use of more than one machine has two other important advantages, however. In the first place, in cooling a flow of gas or liquid the highest efficiency is in theory obtained when cold is supplied to the medium at each temperature in the cooling process. This ideal situation is much better approximated when cold is supplied at four different temperatures — here 100, 65, 30 and 15 °K — than when it is supplied at only one or two. In the second place, when cold is supplied at four temperatures it is possible to remove the impurities from the helium gas

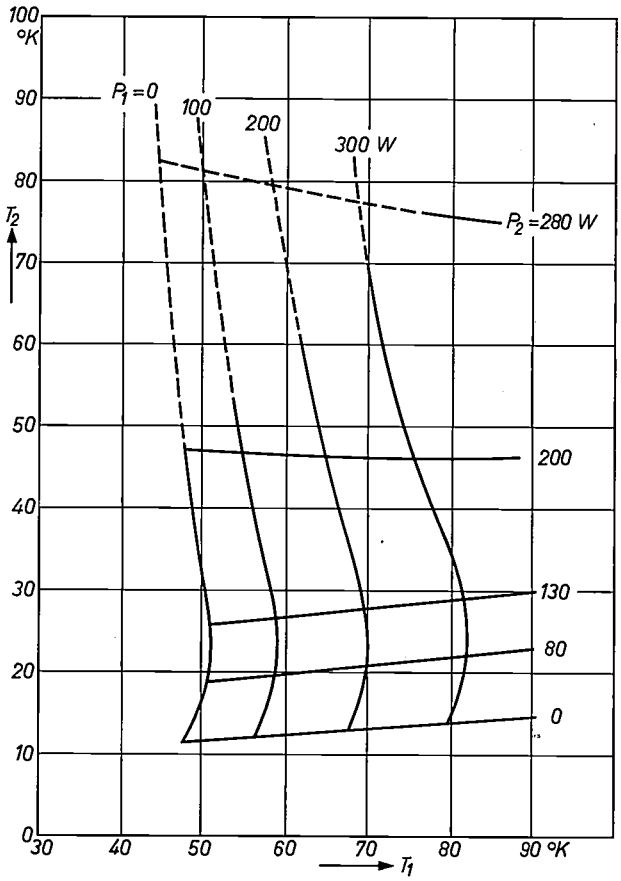


Fig. 3. The relation between the first-stage temperature T_1 and the "head" temperature T_2 of a Philips A20 two-stage gas refrigerating machine which is found when the refrigerating capacity P_1 is varied while that of the head (P_2) is held constant, and vice versa. (These results represent the average of measurements made on a number of machines.) The shape of the curves indicates that variation of P_1 has hardly any effect on the refrigerating capacity delivered by the head at a particular temperature.

(The dashed parts of the curves correspond to situations in which the head is less cold than the other stage; such situations do not of course arise when the machine is in normal use.)

(mainly air) in a very effective way by adsorption. This will be discussed separately in the next section.

A feature of the two-stage machine which we have found very useful is that the cold production of the first stage — i.e. the one with the higher temperature — has hardly any effect on that of the second stage (*fig. 3*). The system can thus be arranged so that a

large part of the cold is delivered at the higher temperature, which improves the efficiency.

We have already mentioned the compressor with rolling diaphragms, specially designed for this liquefier. This compressor, one of a type designed by Ir. H. J. Verbeek of the Philips Industrial Equipment Division, will be the subject of a forthcoming article in this journal. The rolling diaphragms form a hermetic seal between the spaces containing the gas for compression and the spaces containing oil. This is of particular importance: without an oil-free compressor a system like that described here would not be able to work continuously for very long.

Purification of the gas supply

At the temperature of liquid helium all other substances are solids. Impurities in the gas supplied must therefore be removed before this temperature is reached, otherwise they can cause stoppages. In the Philips helium liquefier this is done by *adsorption*, as we noted earlier. With one exception, the adsorbers are all located in the cryostat, each being placed at the coldest possible place consistent with the requirement that the impurity in question is to remain in the gas phase (cf. fig. 1). The new helium liquefier differs in this respect from all existing ones, which require pre-purification of the helium supply to reduce the impurity content to about 0.01 % by volume. This pre-purification requires the use of liquid nitrogen, which is not required in the Philips liquefier.

As can be seen from fig. 1, the adsorber that traps *water vapour* works at room temperature. *Carbon dioxide gas* is trapped by an adsorber situated between the first and second heat exchangers, where the prevailing temperature is about 125 °K. Both adsorbers contain "molecular sieves", and if the helium gas entering through S_1 contains 2 % of air and is saturated with water vapour, they must be regenerated once every 100 hours.

Oxygen, *nitrogen* and *argon* are removed between the second heat exchanger and the third (temperature about 70 °K). This is done by means of two adsorbers consisting of activated charcoal which operate alternately for periods of 40 minutes. While one is working the other is being regenerated, which is done by raising the temperature to about 145 °K and pumping off the desorbed gas. The switch-over is automatic. If the gas supply contains more than 2 % of air, part of it is liquefied in the second heat exchanger. This liquid is collected in a Dewar vessel and from time to time automatically removed from the system (see below).

Finally, between the fourth heat exchangers and the fifth there is another small charcoal adsorber which traps hydrogen and neon; the quantity of these gases contained in the gas supply is usually very small. This adsorber can also work for 100 hours continuously with 2 % of air in the helium supply.

All these adsorbers are located in the supply line (S_1-*Con*) so that the gas circulating in the starting period does not pass through them. For purifying this gas a separate adsorber is included between H_2 and H_3. As the temperature of the system falls, this adsorber traps an increasingly larger fraction of any impurities, and eventually the circulating helium is practically pure.

Technical details

The construction of the new liquefier is illustrated schematically in *fig. 4*. The photograph (*fig. 5*) gives a general view of the system. The cryostat, which contains the five gauze-type heat exchangers, the ejector, the expansion valve and the devices for purifying the gas (except the adsorber for water vapour) are contained inside an evacuated Dewar vessel (1-0.1 torr) of special design. This vessel is divided into three compartments by two refrigerated radiation shields RS_1 and RS_2, coated with silver on both sides; the shields extend into the high-vacuum space between the two walls of the Dewar vessel. Each shield is cooled by a freezer of one of the refrigerating machines; RS_1 has a temperature of about 100 °K, and RS_2 has a temperature of 30 °K. The use of these shields considerably improves the heat insulation of the Dewar vessel. The part of the shield situated between the walls of the Dewar vessel and the part in the open space are made separate as the equipment has to be demountable. The rim of the shield is increased in area by means of a special strip and the gap between the two parts is kept small (about 0.25 mm), thus ensuring sufficient heat transfer from conduction in the gas.

The expansion valve *JK* is automatically controlled. At the top of the cryostat there is a double-bellows system, not shown in fig. 4, which is connected with the line containing the helium that has just passed the expansion valve. The position of the bellows is determined by the difference between the pressure in the line (normally just above 1 bar) and the pressure of the outside air. When the pressure in the line increases, the bellows system closes the expansion valve tighter; when the pressure decreases, the valve opens wider. This simple control method proves eminently satisfactory in practice.

At the centre of fig. 4, next to the air adsorbers, a component *Bl* can be seen which we have not yet mentioned. This starts to operate when the helium gas supplied to the liquefier is so strongly contaminated with air that air begins to condense in the heat exchanger H_2. This liquid air is collected in *Bl*, and automatically blown out again when the liquid reaches a certain level.

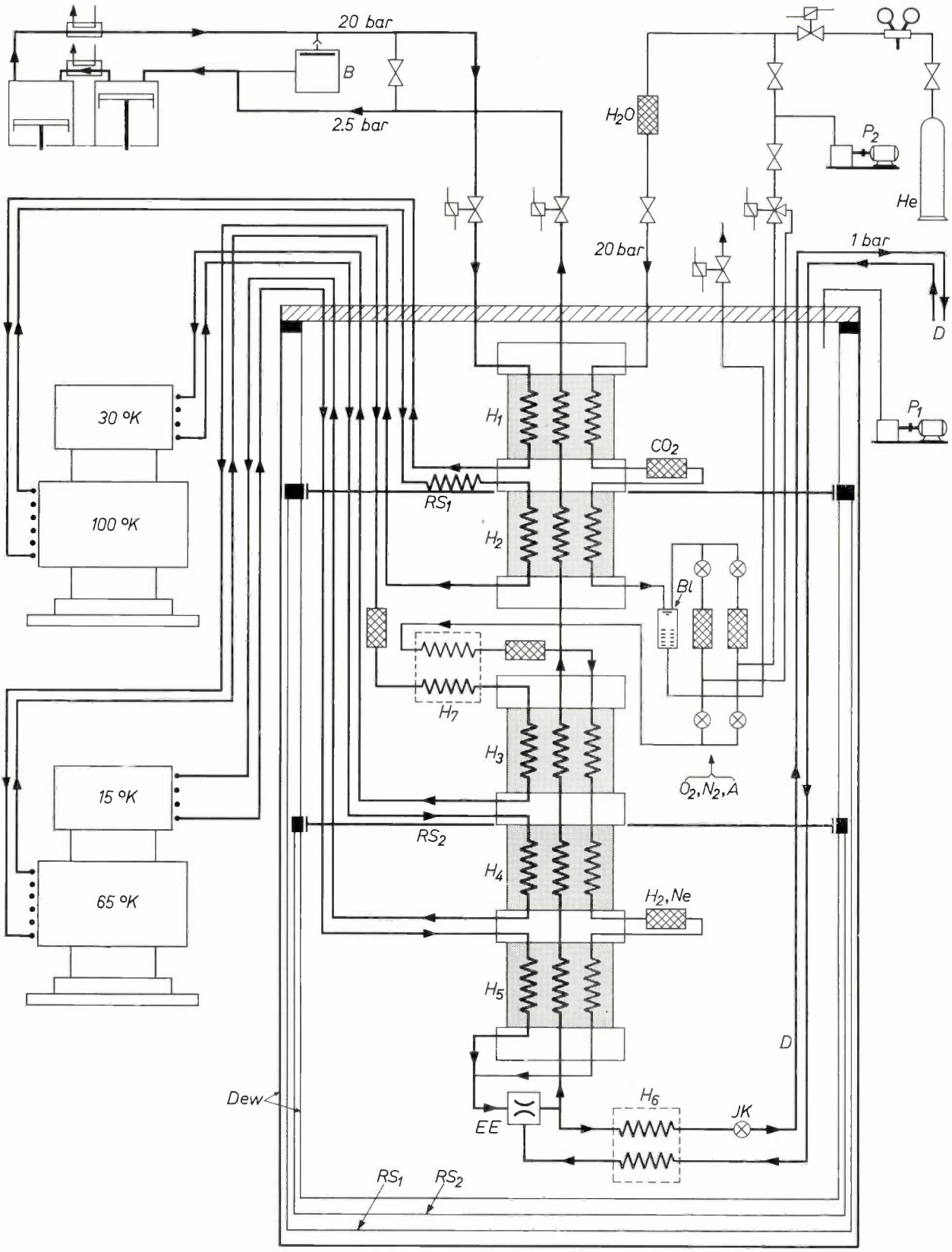


Fig. 4. Complete constructional diagram of Philips helium liquefier. *Dew* walls of Dewar vessel (cryostat). RS_1 and RS_2 refrigerated radiation shields at 100 °K and 30 °K respectively. *D* delivery tube, containing one channel for removing the liquefied helium and one for the returning gas. *B* buffer vessel. P_1 and P_2 vacuum pumps. The various adsorbers for removing impurities are again denoted by the chemical symbol of the gas which they adsorb. *Bl* device for collecting and automatically blowing off liquid air; this comes into operation only when the impurity content of the helium gas is very high (see also fig. 6). The other letters have the same significance as in fig. 1.

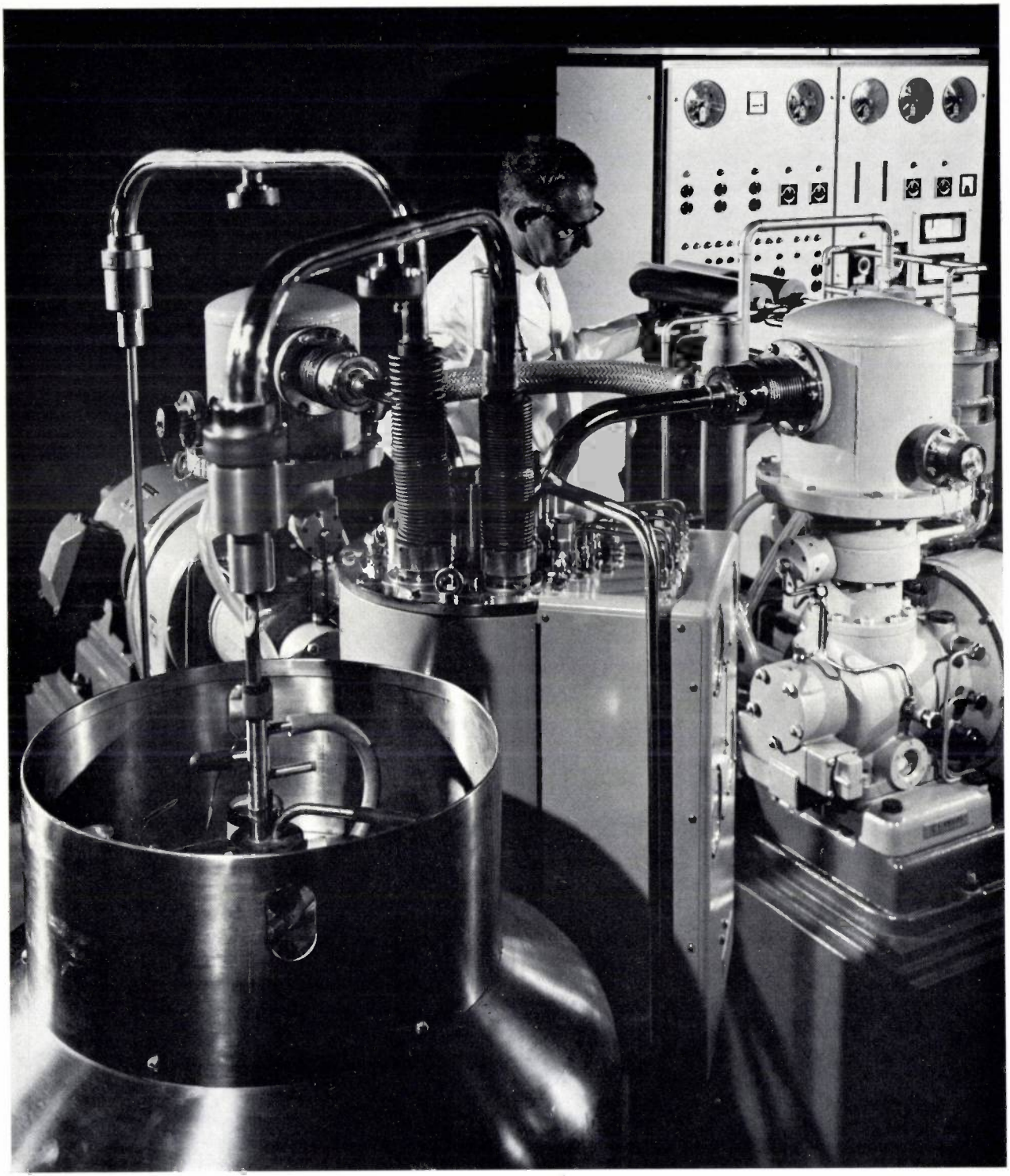


Fig. 5. The Philips helium liquefier. One Philips two-stage gas-refrigerating machine is to be seen on the left in the background and another can be seen at the far right. A Dewar vessel in which the liquid helium is collected and removed can be seen in the foreground. The compressor is behind the right-hand refrigerating machine. The cryostat is at the centre of the picture; the cabinet attached to it on the right contains meters for registering temperatures and pressures.

Details of this automatic system are given in *fig. 6*, together with the air adsorbers and the corresponding valves, feeder tubes, etc. The vessel *Bl* contains a hollow pin, which is open at the bottom and is fitted inside with a reed relay. Floating on the liquid air around the pin is a ball *F* with a permanent magnet inside it. If the liquid level rises there comes a moment at which the magnet is level with the relay reeds, which are ferromagnetic. The

reeds then spring together, the valve V_3 opens and remains open until F has gone down far enough for the relay to open again. Things are so arranged that no helium gas can be blown off. The air adsorbers A_1 and A_2 are brought into the circuit alternately by means of the pneumatically operated double-acting valves V_1 and V_2. When an adsorber is being cleaned, the desorbed gas escapes through the tube shown in the drawing at the top, which leads to a vacuum pump.

A_4 is the adsorber which purifies the gas circulating in the system during the period immediately after the system has been switched on.

The last part of the new liquefier we shall discuss is the *delivery tube*, i.e. the tube (D in fig. 4) through which the liquid helium leaves the cryostat and flows to the Dewar (*fig. 7*). In the sleeve *1*, which can be seen in the foreground of fig. 5, there are two coaxial tubes *2* and *3*, which are surrounded in a part of the delivery

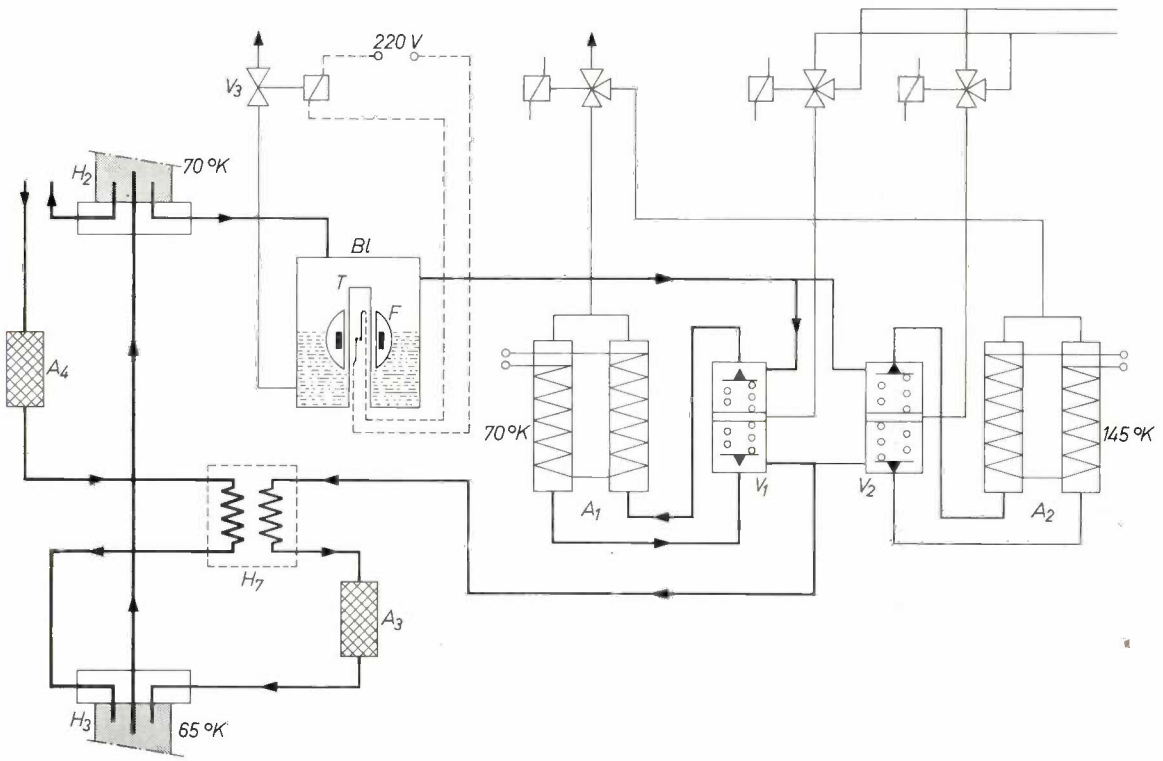
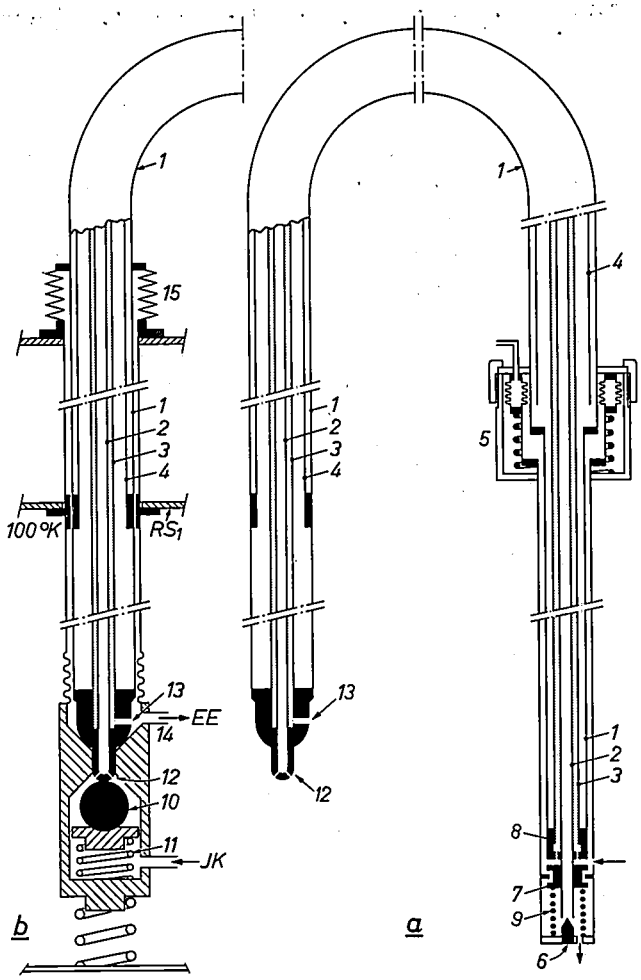


Fig. 6. Apparatus for removing nitrogen, oxygen and argon from the helium gas supply. H_2 and H_3 gauze heat exchangers (see fig. 1). A_1 and A_2 groups of two adsorbers connected in series, one of which is in operation while the other is being regenerated. The adsorbed gas escapes through the tubes shown at the top, which lead to a vacuum pump. V_1 and V_2 are double-acting pneumatically operated valves for switching over from one group of adsorbers to the other. A_3 adsorber for use during the short transitional period after switching over. H_7 heat exchanger. *Bl* collector vessel for the air condensed in H_2, with a device (F, T) for automatically opening the valve V_3.

Adsorber A_3 is also an air adsorber, which comes into operation immediately after A_1 has been switched over to A_2 or vice versa. The adsorber just switched in is then still relatively warm, and during the first few seconds does not trap anywhere near all the air that passes. This air is adsorbed by A_3, which is always sufficiently cold because the gas flow from $A_{1,2}$ (0.35 g/s) first passes the heat exchanger H_7 through which the main flow (2.5 g/s) also passes. The time during which the liquefier can operate continuously is mainly determined by the capacity of A_3.

tube by a refrigerated radiation-shield *4*, silvered on the outside. In the space between *3* and *1* there is a high vacuum. Gas and liquid flow through the inner coaxial tube to the collector vessel (in fig. 1 the part of the line between *JK* and *C*), and gas flows back to the cryostat (line between *C* and *EE* in fig. 1) through the outer tube (shaded). The radiation shield *4* is connected with one of the radiation shields of the cryostat, i.e. the shield at about 100 °K. The end of the delivery tube projecting into the collector vessel (on the right in the figure) contains valves *6* and *7*, which can be pneumatically

Fig. 7. *a*) Delivery tube; the end at the right reaches into the vessel where the liquid product is collected, the other end into the cryostat. *1* outside wall. *2* central tube for the removal of liquid and gas. *3* tube through which gas flows back to the cryostat, coaxial with *2*. *4* radiation shield (100 °K) coaxial with *2* and *3*. *5* bellows system for closing the tube, working pressure 2.5 bar; the pin *6* closes tube *2*, the ring *7* closes the holes in block *8* through which the gas can enter tube *3*. *9* compression spring. *b*) Connection of the tube to the cryostat. As long as the tube is not in the lowest possible position, a ball valve, with ball *10* and spring *11*, prevents the escape of helium gas and liquefied helium coming from the expansion valve (*JK* in fig. 1 and fig. 6). When the tube enters the cryostat to the fullest possible extent, it presses the ball *10* downwards and the helium then has access through holes *12* to the central tube *2*. The gas returning to the cryostat can flow to the expansion ejector (*EE* in fig. 1 and fig. 6) via hole *13* and tube *14*. *15* rubber bellows; this enables the tube to be raised for changing the container, without resulting in an open connection between the ejector and the outside air.



operated by means of the bellows system *5*. The other end of the tube reaching into the cryostat can easily be taken out, which automatically closes the line which supplies the liquid and gas to the central tube. A special method of manufacture is used for making the complex assembly of coaxial walls *1*, *2*, *3* and *4*: the structure is fabricated in rectilinear form, all the spaces are filled with water, and it is then cooled in a suitable way to the temperature of liquid nitrogen and bent to shape.

Summary. The article describes an easily operated, highly automatic helium liquefier which makes use of Joule-Thomson effect cooling and has a pair of two-stage gas-refrigerating machines for precooling. The installation produces about 10 litres of liquid helium an hour at a high efficiency (power consumption 2.8 kWh per litre). The helium gas does *not* have to be carefully purified beforehand. If the gas contains 2% of air, the system can work for 100 hours continuously; a much higher air content can be tolerated for short periods of operation. Factors contributing to the high efficiency are the availability of *four* precooling temperatures (approx. 100, 65, 30 and 15 °K) and the use of special gauze-type heat exchangers and an expansion ejector. The expansion ejector permits the use of a relatively small compressor (suction pressure 2.5 instead of 1 bar, and compression 20 bar). This compressor is equipped with rolling diaphragms. All impurities are removed by adsorption, each at the most appropriate temperature. The connection between the cryostat and the vessel in which the liquid helium is collected is formed by a specially designed delivery tube, containing coaxial feeders for the liquid product and the returning gas.

High-speed printers for numerical data processing equipment

J. Borne

In the present "fan-out" of computer technology much attention is being given to the development of methods of printing — with particular reference to sheer speed, economy, quality, legal requirements, etc. Two interesting principles have been considered and translated into hardware at L.E.P. (Laboratoires d'Electronique et de Physique Appliquée) near Paris.

Introduction

The production of a printed document, whatever its nature and purpose, is divided into three basic phases: editing, typesetting, and printing.

Editing for the press, and in particular for book publication, is invariably a relatively slow operation in which the material is checked to ensure that the original thought or information has been communicated faithfully.

The *typesetting* requires a similar degree of attention to the design of the printed page, both for the make-up or layout and the choice of characters. The speed of this operation is not generally of primary importance, and absolute reliability is not essential owing to the possibility of making improvements and corrections.

The need for speed and reliability only arises in the third phase, in the actual *printing* of the document. Design considerations here are confined purely to the choice of inks and paper.

In numerical data processing equipment however, the machine itself "edits" the document at very high speed, and no human intervention is required to verify the correctness of the information transmitted. To speed up the "typesetting" process, the page make-up is determined during the "editing" phase, and the choice of characters is usually limited to a single series of capital letters, numerals and essential characters or symbols. Setting and printing are virtually simultaneous, and since there is no chance of inserting corrections between the two phases it is essential that these operations should be extremely reliable.

All this may seem obvious to readers familiar with the problems of data processing, but it is useful to recall such matters here in order to evaluate the characteristics of the unfamiliar printing processes which we shall describe. In fact, printing equipment based on these processes should be very suitable for this kind of application.

J. Borne, Ing. I.R.G., is with Laboratoires d'Electronique et de Physique Appliquée (L.E.P.) at Limeil-Brévannes (Val de Marne), France.

The detailed descriptions of equipment that will be given here refer to experimental models built in the laboratory, as an application of one of these processes which appears to combine the maximum number of advantages: speed, reliability, versatility, and the use of a standard paper with the possibility of printing simultaneous copies.

Formation of characters by a dot mosaic

In printers used for information processing the outline of each character is generally formed by a continuous line, of practically constant width, raised in relief on the printing element. This continuous line may be perfectly reproduced by a series of dots sufficiently numerous and close together, but a satisfactory representation can also be obtained when the number of these dots and the positions they can occupy are limited. *Fig. 1* shows some characters formed by dots arranged in a mosaic of 7 rows and 5 columns.

In spite of the small number of points, the definition of this mosaic permits the formation of the 64 characters which are sufficient for carrying the information processed in automatic equipment. This limited definition has been chosen for our experiments, but it can easily be augmented by increasing the number of rows

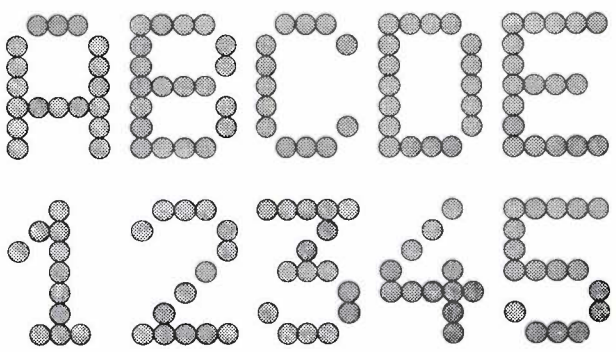


Fig. 1. Formation of some characters in a dot mosaic of 7 rows and 5 columns.

and columns if a larger number of different characters is required, or if a higher quality of presentation is thought to be necessary.

In this process of forming characters the whole of the printing is done with only a single basic element, the *dot*, and "setting" is limited to the positioning of this one element. We shall show how this can be done at very high speed and with very great reliability by electronic means. First, however, we shall describe two printing processes based on the use of dots.

Dot-printing by a thermal process

In the first process the printing element consists of a piece of conducting material arranged as a bridge between two relatively large metal blocks, used as current leads.

When a current pulse passes through this element, the current density in the metal blocks is low but assumes a very high value in the bridge, whose temperature rises rapidly due to the Joule heating. The heat developed can be used to blacken a sheet of thermosensitive paper held against the bridge, thus producing a black dot of the same size as the bridge.

We shall present here some calculations and experiments that clearly demonstrate the possibilities — and limitations — of this printing process.

The energy dissipated is the same at every point of the bridge, but, as it is produced, the heat flows by thermal conduction to the metal blocks which are large enough to dissipate the heat without themselves becoming appreciably hotter. It may thus be assumed that the temperature at the ends of the bridge is constant and equal to the temperature T_0 of the metal blocks.

Let L be the length of the bridge, c the specific heat of the material per unit volume, γ the thermal conductivity and w the power dissipated per unit volume. The temperature T at every point of a normal cross-section (abscissa x, see *fig. 2*) as a function of time t is given by the heat-diffusion equation:

$$c\frac{\mathrm{d}T}{\mathrm{d}t} = \gamma\frac{\mathrm{d}^2T}{\mathrm{d}x^2} + w(t).$$

The power dissipated in the bridge as a whole is $W = wAL$, where A is the surface area of the cross-section. Introducing the quantities τ and Θ, defined by:

$$\tau = \frac{cL^2}{\gamma} \quad \text{and} \quad \Theta = \frac{wL^2}{\gamma} = \frac{WL}{\gamma A},$$

we may write the equation thus:

$$\tau\frac{\mathrm{d}T}{\mathrm{d}t} = L^2\frac{\mathrm{d}^2T}{\mathrm{d}x^2} + \Theta(t).$$

If at time $t = 0$ a current capable of developing a constant power W is suddenly initiated in the bridge,

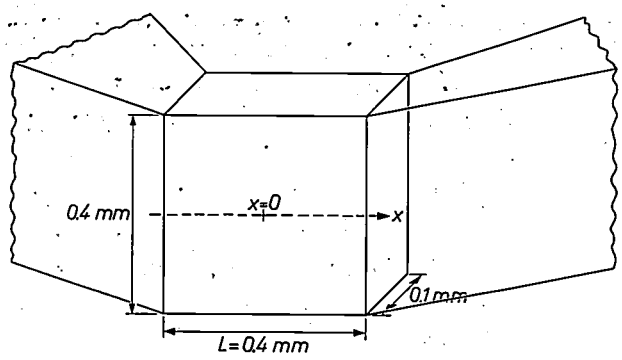


Fig. 2. Thermal process: configuration of conductor (silicon) arranged in the form of a bridge between two current leads. Thermal conductivity $\gamma = 0.84$ watt/cm s deg; specific heat $c = 1.76$ joule/cm³ deg.

the variation of the temperature at the centre of the bridge is given to a good approximation by the first term of the series expansion of the solution of the differential equation [1]:

$$T - T_0 = \frac{4\Theta}{\pi^3}(1 - \mathrm{e}^{-\pi^2 t/\tau}).$$

When the current is switched off after a time t_1, the heat continues to flow to the metal blocks and the temperature at the centre of the bridge decreases in accordance with the equation:

$$T - T_0 = \frac{4\Theta}{\pi^3}(1 - \mathrm{e}^{-\pi^2 t_1/\tau})\,\mathrm{e}^{-\pi^2 t/\tau}.$$

The temperature curve at the centre of the bridge, shown by the oscillograms in *fig. 3*, is therefore represented for every current pulse by two exponentials which have the same time constant τ/π^2.

In our experiments the conducting bridge was a piece of silicon whose dimensions and physical constants are given in fig. 2. The resistance of this element can be fixed at a value of 10 ohms by doping, and a current pulse of 1 A then dissipates in it a power of 10 W. Under these conditions the maximum temperature increase is:

$$T - T_0 = \frac{4\Theta}{\pi^3} = \frac{4WL}{\pi^3\gamma A} = 160\ ^\circ\mathrm{K},$$

and the time constant is:

$$\frac{\tau}{\pi^2} = \frac{CL^2}{10\gamma} = 0.3\ \mathrm{ms}.$$

Fig. 4 shows the arrangement used for checking these results. In the experiments the spacing (from leading edge to leading edge) in a train of pulses was 3 ms, the pulse duration was 1 ms and the energy dissipated during

[1] H. Dormont of L.E.P. gave us valuable assistance with these calculations.

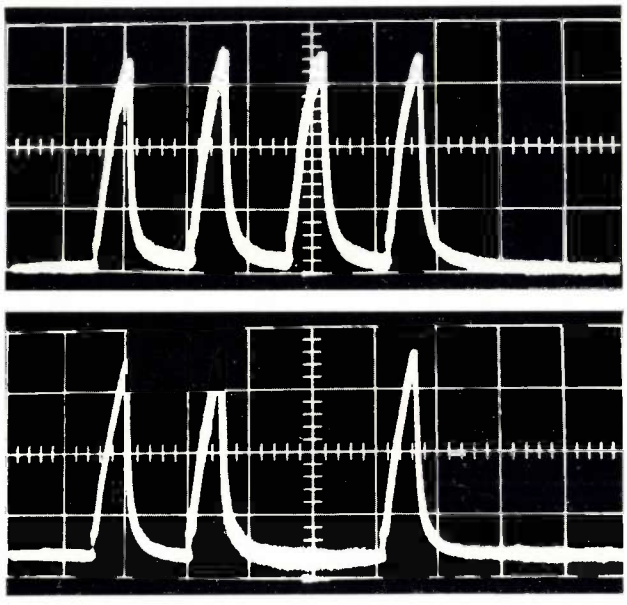
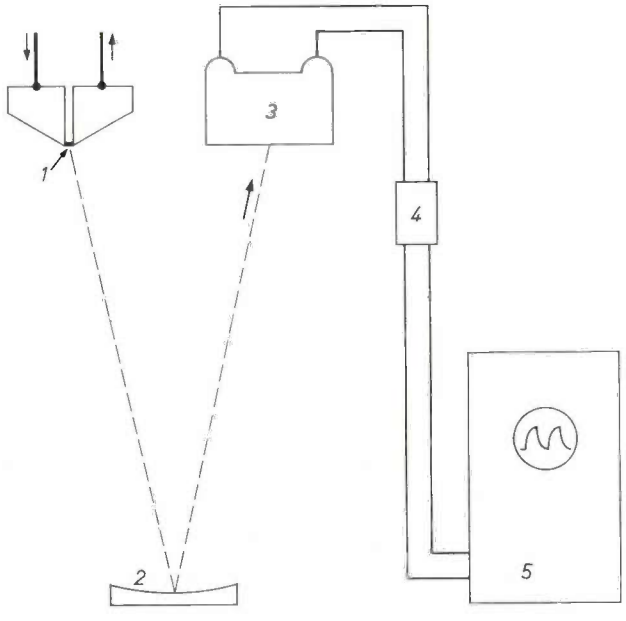


Fig. 3. Oscillograms of the heat radiated by the silicon bridge used in the arrangement of fig. 4 (horizontal scale: 1 square for 2 ms).
a) Recurrent trains of 4 pulses.
b) Recurrent trains of 3 pulses.

a pulse was 10 mJ. It can be seen from the oscillograms that the time constant is short enough for the temperature of the bridge to have returned to a value which is effectively equal to the steady-state temperature before each pulse, and the temperature reached is not affected by the preceding pulse. These elements may therefore be used for printing dots at a rate of 330 per second.

Fig. 5 shows an assembly of 7 heating elements arranged in a vertical column, and an example of the printing results obtained with this process. When this assembly is moved over the surface of a sheet of thermosensitive paper, the desired characters are formed by energizing the elements at the appropriate times. This is done by means of selecting and actuating circuits similar to those we shall describe in the section dealing with the design of a printer using the other process of dot formation.



←
Fig. 4. Arrangement for checking the operating of a thermal printing element. The heat radiated by the bridge *1* when a current pulse passes through it is reflected and focused by a spherical mirror *2* on to an indium antimonide cell *3*, cooled to the temperature of liquid nitrogen. The voltage delivered by the cell is amplified (*4*) and applied to the vertical deflection plates of a cathode-ray oscilloscope *5*.

Dot-printing by a mechanical process

The second of the two processes referred to in the introduction is one whose applications have been studied more thoroughly. In this process the printing element is formed by a needle which runs in a series of perforated plates and is held in its position of rest by a return spring (*fig. 6*). An electromagnet with a plunger core pushes the needle towards the paper when a current pulse passes through its winding, and the impact of the needle causes ink contained in an inking ribbon to be transferred to the paper.

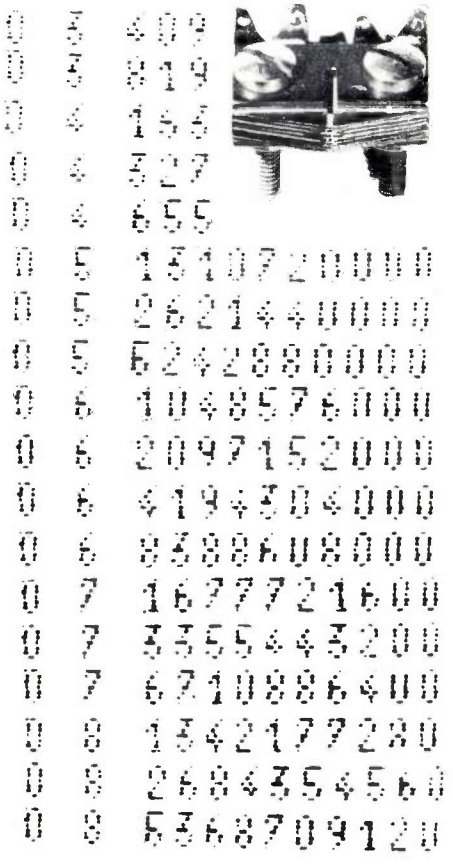


Fig. 5. Thermal printing unit containing 7 printing elements, and figures (results of calculations) printed on thermosensitive paper at a speed of 40 numerals per second in a 4×7 dot mosaic.

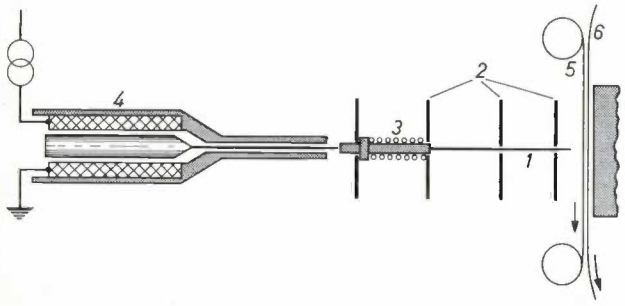


Fig. 6. Printing element in mechanical process. *1* needle. *2* guide-plates. *3* return spring. *4* electromagnet. *5* inking ribbon. *6* paper.

The movement of the needle may be calculated from the general equation:

$$\frac{d^2x}{dt^2} = \frac{F}{m},$$

where F is the force developed by the electromagnet and m the mass of the moving parts (core and needle). The current pulse, whose duration we shall set at 1 ms, is supplied by an electronic switch of low internal resistance, which thus acts as a constant voltage generator V, and the current is given by:

$$i = I_0 (1 - e^{-t/\tau}),$$

where $I_0 = V/R$ and $\tau = L/R$,

and R and L are the resistance and inductance of the winding. The force developed by the electromagnet is given by:

$$F = Ki^2,$$

where K is a characteristic constant of the electromagnet.

We assume that the displacement of the plunger core is small enough for us to neglect the variation of the air gap during the movement, and to simplify the calculations we shall approximate the current curve to the straight line $i = \alpha t$ connecting the points on the exponential curve corresponding to $t = 0$ and $t = 1$ ms; we then have $F = K(\alpha t)^2$ and hence:

$$\frac{d^2x}{dt^2} = \frac{K\alpha^2 t^2}{m},$$

from which we find by integration:

$$v = \frac{dx}{dt} = \frac{K\alpha^2 t^3}{3m} + v_0,$$

and then:

$$x = \frac{K\alpha^2 t^4}{12m} + v_0 t + x_0.$$

The curves representing the movement of the needle may be calculated from this equation and are shown in *fig. 7*. The various constants required were measured with the aid of an experimental model, and are indicated in the caption to the figure.

When the voltage is applied to the coil, the force developed by the electromagnet starts to increase without causing any movement of the core, until, at the time t_1, it reaches the value F_r, equal to the force exerted by the return spring:

$$t_1 = \sqrt{\frac{F_r}{K\alpha^2}};$$

for $F_r = 0.25$ newtons we have $t_1 \approx 200$ μs. The initial conditions for the movement of the needle: $v_0 = 0$, $x_0 = 0$, hold at this moment. Therefore, we obtain:

$$v = \frac{K\alpha^2}{3m}(t - t_1)^3,$$

$$x = \frac{K\alpha^2}{12m}(t - t_1)^4,$$

and at the end of the current pulse, i.e. for $t - t_1 = 800$ μs, we have:

$$v \approx 0.5 \text{ metres/second},$$

$$x \approx 0.08 \text{ millimetres}.$$

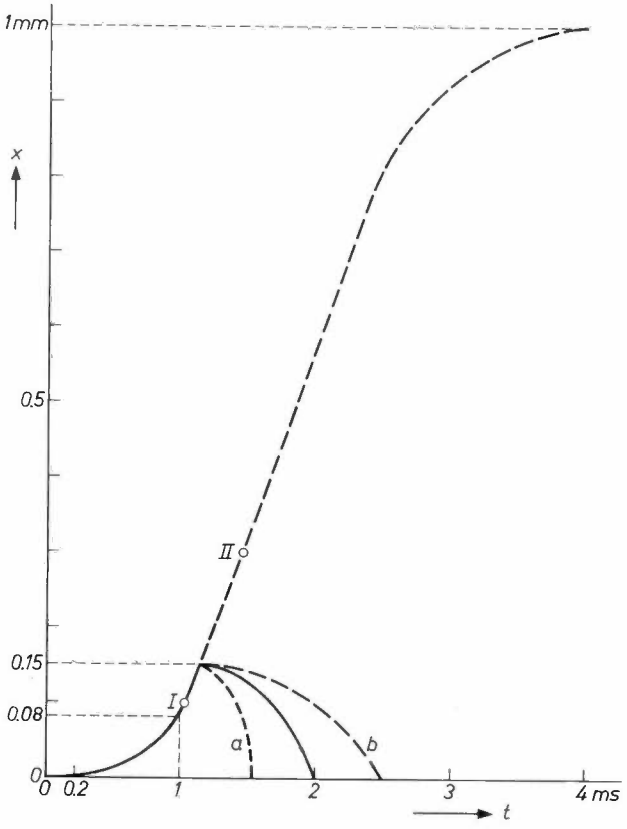


Fig. 7. Movement $x(t)$ of the moving parts for a current pulse of 1 millisecond. The curves are calculated from the equation derived in the text and using values measured with the experimental model: $K = 10$, $\alpha = 700$ A/s and $m = 2$ g. Points *I* and *II* correspond to times measured on the oscillograms in fig. 9 when the needle travels 0.1 and 0.3 mm respectively.

If the moving parts were then able to continue in free motion, the motion could be represented by a parabola as shown in fig. 7. (The force exerted by the return spring remains practically constant over the small excursion made by the needle.) The vertex of the parabola would have the co-ordinates:

$$x = \tfrac{1}{2}\, mv^2/F_r \approx 1 \text{ mm},$$

$$t = mv/F_r \approx 4 \text{ ms}.$$

In fact, however, the movement of core and needle is stopped when the needle strikes the paper, and the moving parts are returned to their position of rest through the action of the return spring and the rebound which occurs after the needle hits the ribbon and paper. This return movement may again be represented by a parabolic arc which will be between the limiting curves *a* and *b* in fig. 7, which correspond to perfectly elastic impact and perfectly inelastic impact respectively.

The movement of the needle may be investigated experimentally by measuring the time between the moment when the voltage is applied to the coil and the moment when the tip of the needle touches a metal plate which has been substituted for the paper. The measuring arrangement is illustrated in *fig. 8*: the voltage appearing at the terminals *c* and *d* of the resistance R_1 when the current passes through the coil is applied to one amplifier channel of a double-beam oscilloscope, while the voltage between points *a* and *b*, which becomes zero when the tip of the needle touches the metal plate, is applied to the input of the second channel.

Fig. 9 shows two oscillograms corresponding to needle displacements of 0.1 and 0.3 mm. For displacements between 1 and 2 tenths of a millimetre, the speed of the needle at the moment of impact on the inking ribbon is virtually constant, of the order of 0.5 m/s. The energy available for transferring the ink on to the paper is thus:

$$\tfrac{1}{2}\, mv^2 = 0.25 \text{ millijoule}.$$

This energy may be compared with that delivered by the impact of a hammer in printing an embossed character [2]: depending on the number of simultaneous copies, this energy should be between 3.1 and 7.2 mJ. In the case of characters formed by a dot mosaic of 7 rows and 5 columns the average number of dots defining a character is 20, and therefore the total energy delivered by the needles for printing the complete character is $20 \times 0.25 = 5$ mJ, which is sufficient for simultaneous printing of four copies. This energy may be increased or reduced to meet the required conditions

[2] B. J. Greenblott, A development study of the print mechanism on the IBM 1403 chain printer, Trans. AIEE **81**, Part I, 500-509, 1962.

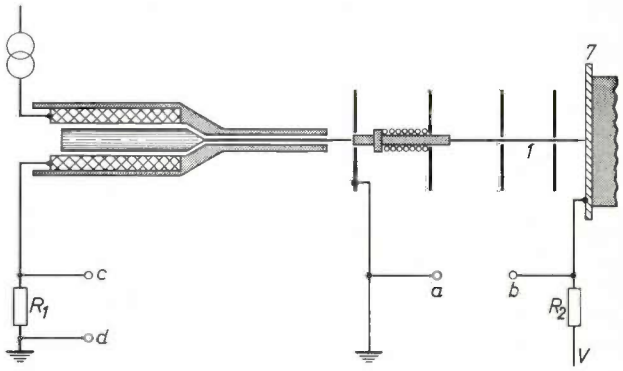


Fig. 8. Diagram of the arrangement used for investigating the movement of the needle in the mechanical printing element. The paper and the inking ribbon are replaced by a metal plate *7*. After travelling the required distance, the needle touches the plate, so that the voltage between points *a* and *b* falls to zero. This voltage and the voltage between terminals *c* and *d* of resistance R_1, which carries the electromagnet current, are applied to a double-beam oscilloscope.

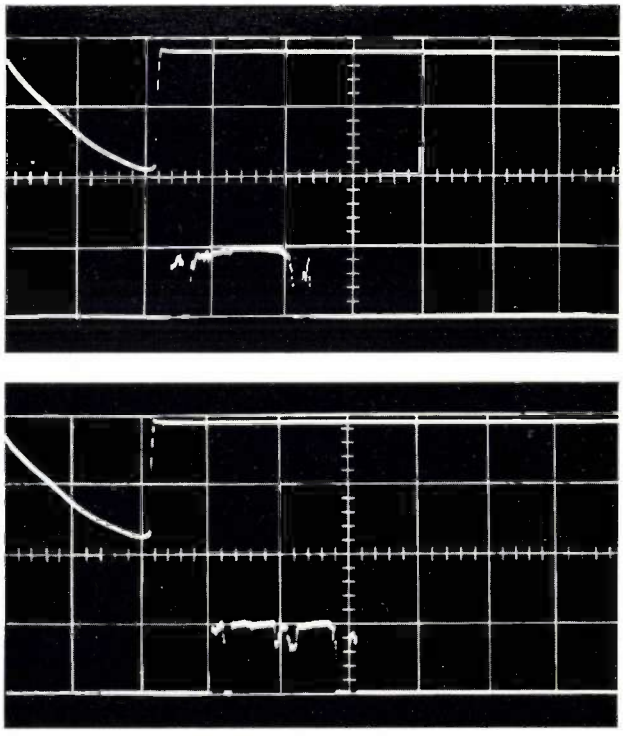


Fig. 9. Checking the time required for the displacement of a needle, using the arrangement of fig. 8 (horizontal scale: one square for 0.5 ms).
a) Distance covered 0.1 mm.
b) Distance covered 0.3 mm.

by varying the power supplied to the electromagnet. In the models described below the energy supplied to the electromagnet for printing a dot is 8 mJ, and the maximum operating rate is 500 dots a second.

Experimental models using the mechanical process

Serial numerical printer

In our first model the printing unit contains 7 needles with their points in a vertical line, and is moved hori-

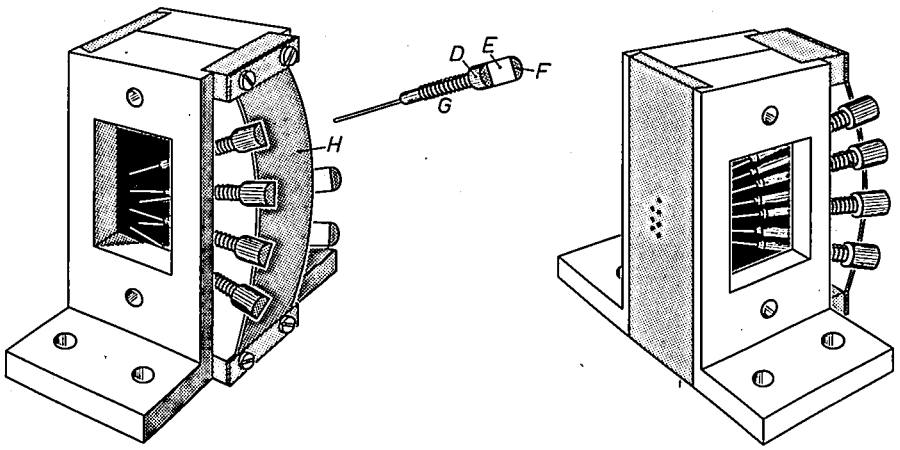


Fig. 10. Printing unit for the serial printer, seen from two angles. The unit contains the 7 needles for a column of the mosaic, arranged in two fan-like arrays on each side of the retaining plate *H*. The needles pass through the 7 holes at the front of the unit and converge to form a straight line at the level of the paper, which passes in front of the face at a distance of 0.2 mm. A needle has been removed from the unit on the left and illustrated separately. Note the head *D* with end-face *F*, the retaining surface *E* preventing the rotation of the needle, and the return spring *G*.

zontally at uniform speed. The printing of each character is thus completed before the printing of the next one starts (serial printer).

Fig. 10 shows the arrangement of the printing unit. The 7 needles are arranged in two fan-shaped arrays on each side of a vertical plane, and their tips converge to form a straight line at the level of the paper at a small distance (about 0.2 mm) in front of the face of the unit, as can be seen in the drawing on the right of the figure. At the other end of each needle there is a specially shaped head, which in the rest position is pressed against a curved retaining plate by a small spring.

When the printing unit moves from left to right, the selective actuation of the needles by means of their electromagnets, which are electronically controlled, results in the formation of the desired characters; see *fig. 11*. The 7 electromagnets are contained in unit *B* and they are directly excited by read amplifiers *AL* (blocking oscillators) associated with a magnetic core matrix *MC* whose rows and columns correspond to those of the mosaic in which the characters are formed. In this matrix each character is represented by a single wire threaded through a number of the cores at the particular junctions which together form the pattern of the desired character, as in the example shown dotted in the figure (the number "2").

In the application of the process described here, the characters to be printed are the ten numerals 0 to 9, which can be represented in any convenient code. When a character is to be printed, a current generator controlled by a decoder *CD* [3] sends a current pulse through the wire corresponding to that character, and the cores crossed by this wire then acquire a state of magnetization called state "1". The same current generates across the resistance *r* a voltage pulse which controls the sampling of the successive columns of the matrix *MC* by a current from a scanning circuit consisting of a magnetic core shift register *CS* [4]. The current from this register first traverses the

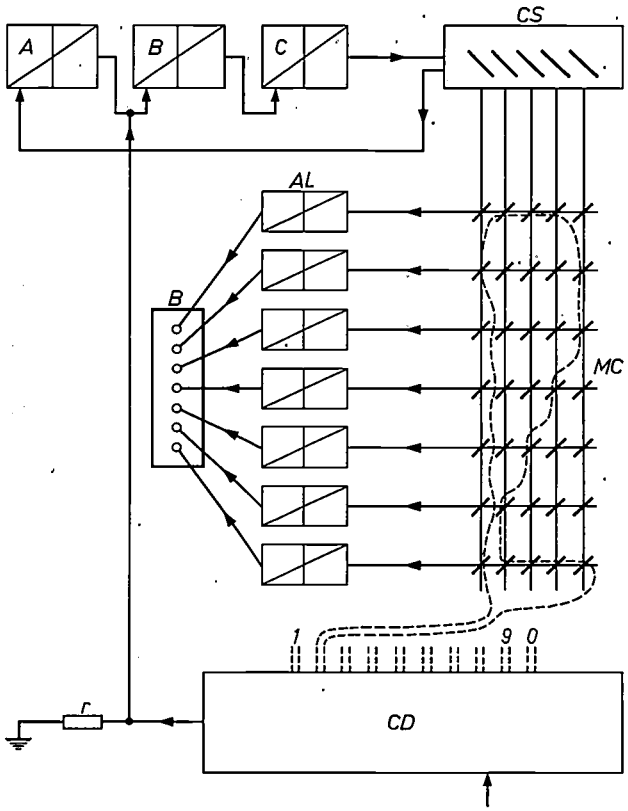


Fig. 11. Diagram of electronic circuits for the serial printer. The wire indicated by the dotted line forms the figure "2" in the mosaic matrix *MC*; there is a corresponding wire for each figure. *B* printing block for a vertical column of 7 dots (illustrated in fig. 10). *AL* read amplifiers, which control the electromagnets of the 7 needles located in *B*. *CD* decoder. *CS* scanning circuit controlled by a series of monostable circuits *A*, *B*, *C*, which give a suitable delay.

first column on the left of the matrix and sets the cores in this column to the state of magnetization "0". For every core that was previously set to state "1", the switch-over induces a voltage pulse in the horizontal lines passing through it, and the associated read amplifiers energize the electromagnets which print the dots appearing in the first column of the mosaic.

The scanning circuit *CS*, which is driven by the group of control circuits *A*, *B*, *C*, then delivers a second current

pulse which traverses the second column of the matrix *MC*, but the circuits *A*, *B*, *C* give a sufficient delay to permit the printing unit to be displaced towards the right. The switch-over of the cores previously set to state "1" in the second column of *MC* again induces a voltage pulse in the horizontal lines associated with the read amplifiers, and the dots which appear in the second column of the mosaic are printed.

This process is repeated until the dots in the last column have been printed and then, all the cores having been reset to state "0", the matrix is available for forming the next character and printing it.

With the operating characteristics of the printing elements described in the previous section, and a mosaic of 5 columns, the time taken to print a character is 10 ms. This means that — when spaces have to be included — 80 characters can be printed per second.

The photograph of *fig. 12* shows a rear view of a printer used with a small desk calculator. The carriage bearing the printing unit of needles and electromagnets is moved along two cylindrical rails by the rotation of a two-start lead-screw machined from a graphitized "Rilsan"[*] cylinder, which engages with a pin on the carriage. The lead-screw is driven by a small synchronous motor which picks up speed quickly enough to ensure uniform displacement, once movement has started. The printing system is located behind the paper. The roll of paper and the inking-ribbon spools, which can be seen in the photograph, give some idea of the small dimensions of this printer.

Fig. 13 reproduces as an example a document printed by the model described above.

Alphanumeric line printer

The second model which we have built can be used to print a text at a speed of 40 lines per second. It is capable of printing 64 different characters, namely letters, numerals and several symbols (*alphanumeric* printer).

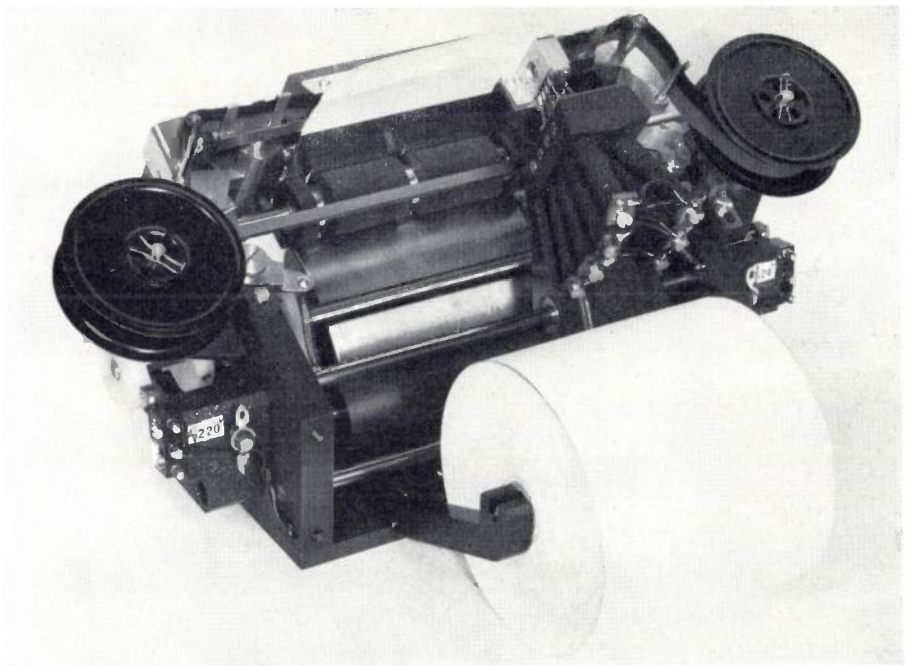


Fig. 12. Experimental model of the serial printer.

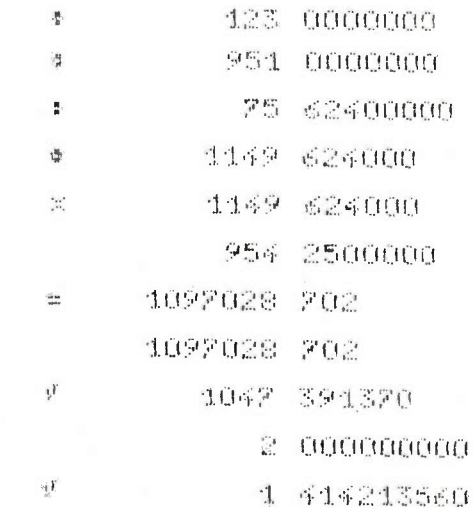


Fig. 13. Results of calculations made with a desk calculator and printed by the serial printer with a 4×7 dot mosaic. (The mosaic matrix in fig. 11 contained only 4×7 cores in this case.)

[3] In our model the numerals are represented in a biquinary code, but this detail is not essential.

[4] The operation of the magnetic core scanning circuit, well known in general applications, is described in more detail in an article by J. Borne, M. Audebert, M. Martin and R. Goudin, Une machine à calculer de bureau électronique, Acta electronica **8**, 203-239, 1964.

[*] Registered trade mark, Organico S.A.

Printing is done by means of a horizontal row of needles, arranged along the whole length of the line. The paper moves vertically at constant speed. The printing elements — needles and electromagnets — are grouped in the form of units placed side by side, each unit containing elements for 15 dots. The photograph in *fig. 14* shows the arrangement of the needles, which converge towards the point of the unit, flush with the outside face of a perforated guide-plate containing 15 holes aligned along an axis perpendicular to the general plane of the unit. Grouping 15 needles in this way over a width of three characters makes it possible to give each unit a thickness of 7.5 mm, equal to the diameter of the electromagnets. With this arrangement electromagnets of conventional construction may be used with no danger of overheating.

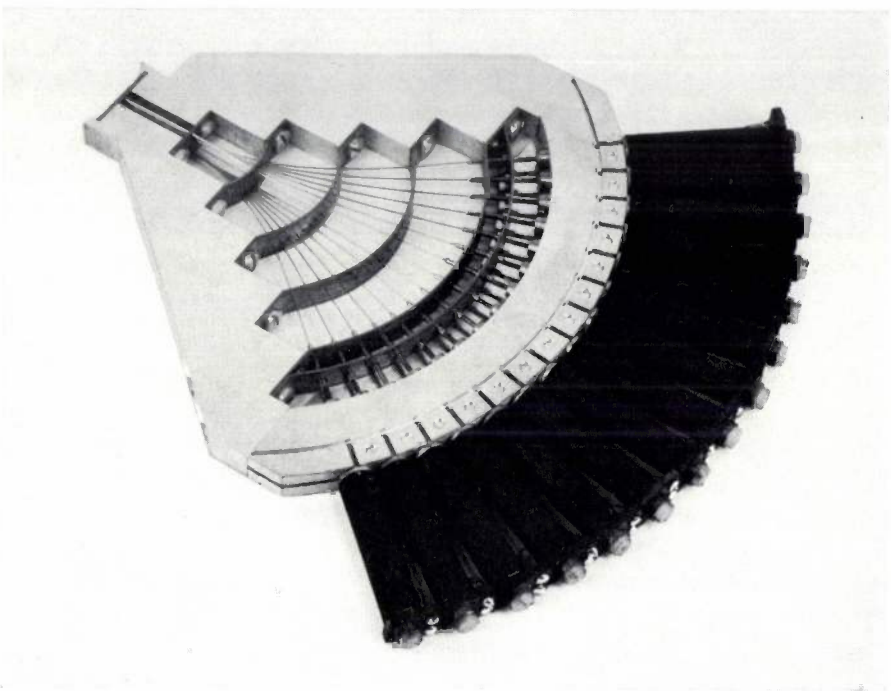


Fig. 14. Unit containing the printing elements for three neighbouring characters in the line printer (15 needles aligned along an axis perpendicular to the general plane of the unit).

Each needle is guided in holes cut into plates which are bent in such a way that the curvature of the needle is virtually the arc of a sine curve, as in natural bending. The small return springs around each needle are compressed between the last curved plate and the heads of the needles, thus keeping the needles in the rest position. The ends of the shafts of the cores of the electromagnets can also be seen, in contact with the needle-heads.

The operation of the printer will now be explained with reference to the diagram in *fig. 15*.

The first operation in printing a line of characters consists in writing the characters into a buffer store (*MT*). This is a magnetic-core matrix containing as many columns as there can be characters on a line of text (including the spaces), and as many horizontal rows as there are "bits" for each character in the code used in the data processing system. In our case this number is 6 bits, since the code is to have $64 = 2^6$ characters.

The cycle time of the buffer store used in the experimental model is 10 ms, and its access elements are entirely conventional:

a) a word register (*RM*) in which each bit may be written either by means of external information or by information derived from a horizontal row of the store;

b) a selection circuit (*X*) supplying currents for sampling the successive columns of the store, which serve both for reading and for writing or rewriting in conjunction with the inhibiting currents, which flow in the horizontal rows of the matrix and are controlled by the word register.

The line of characters is written into the buffer store, prior to printing, by simultaneously controlling, through the data processing system, the inputs of the word register and the selection circuit. The time taken to write in a line of characters in the buffer store can be considerably less than 1 ms.

The printing-control process is governed by the fact that printing line by line, with the paper continuously moving, actually has to be done "row by row", each line

— like the individual characters — being subdivided into 7 rows of dots. This process requires, apart from the buffer store and the word register, a decoder *D*, a mosaic matrix *MC*, and its scanning circuit *S*. The mosaic matrix, as in the first model described, is a magnetic-core matrix of 7 horizontal rows and 5 vertical columns corresponding to the definition of the printing mosaic. Each of the 64 characters is represented by a single wire which threads the cores which form the pattern of the outline of the character. (There are thus 64 wires, and in practice *four* magnetic cores have been used for each junction to reduce the number of wires passing through a core.) Each horizontal row of the matrix may be traversed by a write current or by an opposite read current, as selected by one of the 7 positions of the scanning circuit.

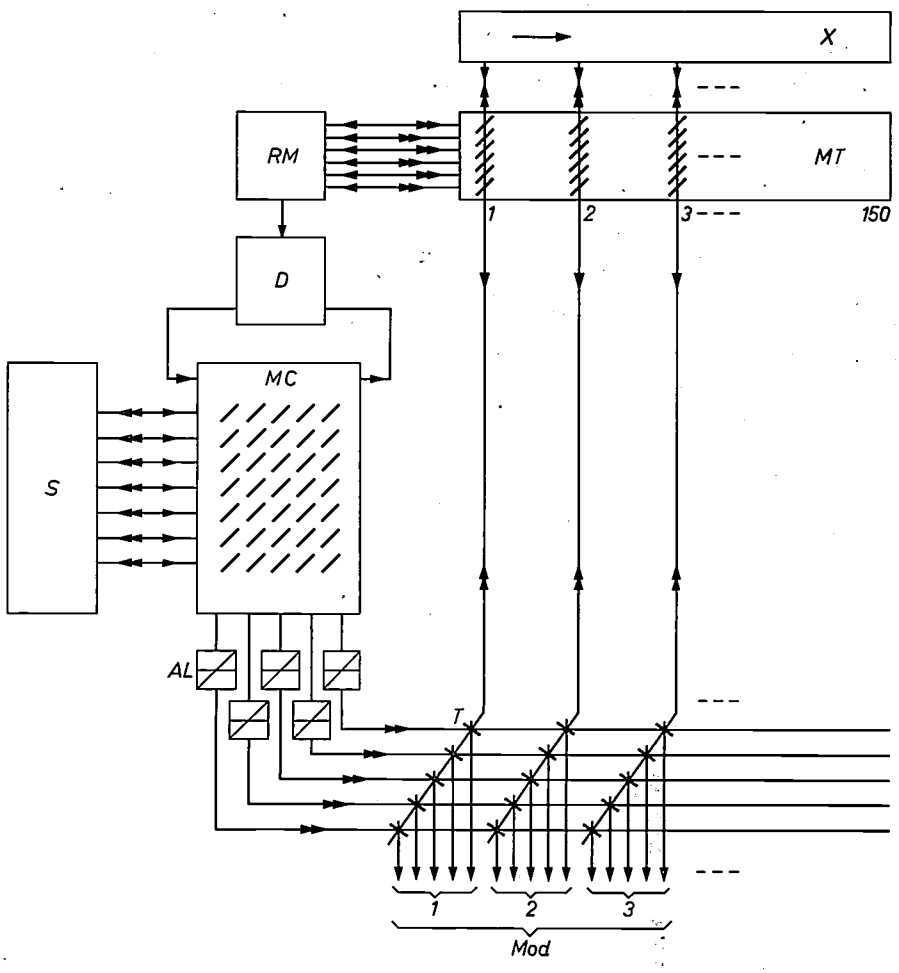


Fig. 15. Diagram of electronic circuits for the line printer. *MT* buffer store using magnetic cores, containing as many columns as there are characters in a line (e.g. 150). *MC* mosaic matrix, similar to the one shown in fig. 11, with scanning circuit *S*. To print a row of dots in a line each character of the line is successively transported to the matrix *MC* with the aid of the word register *RM* and the decoder *D*, and is rewritten in the buffer store *MT* after readout and printing. *X* selection circuit. *AL* read amplifiers. *T* magnetic cores whose change of state energizes the electromagnets of the printing elements. *Mod* unit comprising 15 needles (see fig. 14).

The single arrows in the diagram indicate the direction of the currents through the wires during the read cycle of the buffer store, and the double arrows indicate the direction of these currents during the rewrite cycle of the store.

Let us consider the first character of the line to be printed, written into the buffer store. For printing its first row of dots, the character is transferred to the word register *RM* in a time equal to one reading cycle of the buffer store, i.e. 5 μs. The decoder *D*, using the 6 bits of information identifying the character, selects the corresponding wire of the mosaic matrix *MC* and causes a current of magnitude $I_c/2$ to pass through it (I_c being the value of current which must flow in a single wire through the core to produce a field H_c slightly higher than the coercive field). This current, depending on its path through the junctions of the mosaic matrix, sets one or more of the five cores in the first horizontal row to state "1", with the assistance of another current $I_c/2$ which the scanning circuit *S* supplies during this time to this row only.

Thus, at the end of this first cycle, the binary information of the first character is erased from the buffer store and written into the word register, while the cores of the mosaic matrix which correspond to the dots to be printed for the first horizontal row of this character are in state "1".

In a second cycle, again lasting 5 μs, the needles corresponding to the first character of the line are actuated, and the character is rewritten into the buffer store again — this is obviously necessary for the later printing of the second row of dots, etc., after the first row of dots for all the characters of the line has been printed. For actuating the needles the vertical columns of the mosaic matrix *MC* are connected to 5 read amplifiers *AL*, each consisting of a blocking oscillator. Each oscillator can be triggered by the resetting to state "0" of one of the 7 cores in the column by a read current of magnitude $I_c/2$, in the opposite direction to the previous one and crossing each core in *two* wires; in this second

cycle this current is supplied at the appropriate moment by the scanning circuit to the first horizontal row of the mosaic matrix. Only the cores previously set to state "1" change state and trigger the associated read amplifiers. The separate control circuit of each electromagnet for actuating its needle (located in the unit *Mod*) again

It has thus taken 10 μs to trigger the printing of the dots for the first horizontal row of the first character, and the same operations will now be repeated for printing the dots of the first horizontal row of successive characters. For the second character the selection circuit *X* delivers to the second column of the buffer store

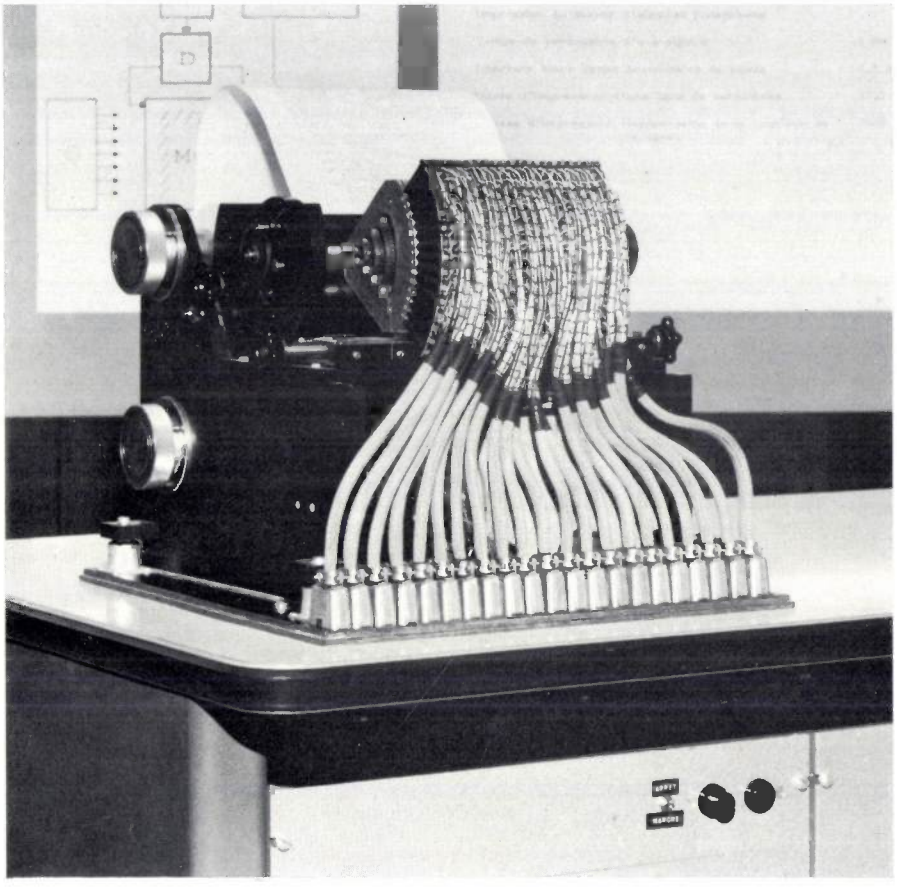


Fig. 16. Experimental model of the line printer, for printing lines of 63 characters at a speed of 2400 lines a minute.

consists of a blocking oscillator, triggered by the change of state of one of the magnetic cores *T*. This change of state is caused by the coincidence of a current $I_c/2$ from the appropriate read amplifier of the mosaic matrix and a current $I_c/2$ which originates at the same moment from a column of the selection circuit *X*. The latter current is also used for rewriting the binary information of the first character into the buffer store by means of the inhibiting currents controlled by the word register.

Thus, at the end of this second cycle, the binary information of the first character has been rewritten into the buffer store, the mosaic matrix has been erased, and the energizing of the electromagnets has been triggered and will be maintained by the blocking oscillators during the time necessary for giving the needles the required momentum, i.e. 1 ms.

and to the second group of 5 cores *T* the currents necessary for transferring the character to the word register *RM* (first cycle) and for triggering the blocking oscillators controlling the appropriate electromagnets and for rewriting the character in the buffer store (second cycle). (Incidentally, the read current delivered by the selection circuit *X* in the first cycle is also used for erasing from the connected group of cores *T* the information they have retained after printing the preceding row of dots.) The capacity of the buffer store, i.e. the length of the line, will typically be 100, 150 or 200 characters. The corresponding time required for triggering the printing of dots on a whole horizontal array will be 1, 1.5 or 2 ms.

Although experience has shown that it is possible to achieve a printing speed of 500 dots a second, with these printing elements, we have limited the speed of this line

printer to 400 dots per second. Under these conditions, the electromagnets, after controlling the printing of the dots for the first horizontal row of the first character of the line, may be re-energized after a time of 2.5 milliseconds for printing the dots of the second horizontal row. The process of printing a row of dots then begins again, controlled this time by the cores situated at the junctions of the second row of the mosaic matrix. If this matrix has 7 rows, the time needed for printing a complete line of characters is 17.5 ms.

Since the paper moves continuously in the vertical direction, the printed lines are tilted with respect to the horizontal line on which the needles lie, but the slope is only 6/10 000, and is therefore imperceptible. The displacement between two horizontal rows of dots is 0.42 mm, equal to the centre-to-centre distance between two neighbouring needles; the paper speed is 16.8 cm per second and the height of a character 2.94 mm.

In continuous printing the space between two successive lines is practically equal to half the height of a character, leaving a time of 7.5 ms between the printing of the last row of dots of a line and the first row of the next line. This time is amply sufficient for writing the characters of a new line into the buffer store.

Thus, the total time taken for printing a line and displacing the paper is 25 ms, giving a printing speed of 40 lines per second, or 2400 lines per minute. This is practically independent of the length of the line, up to 250 characters per line, or even more if the cycle time of the buffer store is less than 10 μs.

Fig. 16 shows a photograph of the experimental model. The line length was limited to 63 characters, and 21 units are therefore mounted vertically side by side so that the tips of the needles are arranged in a horizontal line parallel to the axis of a metal cylinder which drives the paper. The inking ribbon, situated between the guide-plate at the tips of the needles and the paper, is driven by the same movement.

The complete mechanism, comprising the printing elements, the cylinder and its driving motor, and the supplies of white paper and of inking ribbon in rolls of 100 metres, occupies about 40×40×40 cm.

For experimental purposes the buffer store was replaced by a store of larger capacity, containing 4096 words of 7 bits. With this store 64 lines of text could be printed, with 63 characters per line. This store may be supplied with information from a teleprinter, and the logic circuits therefore contain, in addition to the selection circuit *X* (which would also be used in conjunction with the buffer store), a selection circuit *Y* and synchronizing circuits for continuously printing out the contents of the store.

All the electronic circuits are contained in three racks whose total volume corresponds roughly to a cube of 45 cm side. The average power consumption of the electromagnet control circuits is about 300 W.

Fig. 17 is a reproduction of a piece of text printed with this experimental model.

IMPRIMANTE RAPIDE 2400 LIGNES / MINUTE

PROCEDE D IMPRESSION

LA MACHINE UTILISE UNE METHODE D ECRITURE DES SIGNES ET CA - RACTERES ALPHANUMERIQUES PAR POINTS DISPOSES SUIVANT UNE MOSA- IQUE COMPORTANT ICI 5 COLONNES VERTICALES ET 7 RANGEES HORIZON- TALES . LES POINTS SONT IMPRIMES EN TRANSFERANT SUR LE PAPIER L ENCRE CONTENUE DANS UN PAPIER ENCREUR PAR PERCUSSION A L AIDE D AIGUILLES FINES ACTIONNEES PAR DES ELECTROAIMANTS

UN GROUPE DE 5 AIGUILLES PEUT FRAPPER 5POINTS SUR UNE LIGNE HORIZONTALE ET IMPRIMER UN CARACTERE EN 7 ETAPES SUCCESSIVES LORSQUE LE PAPIER DEFILE DANS LE SENS VERTICAL AVEC UNE VITESSE UNIFORME

Fig. 17. Some text printed in a 5×7 dot mosaic at a speed of 2400 lines a minute.

Curve plotter in Cartesian co-ordinates

The process of forming an outline from adjacent dots, used here for forming characters by means of a horizontal row of needles, can also be employed for plotting curves. The only difference is that the gaps between the groups of needles, which were required for spacing the characters, must now be omitted. A curve plotter of this type has been built in our laboratories. The units, which, in the printer described above, contained 15 needles in three groups for printing three characters, have been slightly modified to contain 18 needles, whose regularly spaced tips are guided by

a single plate perforated with holes spaced at 0.42 mm intervals. A photograph of this curve plotter can be seen in *fig. 18*, and its operation will be explained with reference to *fig. 19*.

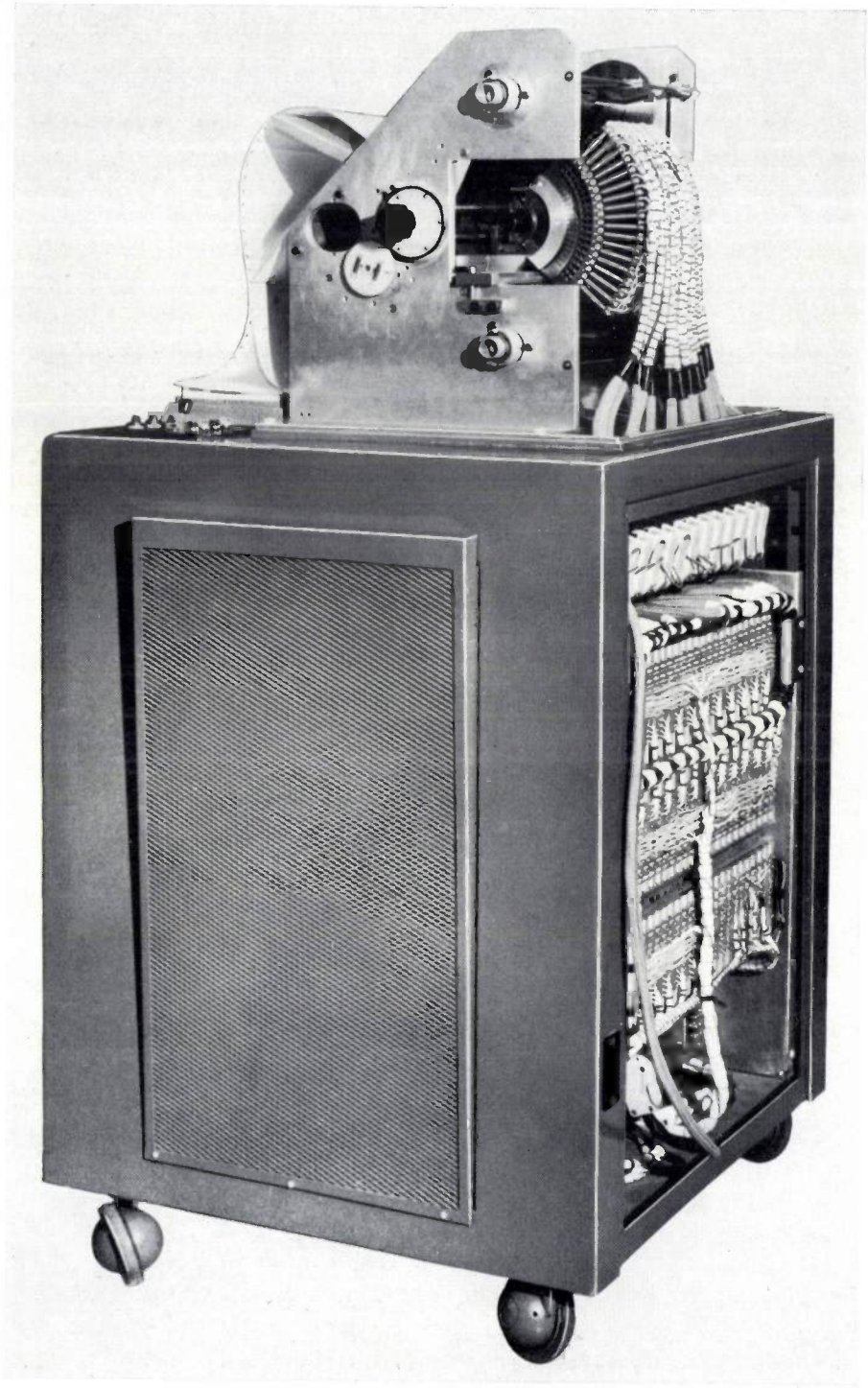


Fig. 18. Experimental model of the curve plotter.

In the experimental model described here, the co-ordinates x and y of all the points of the curve or of the family of curves to be plotted are computed and classified in order of ascending abscissa values in a memory M. This could if necessary be a computer store, and the classification could be performed by a programme either before or during the printing. One word of the store contains 18 bits and represents the co-ordinates of one point, that is to say 9 bits ($2^9 = 512$ steps) for the abscissa and 9 bits for the ordinate. The direction x is the direction in which the paper moves. Each needle with its electromagnet corresponds to a particular value

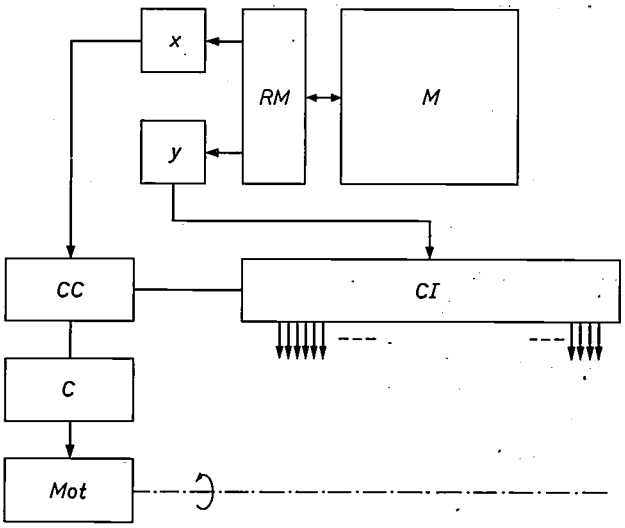


Fig. 19. Block diagram of the electronic circuits of the curve plotter. *M* store in which the co-ordinates of all the points of the family of curves to be plotted are entered. *RM* word register. *CI* printing control circuits. *CC* comparator. *C* counter. *Mot* paper-drive motor.

point every 10 μs, determined by the reading and writing cycle time of the store, for as long as the co-ordinates analysed in the successive addresses all contain the abscissa value $x = 0$. If the abscissa value contained in an address differs from zero, and the counter is still at zero — this could obviously be the case at the beginning of the analysis of the store —, the comparator prevents the printing of the corresponding point and its co-ordinates x and y remain stored in the word register.

After a time of 2.5 ms, the paper is advanced one step and the counter advances one position. If the abscissa value stored in the word register is $x = 1$, the corresponding point can be printed and the analysis of the store continued. If the value x is different from 1, the co-ordinates remain stored in the word register until the position of the counter reaches this value, i.e. until the paper, advancing step by step at the rate of 2.5 ms, has traveled the proper distance; the comparator then permits the printing of the corresponding point.

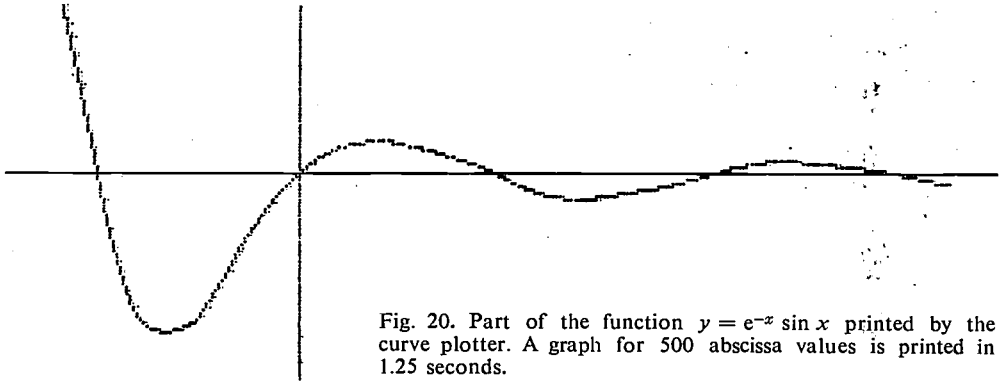


Fig. 20. Part of the function $y = e^{-x} \sin x$ printed by the curve plotter. A graph for 500 abscissa values is printed in 1.25 seconds.

of the ordinate y. The electromagnets are energized when the comparator circuit *CC* establishes that the abscissa value of a co-ordinate pair entered in the word register is equal to the state of a counter *C* which receives the successive driving pulses from the motor *Mot* for advancing the paper.

At the beginning the counter is set to zero. The co-ordinates contained in the first address of the store are transferred to the word register *RM*, and the abscissa value x is compared with the state of the counter. If $x = 0$, the value of y is decoded in the electromagnet control circuits *CI* and the point whose co-ordinates were stored in this address is immediately printed. These co-ordinates are rewritten into the store (for later use if required) and the contents of the word register are erased. The next address in the store is then transferred to the word register, and the value of the abscissa is again compared with the state of the counter, which has remained at zero. If the comparison again shows $x = 0$, the value of y is decoded and the corresponding point printed. This process may be repeated at the rate of one

Thus, for every step of the paper corresponding to a state of the counter, all the points of the same abscissa are printed virtually simultaneously (at the rate at which the store is analysed). The maximum number of these points is given by the relation:

$$N_{\max} = \frac{\text{period of driving pulses}}{\text{store cycle time}}.$$

In the experimental model this is:

$$N_{\max} = \frac{2.5 \text{ ms}}{10 \text{ μs}} = 250.$$

In practice the time necessary for plotting a curve will depend only on the speed of the motor used for driving the paper and on the maximum value of the abscissa. In the laboratory model this time is 1.25 seconds for a curve or a family of curves plotted in a rectangle limited to 500 abscissa values and 250 ordinate values. Part of a single curve plotted in this way is shown in *fig. 20*.

Conclusion

The results obtained in the experiments carried out in our laboratories confirm the great versatility of the process of forming characters by a dot-mosaic pattern and its adaptability to the different requirements of automatic data processing. Examples of documents printed with the models described here and reproduced in figs. 13, 17 and 20 demonstrate the printing quality and the definition obtained with what was still in effect an experimental version.

While retaining the important features of printing by mechanical impact — the use of a standard paper and the facility of making simultaneous copies — a complete range of printers can be built for data-handling equipment ranging from the simplest to the most elaborate.

The quality of the printing is amply sufficient for transmitting the information with great reliability; a further substantial improvement can be obtained by using a closer horizontal spacing of the needles, or by increasing the definition as illustrated by the document shown in *fig. 21*. This was printed at the Industrial Centre of S.A. Philips at Bobigny with a dot-mosaic pattern of 7 rows and 7 columns.

CE TEXTE A ETE IMPRIME SUR NOTRE MODELE EXPERIMENTAL

60 CARACTERES PAR SECONDE, CARACTERE PAR CARACTERE.

Fig. 21. Document printed with a 7×7 dot mosaic, by a line printer of improved definition, built by S.A. Philips at Bobigny.

The author wishes to thank the Délégation Générale à la Recherche Scientifique et Technique (France) who supported the study of the curve plotter.

Summary. The essential characteristics of printers for numerical data processing equipment are speed, reliability and adaptability. Methods have been devised to form characters by means of a dot mosaic, which makes it possible to combine these characteristics. Two such dot-printing processes have been studied, one using rapid heating and the other mechanical impact of sliding needles. Both processes have been tried out in practice. The second process has been studied in greater detail and several experimental models have been produced: a serial numerical printer capable of a speed of 80 numerals per second; an alphanumeric line printer operating at 2400 lines per minute; and a curve plotter which can be used to plot a family of curves, defined by the Cartesian co-ordinates of all their points, in a time of 1.25 seconds.

Electron-microscope investigation of the domain structure of magnetic particles

In an electron microscope with magnetic lenses the electrons in the beam are deflected by the Lorentz force acting upon them in the circular-symmetric field of a lens. The electrons pass through a very thin specimen in which the scattering varies from point to point, and conditions are arranged so that they are focused upon the screen to give an image of the specimen. The desired differences in photographic density are mainly obtained through the interception of the most strongly scattered electrons by a diaphragm, and also from the differences in the absorption of the electrons in the specimen, and from interference between direct and scattered electron rays in the image plane (phase contrast).

A departure from this simple situation is found when specimens of magnetic material are under investigation. In this case the field in or around the specimen will also exert Lorentz forces on the electrons and thus cause additional deflections of some parts of the beam. This effect, which shows its presence when the image is defocused, can be used for investigating certain magnetic properties of the specimen, such as the domain structure of thin films of ferromagnetic material [1] or the expulsion of the magnetic flux from a superconducting specimen [2].

In the investigation described here the specimens were not thin films but particles of magnetic material opaque to electrons. An image of the edge of the particle was obtained by means of an electron beam which grazes the edge of the particle. The magnetization of the particle causes a distortion of this image, and this distortion can give information about the magnetization. *Fig. 1* illustrates the principle of the method,

Conclusion

The results obtained in the experiments carried out in our laboratories confirm the great versatility of the process of forming characters by a dot-mosaic pattern and its adaptability to the different requirements of automatic data processing. Examples of documents printed with the models described here and reproduced in figs. 13, 17 and 20 demonstrate the printing quality and the definition obtained with what was still in effect an experimental version.

While retaining the important features of printing by mechanical impact — the use of a standard paper and the facility of making simultaneous copies — a complete range of printers can be built for data-handling equipment ranging from the simplest to the most elaborate.

The quality of the printing is amply sufficient for transmitting the information with great reliability; a further substantial improvement can be obtained by using a closer horizontal spacing of the needles, or by increasing the definition as illustrated by the document shown in *fig. 21*. This was printed at the Industrial Centre of S.A. Philips at Bobigny with a dot-mosaic pattern of 7 rows and 7 columns.

CE TEXTE A ETE IMPRIME SUR NOTRE MODELE EXPERIMENTAL

60 CARACTERES PAR SECONDE, CARACTERE PAR CARACTERE.

Fig. 21. Document printed with a 7×7 dot mosaic, by a line printer of improved definition, built by S.A. Philips at Bobigny.

The author wishes to thank the Délégation Générale à la Recherche Scientifique et Technique (France) who supported the study of the curve plotter.

Summary. The essential characteristics of printers for numerical data processing equipment are speed, reliability and adaptability. Methods have been devised to form characters by means of a dot mosaic, which makes it possible to combine these characteristics. Two such dot-printing processes have been studied, one using rapid heating and the other mechanical impact of sliding needles. Both processes have been tried out in practice. The second process has been studied in greater detail and several experimental models have been produced: a serial numerical printer capable of a speed of 80 numerals per second; an alphanumeric line printer operating at 2400 lines per minute; and a curve plotter which can be used to plot a family of curves, defined by the Cartesian co-ordinates of all their points, in a time of 1.25 seconds.

Electron-microscope investigation of the domain structure of magnetic particles

In an electron microscope with magnetic lenses the electrons in the beam are deflected by the Lorentz force acting upon them in the circular-symmetric field of a lens. The electrons pass through a very thin specimen in which the scattering varies from point to point, and conditions are arranged so that they are focused upon the screen to give an image of the specimen. The desired differences in photographic density are mainly obtained through the interception of the most strongly scattered electrons by a diaphragm, and also from the differences in the absorption of the electrons in the specimen, and from interference between direct and scattered electron rays in the image plane (phase contrast).

A departure from this simple situation is found when specimens of magnetic material are under investigation. In this case the field in or around the specimen will also exert Lorentz forces on the electrons and thus cause additional deflections of some parts of the beam. This effect, which shows its presence when the image is defocused, can be used for investigating certain magnetic properties of the specimen, such as the domain structure of thin films of ferromagnetic material [1] or the expulsion of the magnetic flux from a superconducting specimen [2].

In the investigation described here the specimens were not thin films but particles of magnetic material opaque to electrons. An image of the edge of the particle was obtained by means of an electron beam which grazes the edge of the particle. The magnetization of the particle causes a distortion of this image, and this distortion can give information about the magnetization. *Fig. 1* illustrates the principle of the method,

sometimes referred to as Lorentz microscopy. In fig. 1*a* the specimen *S* is situated in the object plane *O*. All the scattered electrons from point *A* are focused at *A'* on the image plane *F*, even if the direction of the electron beam is changed by a Lorentz force introduced by the magnetization of the specimen and acting at point *A* (fig. 1*b*). If, however, the lens is made weaker (fig. 1*c*), the specimen *S* is then no longer in the object plane, and the image of the edge (now blurred but still easily recognizable) is displaced under the action of the Lorentz force, say towards *B'*. The location of *B'* is related to the direction and magnitude of the Lorentz force. The same kind of thing happens when the lens is made stronger (fig. 1*d*). If the magnetization of the particle varies from point to point, then when the electron microscope is defocused the image of the edge of the particle shows a distortion that corresponds to the local magnetization [3].

The domains we referred to earlier are small areas in which the average magnetization has the same direction everywhere. In seeking the configurations which are energetically favourable, these domains arrange themselves into certain characteristic structures. Some electron-photomicrographs are shown here, obtained from a study of domain behaviour in particles of MnAl, a ferromagnetic material [4].

Fig. 2 shows pictures of the edge of such an MnAl particle. These pictures can be explained by reference to *fig. 3*, which illustrates schematically the effect of the Lorentz force (large arrows) exerted by the magnetic field of the domains (small arrows) on electrons grazing a specimen in the edge of which north and south poles alternate. The curved line shows the shape of the edge as it is seen by the defocused microscope. In fig. 2*b* and *c* the arrow indicates the position of the same irregularity in the specimen film. It can be seen that the distortions obtained on increasing and decreasing the focal length are displaced so that the peaks of the one correspond with the troughs of the other. This is what we should expect from fig. 1.

The pictures were all taken with a Philips EM 200 electron microscope. The image in this microscope is produced by means of four magnetic electron-lenses. The first of these, the objective lens, is an immersion lens, i.e. the specimen is located in the magnetic field of the lens. It is possible, however, to switch off the

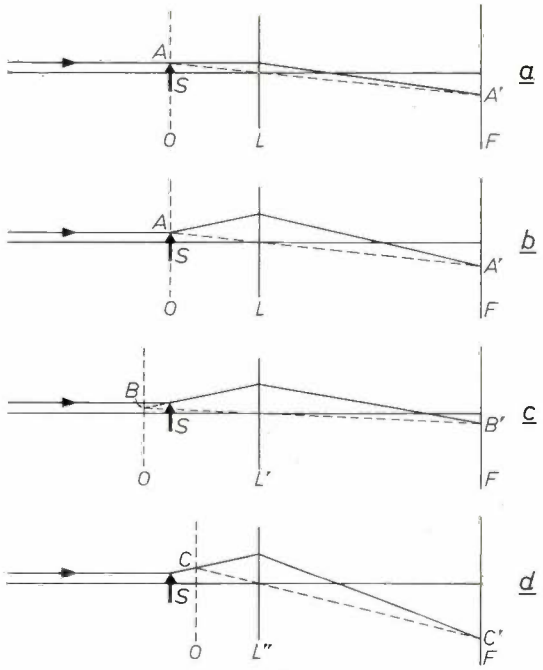


Fig. 1. Illustrating the formation by an electron beam of an image of the edge of a specimen *S* with magnetic properties, causing Lorentz forces to act on the electrons. *F* image plane. *O* object plane. *L* electron lens.

The figure does not take into account the image rotation about the optical axis which occurs with magnetic electron lenses.

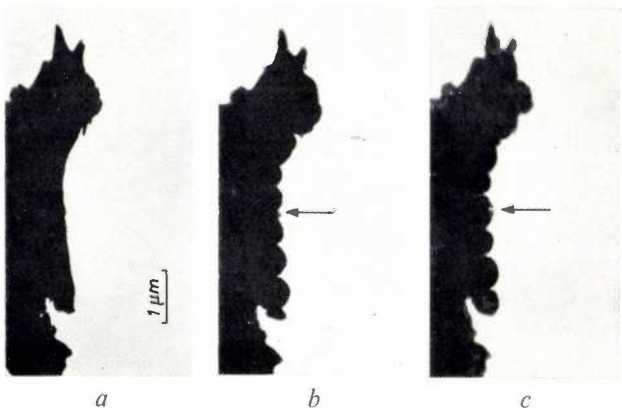


Fig. 2. Electron-photomicrographs of the edge of an MnAl particle (dark). Image *a*) with sharply focused microscope, *b*) with increased focal length, *c*) with reduced focal length.

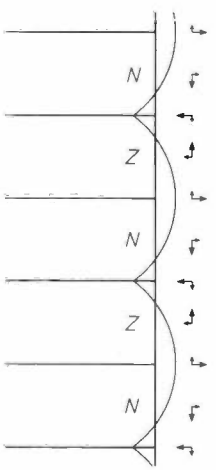


Fig. 3. Domain structure at the edge of a magnetic particle and the resultant image distortion. The electrons pass along the edge in paths perpendicular to the plane of the drawing.

[1] H. Boersch and H. Raith, Naturwiss. **46**, 574, 1959.
[2] M. J. Goringe and U. Valdrè, Phil. Mag. **8**, 1999, 1963.
[3] A comparable effect can be obtained with thin magnetized films which are transparent to the electron beam. The effect in this case can sometimes be explained more adequately with the aid of wave optics (see R. H. Wade, J. appl. Phys. **37**, 366, 1966).
[4] A. J. J. Koch, P. Hokkeling, M. G. van der Steeg and K. J. de Vos, New material for permanent magnets on a base of Mn and Al, J. appl. Phys. **31**, 75S-77S, 1960.

objective lens, and obtain an image of the specimen by means of the other three lenses. The specimen is not then in an external field. (In this application it does not usually matter that the maximum magnification of the instrument cannot be obtained.) Now, on the other hand, by increasing the current in the objective lens, it is possible to subject the specimen to an external magnetic field which can be varied from 0 to 8000 oersteds. The image distortion due to the magnetization of the specimen is not affected by this field variation — provided, that is, that the domain structure itself does not change. Although in this case the objective lens contributes again to the formation of the image, the magnification can be kept constant by appropriately adjusting the strength of the other lenses.

This experimental technique has proved very useful in our investigation of MnAl particles. One very interesting effect observed was a domain ordering that must be magnetostatically unstable (it had identical poles end-on, *fig. 4*). This arrangement can be explained by assuming that the domain walls, which in principle can be displaced under the influence of an external magnetic field, are pinned to certain defects in the crystal lattice [5]. This explanation was supported by observations on the MnAl needle shown in *fig. 5*. This needle contains a number of domains whose magnetic vector is perpendicular to the long axis of the needle (the same structure in fact as in fig. 3). When an external magnetic field is applied, this domain structure remains unchanged, irrespective of the orientation of the particle in the magnetic field, even when the external field is increased to 8000 oersteds. This implies that the material is magnetically very "hard", i.e. it has a high coercivity.

The examples described demonstrate that the electron microscope can be used for a variety of magnetic investigations without requiring special modifications.

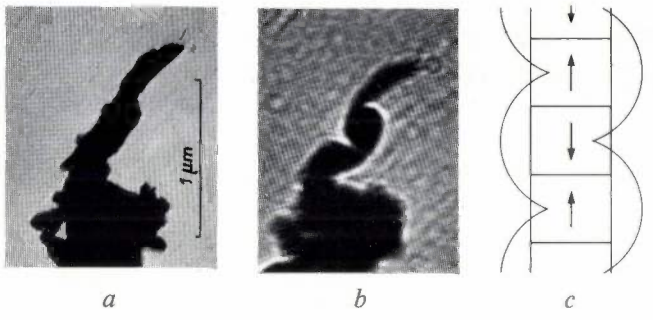


Fig. 4. Electron-photomicrographs of an MnAl particle (dark), *a*) with sharp focus, *b*) with increased focal length. *c*) The domain structure derived from the distortion.

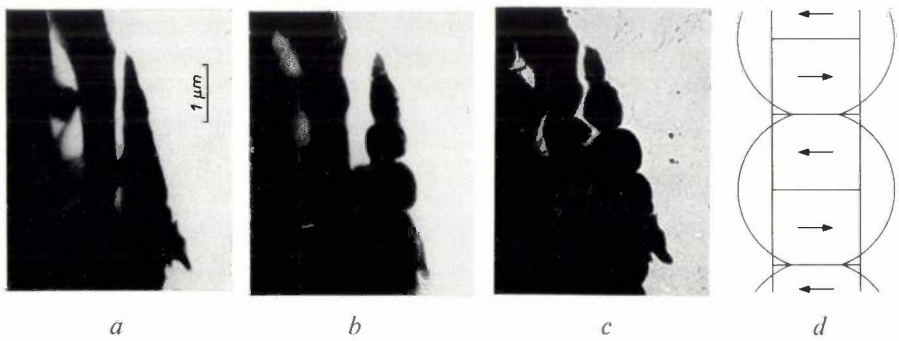


Fig. 5. Electron-photomicrographs of an MnAl particle (dark), *a*) with sharp focus, *b*) with increased focal length. *c*) As (*b*) but with a field of 8000 oersteds, perpendicular to the plane of the drawing. *d*) Domain structure corresponding to the distortion of the needle.

H. B. Haanstra
H. Zijlstra

[5] H. Zijlstra and H. B. Haanstra, Evidence by Lorentz microscopy for magnetically active stacking faults in MnAl alloy, J. appl. Phys. **37**, 2853-2856, 1966.

H. B. Haanstra and Dr. Ir. H. Zijlstra are with Philips Research Laboratories, Eindhoven.

Physical principles of photoconductivity

III. Inhomogeneity effects

L. Heijne

A number of important photo effects and associated phenomena in semiconductors are due to the presence of inhomogeneities of one sort or another, such as non-uniform illumination, inhomogeneity inside the material, the finite dimensions of a sample, etc. Some of these, like the normal photovoltaic effect, are of great practical significance, others are important in semiconductor research, and others again are of importance for a proper understanding of the processes that take place in transistors and diodes. These inhomogeneity effects are the subject of this third and last article on the physical principles of photoconductivity. Although inhomogeneities often give rise to complicated situations, it is shown that a comparatively simple treatment may be used in cases where the "ambipolar" approximation is applicable.

In the articles I and II of this series [1] it was consistently assumed that the photoconducting material and the excitation density were both homogeneous. The only inhomogeneity we took into account was the junction with the contact metal. Now in most photoconducting circuit devices the material is in fact homogeneous. However, there are cases where inhomogeneities are deliberately introduced to give certain desirable special effects, such as the photovoltaic effect.

Homogeneous *excitation* will seldom be found in practice: the light that penetrates into a substance is more or less strongly absorbed and its intensity therefore gradually decreases. The excitation density is approximately homogeneous only when the absorption is relatively weak.

Another unavoidable inhomogeneity is the surface of the material. If one of the processes determining the dynamic equilibrium in a photoconductor, for example recombination, occurs much more strongly at the surface than elsewhere, this may have a significant effect on the characteristics of a device made from that material.

Although in practice two or more inhomogeneity effects usually occur at the same time, we shall treat them separately in this article. For simplicity, we shall generally confine ourselves to special cases, as in the previous articles.

In the following we shall first discuss the photovoltaic effect. We shall then deal with some effects connected with the non-uniform excitation density that results when a photoconducting substance is illuminated with light that is strongly absorbed in it. In such a case the surface effects have a marked influence on the characteristics of a photoconducting device. These effects will be discussed at the end of the article, and it will be shown how the characteristics can be influenced through the surface by the nature of the surrounding gas.

The photovoltaic effect

In the photovoltaic effect illumination causes a potential difference between different parts of a semiconductor or of a semiconductor combined with a metal. Photocells based on this effect are thus capable themselves of supplying energy, e.g. to a measuring instrument, so that no external voltage source is required. This practical advantage has led to the widespread use of such cells, for example in photographic exposure meters. Solar batteries are also built up from these cells.

There are various mechanisms that can give rise to a potential difference of this kind. In the principal one a barrier is present, as in the *P-N* junction or the junction between a metal and a semiconductor. In both cases there is a "built-in" electric field, which appears from the localized curvature of the energy bands (band bending). In what follows we shall be concerned only with *P-N* junctions. The energy band scheme is shown

Dr. L. Heijne is with Philips Research Laboratories, Eindhoven.

[1] L. Heijne, Physical principles of photoconductivity, I. Basic concepts; contacts on semiconductors, Philips tech. Rev. **25**, 120-131, 1963/64; II. Kinetics of the recombination process; sensitivity and speed of response, Philips tech. Rev. **27**, 47-61, 1966.

in *fig. 1a*. We assume throughout this section that the excitation is homogeneous.

Pairs of charge carriers freed in the region of the barrier are separated by the field. In a closed external circuit a current I_f will then flow which increases with the number of light quanta absorbed in the barrier region and hence with the illumination. It should be noted here that a contribution to this current also comes from those charge carriers that are formed outside the barrier but so close to it that they are able to reach it within their lifetime. In the *P* region these carriers are electrons, in the *N* region holes; i.e. in general these are the minority charge carriers (fig. 1*b*). The average width of the region where these charge carriers originate is the diffusion-recombination length L. Quite often L is much greater than the width of the *P-N* junction itself, and in the following we shall assume that this is true in all cases.

When the circuit is open, the displacement of the charge carriers gives rise to a potential difference U_f whose direction is such that the internal field is partly compensated (fig. 1*c*). This potential difference cannot be greater than that corresponding to the jump in E_c and E_v in the barrier region of an illuminated material. (In practice this maximum cannot usually be reached.) We shall now derive the relation between the value of the electromotive force U_f and the short-circuit current generated in a *P-N* junction at a particular level of illumination. We shall confine ourselves to a value of illumination low enough for the photo-e.m.f. to be considerably smaller than the limiting value. In this case there is always a distinct barrier.

The starting point for the calculation is as follows. When the photocurrent begins to flow, e.g. in the case of fig. 1*c*, E_c increases in the *N* region. As a result, a current begins to flow in the barrier in a direction opposite to that of the photocurrent, in exactly the same way as happens when an external voltage is applied in the forward direction. The photo-e.m.f. U_f reaches its final value when this opposing current reaches the same magnitude as the photocurrent. The relation between the short-circuit current and U_f can therefore be derived directly from the current-voltage characteristic of the non-illuminated device, from the "forward" part of the characteristic. What then takes place in the *P-N* junction is illustrated in *fig. 2*.

As shown in the first article (see equation I,11 [*] and I,14) this characteristic can be written in the form:

$$j = j_s \left(\exp \frac{eU}{kT} - 1 \right). \quad \ldots \quad (1)$$

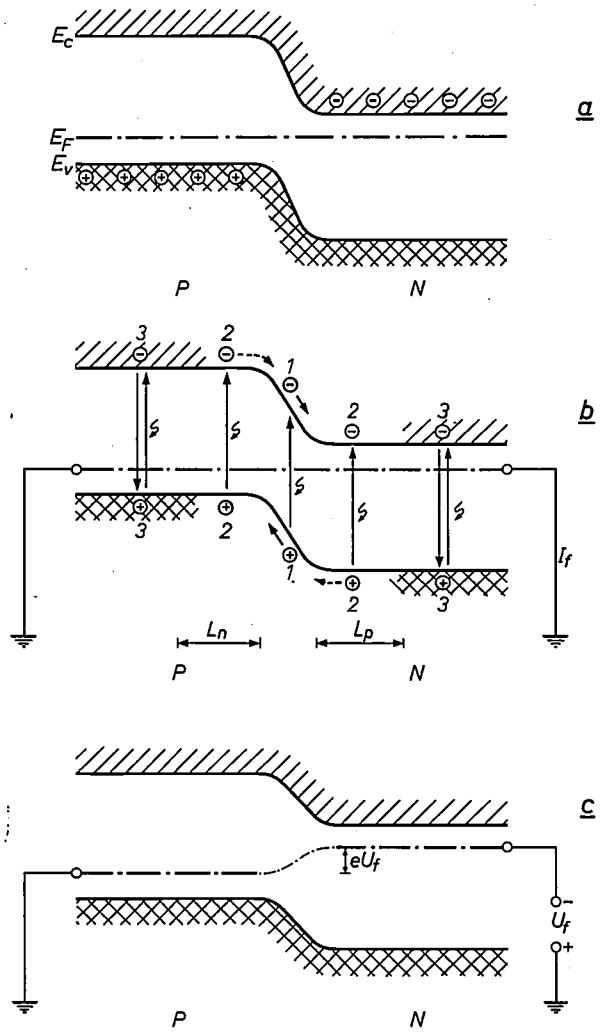


Fig. 1. Illustrating the photovoltaic effect. *a*) Edges of the conduction band (lower edge E_c) and the valence band (upper edge E_v) near a *P-N* junction. There is a jump in E_c and E_v at the location of the barrier. The chain-dotted line E_F represents the Fermi level. *b*) Charge carriers freed by light in the barrier zone (*1*) are separated by the "built-in" field, giving a photocurrent I_f. Charge carriers (*2*) freed at an average distance of less than one diffusion-recombination length (L_n or L_p) from the barrier diffuse towards it and also contribute towards this current. Charge carriers freed at a greater distance from the barrier (*3*) recombine without having made any contribution to the photocurrent I_f. Only the charge carriers freed by illumination are shown. *c*) When the *P* and *N* regions are not connected by an external circuit, a photo-e.m.f. U_f arises when the barrier is illuminated. The energy eU_f can be no greater than the jump made by E_c and E_v in the barrier zone in (*a*).

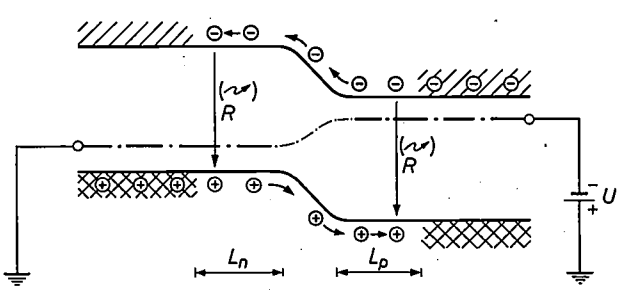


Fig. 2. The photo-e.m.f. U_f has the same magnitude as the external voltage U that must be applied in the forward direction to cause a current equal and opposite to the photocurrent to flow through the barrier. The arrows *R* represent recombinations, which may be accompanied by the emission of light.

[*] I or II in a reference to an equation shows that it is taken from one of the two earlier articles.

Here e is the electronic charge, U the applied voltage, k Boltzmann's constant and T the absolute temperature. The constant j_s is the absolute saturation value of the density of the reverse current at a high reverse voltage (U strongly negative). We thus find the open circuit photovoltage U_f, which corresponds to a short circuit current of density j_f, by substituting U_f for U in equation (1) and $-j_f$ for j, and then solving for U_f:

$$U_f = \frac{kT}{e} \ln\left(\frac{-j_f}{j_s} + 1\right). \quad \ldots\ldots \quad (2)$$

(We take the photocurrent as negative because it flows in the opposite direction to the forward current. The quantity $-j_f$ is therefore positive.) If required, j_f may also be expressed in terms of the excitation density G;

$$j_f = -eG(L_n + L_p). \quad \ldots\ldots \quad (3)$$

This equation follows immediately from the fact that the photocurrent, as explained above, is determined by the number of electrons freed by the light in regions of width L_n and L_p adjoining the barrier zone; the contribution from this zone itself has been neglected.

The quantity j_s may be derived by applying the same kind of reasoning as that used in determining the magnitude of the photocurrent under short-circuit conditions. This shows that j_s is determined by the thermal generation of charge carriers in the barrier zone and in regions of widths L_n and L_p on each side of the barrier (fig. 1*b*). If the width of the barrier is again negligible compared with $L_{n,p}$, then following this argument we can write:

$$j_s = e\left(\frac{L_n n_m}{\tau_n} + \frac{L_p p_m}{\tau_p}\right). \quad \ldots\ldots \quad (4)$$

In this expression τ_n and τ_p are the lifetimes of electrons and holes respectively, while n_m and p_m are their concentrations in the regions where they are minority charge carriers. Thus, in accordance with (I,5), the ratios n_m/τ_n and p_m/τ_p are the respective densities of the thermal generation.

We see that the diffusion-recombination lengths L_n and L_p are important material parameters both in j_f and j_s (and hence also in U_f). It can be shown (see below) that the diffusion lengths depend upon the lifetime τ and the diffusion constant D as follows:

$$L_n = \sqrt{D_n \tau_n}; \quad L_p = \sqrt{D_p \tau_p}. \quad \ldots \quad (5)$$

Since, according to Einstein's relation (I,8), the diffusion constant is proportional to the mobility μ, we see that, just as in ordinary photoconductivity, the magnitude of the effect is determined by the material constant $\mu\tau$ (cf. I,7).

The relation between the photocurrent I_f and the photo-voltage U_f of a photocell with an external energy-consuming load may be obtained by combining equations (1) and (3) [2]:

$$I_f = Sj_f = Sj_s\{\exp(eU_f/kT) - 1\} - eSG(L_n + L_p). \quad (6)$$

Here, S is the cross-sectional area of the photocell. Since U_f can be written as the product of $-I_f$ and the load resistance, the photocurrent in the load at a given excitation density G is entirely determined by (6), and the same therefore applies for the power delivered to the load.

Anomalous photovoltaic effect

In most cases the photo-e.m.f. generated across normal *P-N* junctions can be described very satisfactorily by equation (2). A few cases are known, however, where very much higher photo-e.m.f.'s are found. These may even be many times the energy gap, reaching perhaps 100 V. This anomalous photovoltaic effect, as it is called, is found for example in evaporated films of cadmium telluride [3] and in single crystals of zinc sulphide [4]. In all probability these high e.m.f.'s are due to some kind of series connection of a large number of barriers. This assumption is supported by the fact that with the evaporated CdTe films the strength of the effect was found to depend closely on the angle of deposition. In the ZnS crystals the barriers are presumably to be identified with the interfaces between regions with cubic structure and those with hexagonal structure, from which these crystals are built up.

Injection luminescence

To close this section we should point out that in barriers there is often a relation between photoelectric properties and luminescence. In principle, a photocell can also function as a light source when a voltage is applied in the forward direction. The current that then flows consists of majority charge carriers on one side of the barrier zone; these become minority carriers after crossing the barrier, and recombine. This recombination may be associated with the emission of light (fig. 2). *P-N* light sources of this kind have been produced, e.g. with gallium phosphide [5], and light produced in this way has in fact been used to stimulate laser action in gallium arsenide [6].

---

[2] J. Tauc, Photo and thermoelectric effects in semiconductors, Pergamon Press, Oxford 1962, section 3.6.
[3] L. Pensak, Phys. Rev. **109**, 601, 1958.
[4] S. G. Ellis, F. Herman, E. E. Loebner, W. J. Merz, C. W. Struck and J. G. White, Phys. Rev. **109**, 1860, 1958. A survey will be found in the book by Tauc [2].
[5] See W. Glässer, H. G. Grimmeiss and H. Scholz, Philips tech. Rev. **25**, 20, 1963/64.
[6] M. I. Nathan, W. P. Dumke, G. Burns, F. H. Dill, Jr., and G. Lasher, Appl. Phys. Letters **1**, 62, 1962. See also IBM J. Res. Devel. 7, 62, 63, 66, 70 and 72, 1963.

Effects of non-homogeneous excitation

We shall now deal with effects that occur when the excitation density in a homogeneous photoconductor is not the same throughout the material. In practice this is frequently the case. Often, for example, part of the light will be intercepted by a screen of some kind, so that a shadow is cast on the photoconductor. Another case of non-homogeneous excitation density noted earlier, occurs when the light incident on the substance is strongly absorbed in it and therefore decreases in intensity as it penetrates the substance.

The external effects that occur, or may occur, in such cases all arise because the density of the charge carriers is inhomogeneous when the excitation is inhomogeneous, and this inhomogeneity gives rise to diffusion. The nature of this diffusion is somewhat complicated because of the existence of *two* kinds of charge carrier, the electrons and the holes, whose charge is of opposite sign. They usually also have different mobilities. This diffusion, which is also of importance in various other sorts of semiconductor devices, will first be dealt with separately. We shall then discuss two of the external effects, the Dember effect and the photomagnetoelectric effect.

Ambipolar diffusion of charge carriers

We now consider the case where part of a photoconducting element is screened and the remainder of it is uniformly illuminated for a short time. There is thus a region in which the excitation density G is zero, which is sharply distinct from a region in which G is not equal to zero for a short time and has the same value everywhere. After the light pulse, the most mobile charge carriers initially penetrate into the screened region in larger numbers than the less mobile ones. Let us assume that the more mobile charge carriers are the electrons. In the screened region a negative space charge is then formed, and a positive space charge is formed in the unscreened region. This is accompanied by an electric field which decelerates the electrons — which have got ahead — and accelerates the holes. If the concentrations of the charge carriers are so high that, even where the difference in concentration is relatively small, there is already a high space charge present, and hence a strong field, this field will then tend to correct the movement of the charge carriers in such a way that both kinds diffuse into the screened region to virtually the same extent. This is the process known as ambipolar diffusion (*fig. 3*). In formal terms, ambipolar diffusion can be considered to a good approximation as an ordinary diffusion of neutral pairs of charge carriers, the pairs being assigned their own diffusion coefficient, called the coefficient of ambipolar diffusion D_a.

It can easily be shown (see below) that under normal

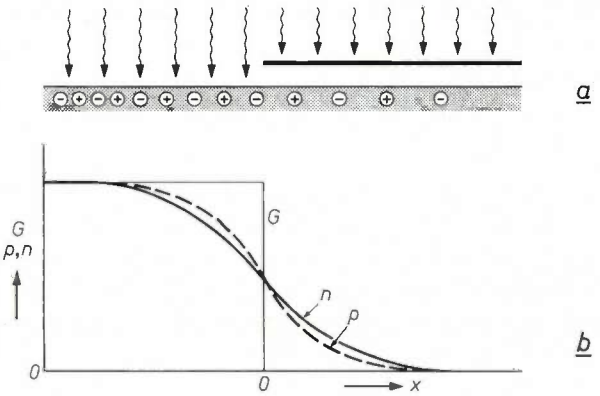


Fig. 3. Ambipolar diffusion. *a*) When part of an illuminated photoconductor, is in the shadow of a screen, charge carriers from the boundary region between light and dark diffuse into the unilluminated region. *b*) Plot of the excitation density G, electron concentration n and hole concentration p, present at a certain moment in the direction at right angles to the boundary (the x-co-ordinate). If the concentrations in the illuminated region are high enough, ambipolar diffusion takes place and the curves for n and p nearly coincide. The figure refers to the case where the charge carriers are freed by a very short pulse of light as well as to the case of continuous illumination.

conditions this coefficient is given by the general expression:

$$D_a = \frac{D_n D_p (n + p)}{n D_n + p D_p}. \qquad (7a)$$

In cases where $n = p$, as in intrinsic material or in a photoconductor which is strongly illuminated by light of quantum energy $E_c - E_v$, the expression for D_a becomes simply:

$$D_a = \frac{2 D_n D_p}{D_n + D_p}. \qquad (7b)$$

In this case the coefficient of ambipolar diffusion is therefore independent of the concentrations. If there is much difference between D_n and D_p, the *smaller* quantity is the significant one. For example, if $D_n \gg D_p$, we find:

$$D_a \approx 2\, D_n D_p / D_n = 2\, D_p. \qquad (7c)$$

The charge carriers then diffuse together with a velocity which is twice the velocity at which the slower holes would diffuse in a field-free environment.

If $n \neq p$ (i.e. the semiconductor is extrinsic), then D_a is equal to the diffusion velocity of the minority charge carriers, which is immediately evident when we consider the form assumed by (7a) in the case where $n \gg p$. We then find:

$$D_a \approx D_n D_p\, n / n D_n = D_p. \qquad (7d)$$

Here also D_a is independent of the concentrations.

Up to now we have tacitly assumed that no electric field was applied to the photoconducting material. Even

in the presence of a field, separation of the charges would give rise to an additional "internal" field which attempts to keep the charges of opposite sign together. This is illustrated in *fig. 4* for the hypothetical case of a suddenly created "cloud" of electrons and holes. Apart from the expansion of the region of increased concentration as a result of ambipolar diffusion, there is now a displacement of the whole cloud, which moves in the direction in which the faster charge carriers would move in the applied field. The movement of the cloud can again be described by taking into account

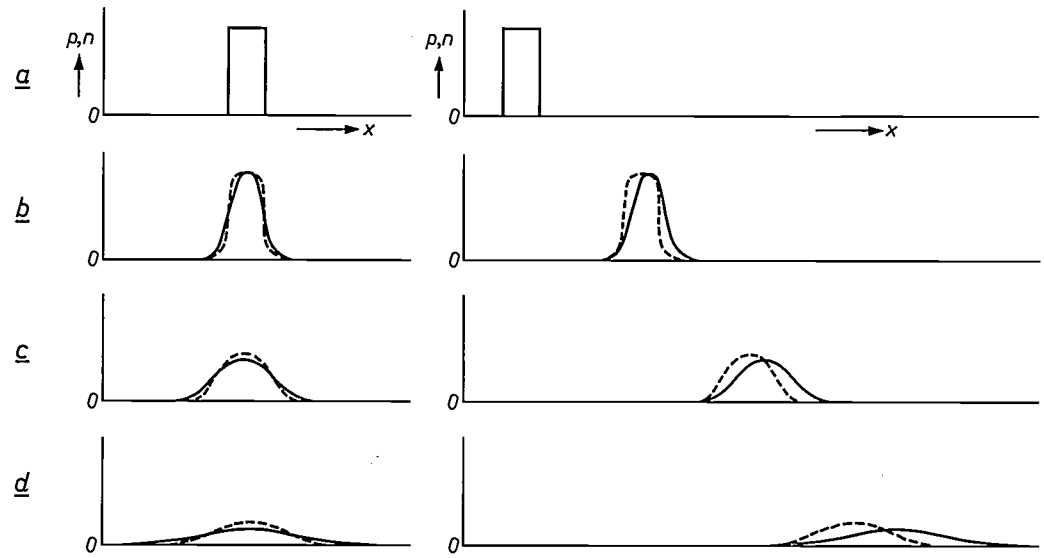


Fig. 4. *a*) Concentration of the electrons (n) and of the holes (p) after a pulsed local excitation. *b*), *c*) and *d*) The situation some time later: on the left, with an electric field; on the right, with no electric field. This field draws the charge carriers with the greater mobility towards the right. The solid lines apply to charge carriers with the greater mobility, the broken lines to those of the other kind. The area under both curves is always the same, but gradually becomes smaller as a result of recombination.

only the externally applied field, allowing for the additional "internal" field produced by the space charges by assigning an ambipolar mobility to the cloud.

In deriving the expression (7a) for the coefficient of ambipolar diffusion it is convenient to start with the expressions for the density j_n of the electron current and the density j_p of the hole current. In a unidimensional case these are:

$$j_n(x) = e\,n(x)\mu_n F(x) + eD_n\,\mathrm{d}n(x)/\mathrm{d}x, \quad . . \quad (8a)$$

$$j_p(x) = e\,p(x)\mu_p F(x) - eD_p\,\mathrm{d}p(x)/\mathrm{d}x. \quad . . \quad (8b)$$

Here F is the electric field strength, n the electron concentration, p the hole concentration and μ the mobility. Now let us consider the case where the electrical circuit which includes the semiconductor is open. Since no net current can flow anywhere when the circuit is open, the electron and hole currents must exactly compensate each other everywhere in the circuit. Expressed as an equation:

$$j_n = -j_p. \quad . . . . . \quad (9)$$

In ambipolar diffusion the concentrations of the additional charge carriers are everywhere virtually identical (cf. fig. 3), so that, neglecting statistical fluctuations:

$$\mathrm{d}n/\mathrm{d}x = \mathrm{d}p/\mathrm{d}x. \quad . . . . \quad (10)$$

Using equations (9) and (10) and Einstein's relation (I,8), and eliminating F from (8a) and (8b) (here F is the strength of the internal field tending to keep electrons and holes together), we find:

$$j_n = \frac{eD_nD_p(n+p)}{nD_n + pD_p}\,\frac{\mathrm{d}n}{\mathrm{d}x}. \quad . . . \quad (11)$$

If this process is to be described as one of pure diffusion, with the ambipolar diffusion coefficient D_a, we must write:

$$j_n = eD_a\,\mathrm{d}n/\mathrm{d}x. \quad . . . . . \quad (12)$$

Equation (7a) then follows directly from (11) and (12).

Of course, (7a) only applies when, in thermal equilibrium, the concentrations n and p are not dependent upon x, and when there are no space charges. The presence of contacts and surfaces has thus been tacitly

excluded. In these regions there is often no electroneutrality, and electric fields and concentration gradients exist here even in thermal equilibrium.

The diffusion-recombination length

Let us now consider the distance over which charge carriers penetrate by ambipolar diffusion into a non-illuminated region of a photoconductor. There are limits to this distance since the charge carriers do not have an unlimited lifetime because of recombination. We assume now that the illuminated region is *continuously* illuminated and that the remaining conditions are such that the diffusion can be described by a coefficient of ambipolar diffusion (see above).

According to equation (I,5) the recombination density R of the electrons is given by:

$$R = n(x)/\tau, \qquad (13)$$

where τ is the lifetime of the charge carriers and $n(x)$ is their concentration at x (fig. 3*b*). The density of the (electron) diffusion current is:

$$j_n = -j_p = eD_a \, dn/dx. \qquad (14)$$

In view of the continuity of the current its divergence (i.e. in our unidimensional case the derivative with respect to x) must be equal to the recombination current eR, since no charge carriers are liberated in the non-illuminated region. This gives:

$$dj_n/dx = eR, \qquad (15)$$

or, with (13) and (14):

$$d^2n/dx^2 = n/D_a\tau, \qquad (16)$$

whose solution is:

$$n(x) = n(0) \exp(-x/L_a), \qquad (17)$$

where $L_a = \sqrt{D_a\tau}$ and $n(0)$ is the value of n at the boundary between the illuminated and non-illuminated regions. The quantity L_a, which has the dimension of a length, is a measure of the distance over which the charge carriers penetrate into the non-illuminated region. We should note here that the quantities L_n and L_p in eq. (5) in the first section are identical with the values of L_a in the *P*- and *N*-type regions. The quantities D_n and D_p in (5) correspond in fact to D_a, since the relationship (7d) applies to the case there discussed ($n \neq p$).

It will be evident that what has been stated above does not only apply to the local liberation of charge carriers by *light*. It can equally apply for example to the diffusion of electrons towards the *P* region and of holes towards the *N* region which occurs when a forward voltage is applied to a *P-N* junction (fig. 2), as in a diode or in the base of a junction transistor [7].

In the theory given so far it has been tacitly assumed that all the charge carriers are completely free, in other words that no traps are present. If a substantial proportion of electrons or holes are held in traps, an effective diffusion-recombination length must be used. This is found by multiplying the normal diffusion length by a factor equal to the ratio of the concentrations of the free and the trapped electrons. A situation of this kind is found in CdS, where the holes are very quickly captured by traps, so that when the surface is excited the charge carriers can only penetrate slightly into the crystal.

The Dember effect

If the two kinds of charge carrier have different mobilities, the diffusion can give rise to a type of photovoltaic effect, called the Dember effect after its discoverer [8]. *Fig. 5* illustrates schematically the origin of this photo-e.m.f. A piece of photoconducting material

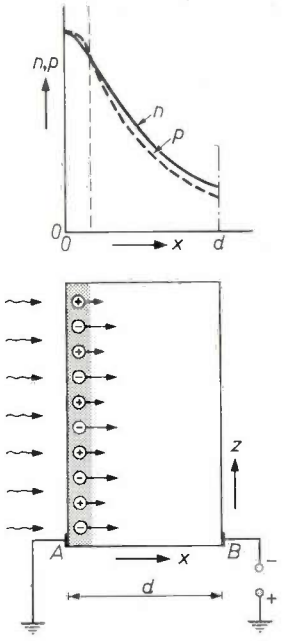


Fig. 5. The Dember effect. Upon excitation in a surface layer (shaded) the difference in the diffusion rate of electrons and holes gives a potential gradient in the direction perpendicular to the surface, and hence to a potential difference between the electrodes *A* and *B* on the front and back of a crystal plate illuminated from one side only. The figure relates to the case where the electrons are the more mobile charge carriers. Above: illustrating the concentration distribution over the thickness of such a crystal plate; the solid curve relates to the electrons.

can be seen which is provided on both sides with an electrode; the left-hand electrode transmits light, so that the light incident form the left can penetrate into the photoconductor. The depth of penetration of the light used is very small, so that excitation only occurs in the shaded area. Assuming again that the electrons diffuse faster than the holes, the right-hand electrode will be negatively charged and a potential difference will appear between *A* and *B*.

This potential difference is called the Dember voltage, and its value can be found directly from the equations given. We again assume that the concentrations of the charge carriers freed by the light are sufficiently high throughout in the sample to ensure an "ambipolar" flow. We assume further that the substance is an insulator in the dark, so that the charge carrier concentrations present in thermal equilibrium are entirely negligible compared with those found on exposure to the light. In that case, at every point x we can take the concentrations n and p as being identical, and also their derivatives dn/dx and dp/dx. The intensity F_D of the field which keeps the charge carriers together is found by solving (8a) and (8b) for F for the case $j_n = -j_p$ (open circuit). We then find:

$$F_D = -\frac{kT}{e}\,\frac{D_n - D_p}{n(D_n + D_p)}\,\frac{dn}{dx}. \quad . \quad . \quad (18)$$

Substituting in this expression

$$dn/dx = -n/L_a, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (19)$$

obtained by differentiation of (17), we can write the Dember field strength F_D as:

$$F_D = \frac{kT}{e}\,\frac{D_n - D_p}{D_n + D_p}\,\frac{1}{L_a}. \quad . \quad . \quad . \quad . \quad (20)$$

The Dember voltage U_D is found from this by integrating over the thickness d of the sample:

$$U_D = -\int_0^d F_D\,dx = \frac{kT}{e}\,\frac{D_n - D_p}{D_n + D_p}\,\frac{d}{L_a}. \quad (21)$$

We see that in the case which we are considering here of a strongly illuminated insulator, the Dember voltage is independent of the charge-carrier concentration and therefore independent of the illumination intensity. The sign of this photo-e.m.f. depends on the difference between the diffusion coefficients and thus indicates directly which of the two kinds of charge carrier is the faster.

If the values of D_n and D_p are known from other measurements, the ambipolar diffusion length can be found from the saturation value of the Dember voltage under strong illumination, given by (21). If D_n and D_p are unknown, but it is known that one diffusion constant is large compared with the other, then $(D_n - D_p)/(D_n + D_p) \approx 1$ (or -1) and the measurement yields directly the value of the diffusion length of the slower charge carriers (see equation 7c).

Measurements of the Dember voltage have up to now been carried out almost exclusively for *qualitative* investigations. Since the effect is located in the layer immediately below the surface of the photoconductor, there are usually fairly serious disturbances due to the space charge present there. As already noted, the associated marginal fields, which are already present when there is no illumination, also give rise to photo-e.m.f.'s and it is difficult to correct for these.

The photomagnetoelectric effect

The Dember effect just discussed is the occurrence of a photo-e.m.f. in the direction of diffusion of the charge carriers. For reasons of symmetry it is clear that *perpendicular* to this direction, e.g. in the z-direction in fig. 5, no potential difference is to be expected, at least if the illumination in this direction is uniform.

The situation changes, however, if a magnetic field is applied, for instance, in the y-direction (*fig. 6*). The

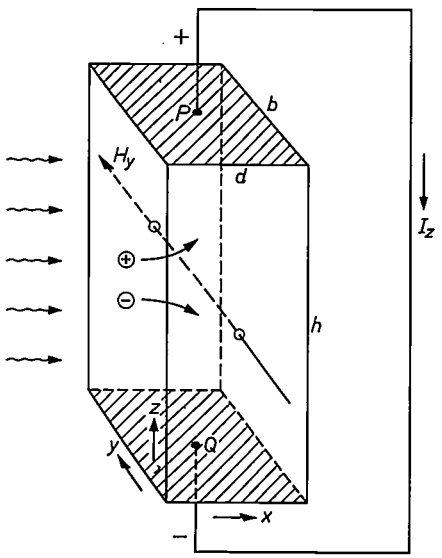


Fig. 6. The photomagnetoelectric effect (PME effect) occurs under the same conditions as give the Dember effect (fig. 5) if there is also a magnetic field parallel to the illuminated surface. The PME effect may be regarded as the Hall effect of the two diffusion currents (of holes or electrons respectively). A potential difference appears between the electrodes P and Q or, if these are connected, a current I_z flows in the external circuit.

electrons and holes flowing in the x-direction are then subject to a Lorentz force which, in our example, deflects the holes upwards and the electrons downwards. The contact P thus acquires a positive potential with respect to the contact Q. If there is an external connection between P and Q, a photocurrent will therefore flow in it [9]. This effect is known as the photomagnetoelectric effect (PME effect for short) and may be regarded as a Hall effect of the two diffusion currents. It will always appear when there are two diffusion currents and a magnetic field has been applied, and will

[7] See the treatment of this subject in W. Shockley, Electrons and holes in semiconductors, Van Nostrand, New York 1950, and in E. Spenke, Elektronische Halbleiter, Springer, Berlin 1965 (2nd edn.).

[8] H. Dember, Phys. Z. **32**, 554, 1931 and **33**, 207, 1932.

[9] T. S. Moss, L. Pincherle and A. M. Woodward, Proc. Phys. Soc. **B 66**, 743, 1953; P. Aigrain and H. Bulliard, C.R. Acad. Sci. Paris **236**, 595, 1953.

therefore also appear in the special case where μ_n and μ_p are equal, i.e. the case where the Dember voltage is zero.

The magnitude of the effect can easily be calculated. As for the Dember effect, we shall consider the case of an insulator which is strongly excited only at the surface. The z-component of the Lorentz force acting on the moving particles is given by:

$$K_z = \pm eH_y v_x, \quad \ldots\ldots \quad (22)$$

where v_x is the velocity of the particles in the x-direction. The plus sign relates to holes, the minus sign to electrons. If we describe the Lorentz force by a field strength F_z, the plus sign suffices and we can write:

$$F_z = H_y v_x. \quad \ldots\ldots \quad (23)$$

The velocity v_x is found by dividing the density of the particle current by the concentration of the particles. The current density follows immediately from (14):

$$j_p/e = -D_a \, dn/dx. \quad \ldots\ldots \quad (24)$$

If in this expression we substitute $-n/L_a$ for dn/dx (see eq. 19), we find:

$$v_x = j_p/ne = D_a/L_a. \quad \ldots\ldots \quad (25)$$

The Lorentz field strength (23) is then:

$$F_z = H_y \, D_a/L_a. \quad \ldots\ldots \quad (26)$$

If the magnetic field is homogeneous, F_z is independent of the co-ordinates. The PME voltage U_{PME} is then simply the product of this field strength and the height h of the crystal (see fig. 6):

$$U_{PME} = hH_y \, D_a/L_a. \quad \ldots\ldots \quad (27)$$

We see that under the conditions considered, the magnitude of the PME effect is no longer dependent upon the excitation but only on the "ambipolar" quantities D_a and L_a. As already discussed, these are often largely determined by the properties of the minority charge carriers or of the less mobile charge carriers (cf. 7c and d). We have also seen that $L_a = \sqrt{D_a \tau}$. If the lifetime τ is known from other measurements, the PME effect therefore makes it possible to determine the mobility of the minority charge carriers. This is very welcome, because in conventional measurements of conductivity and Hall effect the properties which are most in evidence are those of the charge carriers of the other type.

The PME effect is also very useful in the study of surface phenomena. If light is used which is very strongly absorbed, so that nearly all the electrons and holes are freed immediately under the surface, the influence of the surface is reflected in the magnitude of the PME short-circuit current.

Calculation of the PME short-circuit current

If, in the case shown in fig. 6, contacts P and Q are interconnected by a conductor of very low resistance, an electric current will then flow in the z-direction of the crystal. The density of the current is given by:

$$j_z = n(x)e(\mu_n + \mu_p)F_z = n(x)e(\mu_n + \mu_p)H_y D_a/L_a. \quad \ldots \quad (28)$$

Here n, the concentration of the charge carriers, is again a function of x and the same therefore applies for j_z. To find the total current I_z in the z-direction we must therefore multiply by the width b in the y-direction and integrate over the thickness d, after replacing $n(x)$ by the expression found in the right-hand side of equation (17), where $n(0)$ now represents the concentration at the surface:

$$I_z = b \int_0^d j_z \, dx = bn(0)e(\mu_n + \mu_p) \frac{H_y D_a}{L_a} \int_0^d \exp \frac{-x}{L_a} \, dx$$

$$= bn(0)e(\mu_n + \mu_p)H_y D_a \{1 - \exp(-d/L_a)\}. \quad \ldots \quad (29)$$

The diffusion coefficient D_a is given in this case by eq. (7a). We see that the PME short-circuit current, like the PME voltage, is dependent on the quantities D_a and L_a. Furthermore, the current is proportional to the value $n(0)$, the surface concentration of the charge carriers freed by the light.

Surface effects

The general considerations on the quantum states of electrons in a crystal, as discussed in the first article, apply only to a crystal with no boundary faces. The existence of a crystal surface which bounds the crystal does not fit into the picture of a strictly periodic lattice. Although, for practical purposes, the picture of the quantum states in the interior of the crystal does not require correction, it will be evident that electrons situated at the surface may occupy quantum states that do not correspond to those inside the crystal. These surface states, as they are called, differ therefore from the states in the conduction and valence bands as well as from the quantum states of electrons that are bound to localized impurity centres in the crystal. Surface levels also usually have a capture probability for electrons and holes which is different from that of the impurity centres.

Apart from the surface states which are a direct consequence of the finiteness of the dimensions of the crystal lattice, states also arise at the surface of a crystal which are connected with foreign molecules, atoms or ions present at the boundary face. These particles may originate from the surrounding gas, e.g. atmospheric air. The properties of the corresponding surface states, such as the energy and the capture probability for electrons or holes, are closely dependent on the nature of the adsorbed particle. The situation is even more complicated if the charge carriers present inside the crystal influence the state of the absorbed foreign particles (whether ionized or not) and their number per unit area. We may note here that the total number of surface states per unit area is usually relatively high, and

may even be of the same order of magnitude as the number of lattice atoms per unit area.

Surface states can profoundly affect the properties of a crystal as a photoconductor or semiconductor. In the first place, their presence makes it possible for a charge to be present at the surface. This surface charge affects the thermionic work function, discussed in the first article. Moreover, in the adjoining part of the crystal a counter-charge may occur in the form of a space charge. This causes, just as we have seen for the semiconductor-metal junction, curvature of the energy bands near the boundary face. A case of this kind is illustrated schematically in *fig. 7*. It relates to a *N*-type semiconductor with surface states which are half occupied by electrons. In accordance with the theory given in I, in thermal equilibrium the Fermi level E_F must then coincide at the surface with the energy level E_s of the surface states, which is tied to the lower limit E_c of the conduction band and the upper limit E_v of the valence band. Inside the crystal, on the other hand, the location of E_c and E_v relative to E_F is determined by the donor concentration; in general therefore this will differ from that at the surface. In the surface region there is a positive space charge in the form of ionized donors, which in a neutral crystal, has the same magnitude as the total negative charge on the surface.

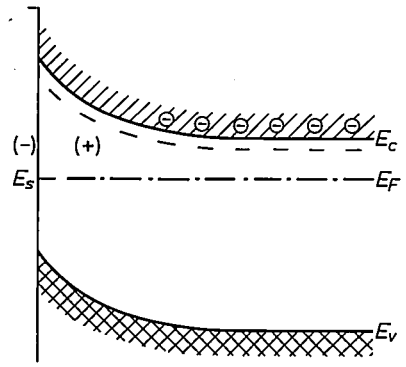


Fig. 7. Energy-band scheme of the edge region of an *N*-type semiconductor with a negative charge at the surface. This charge has driven the electrons away from the edge region, giving a positive space charge there, which is carried by ionized donors. In the case illustrated the energy of the surface states E_s is equal to the Fermi energy E_f; half the surface states are then occupied by electrons.

A second reason why surface states may have a considerable effect on the semiconducting and photoconducting properties of a crystal is that they are effective recombination centres. The preferential recombination at the surface is often particularly pronounced in very pure crystals, since these contain few internal recombination centres.

As a result of the relatively strong recombination at the surface, the sensitivity of a photoconducting crystal will be less for greater absorption of the light, i.e. the closer to the surface the light is absorbed. In general the absorption of light whose quantum energy is slightly greater than $E_c - E_v$, increases with decreasing wavelength, and the sensitivity will therefore decrease with decreasing wavelength. This explains why the sensitivity of a photoconducting crystal is often greatest for light whose wavelength corresponds approximately to the energy gap; see I, fig. 3.

Because of the existence of surface states which are connected with the adsorption of gas, it can be understood why the properties of some photoconducting crystals depend on the surrounding gaseous atmosphere.

We shall now deal in somewhat more detail with the properties of the surface states and their influence on the properties of photoconducting crystals.

Surface recombination velocity

In the theoretical treatment of the recombination process at the surface it is generally assumed that the process may be considered as a (pseudo-)monomolecular one (see II). The "recombination current density" R_s, i.e. the number of electrons or holes recombining in unit surface and in unit time, is taken to be proportional to $n(0)$, the concentration of the electrons or holes. Expressed as an equation:

$$R_s = s\,n(0). \quad \ldots\ldots \quad (30)$$

The proportionality factor s is the surface recombination velocity.

In I and II, where the recombination processes were always assumed to take place in the bulk of the crystal, we were able to calculate the charge-carrier concentrations generated by homogeneous excitation from the excitation density G and the lifetimes τ_n and τ_p (see equation I,5). If we take into account the recombination at the surface, the quantity s will also appear in the equations, and it will become more significant the closer to the surface the light is absorbed. In all cases the surface recombination causes an inhomogeneous charge-carrier distribution, since even with homogeneous excitation, the stronger recombination at the surface makes the concentration dependent on location.

We shall now calculate, for steady state conditions, the way in which $n(0)$, the concentration of the photoelectrons at the surface, depends on the excitation density, for the extreme case where the absorption of the light takes place entirely at the surface. This is the same case as we considered for calculating the Dember and PME effects. Here again we shall assume that the ambipolar approximation may be used.

Let us assume that G_s electrons are freed at the surface per second and per unit area, G_s holes thus being created at the same time. These charge carriers dis-

appear from the surface zone either by recombining at the surface (eq. 30) or by diffusing away from the surface. This diffusion is ambipolar and may thus be described by (24), in which, from (19), we may substitute $-n(0)/L_a$ for the gradient of n at the surface ($x = 0$):

$$j_p/e = -D_a \, dn/dx = n(0) \, D_a/L_a. \quad . \quad . \quad (31)$$

Taking both mechanisms into account we thus obtain for the steady state at the surface:

$$G_s = sn(0) + n(0) \, D_a/L_a, \quad . \quad . \quad . \quad . \quad (32)$$

from which the surface concentration is found to be:

$$n(0) = G_s/(s + D_a/L_a). \quad . \quad . \quad . \quad . \quad (33)$$

With the help of (17) we can also find from this the concentration of the photoelectrons (and hence the concentration of the holes, which is equal to it for ambipolar diffusion) at any location $x \neq 0$, as a function of the excitation density.

If we compare equation (33) for the photoelectron concentration at the surface with the equation $n = G\tau$ (I,5), which holds for the bulk of a crystal under homogeneous excitation, we see that instead of τ we have the expression $1/(s + D_a/L_a)$. In the denominator of this expression we can distinguish a term relating to recombination at the surface and another relating to diffusion away from the surface with subsequent recombination. It may be added that $1/(s + D_a/L_a)$ does not have the same dimension as τ, because G_s is an excitation density per unit area and G is an excitation density per unit volume.

With the exception of s we can determine experimentally all the quantities occurring in (33), so that s can be calculated. In a strongly absorbing substance the excitation density follows directly from the illumination intensity, D_a/L_a can be found by measurements of the PME effect (cf. 27) and $n(0)$ can be found from the surface conductivity, provided the mobilities are known.

We should point out that the recombination at the surface is by no means always monomolecular, as was assumed for eq. (30). This relation may still be applied if the recombination is not monomolecular, but s is then no longer a constant of the material but a quantity dependent on the illumination [10].

The adsorption of gas

Before starting our description of the influence of adsorbed gas on the properties of a semiconductor (or a photoconductor) and of the influence of illumination on the adsorption, we ought to say something about the adsorption itself, i.e. adsorption in the absence of light that can generate free charge carriers.

Gases can be either inert, electropositive, or electronegative. Given an opportunity, a molecule of an electropositive gas will give up an electron, whereas a molecule of an electronegative one will try to capture an electron. Oxygen is a familiar example of a strongly electronegative gas.

Atoms of an electronegative gas that are adsorbed at the surface of a semiconducting crystal can acquire electrons from the crystal, thus changing into negative ions. The extent to which this happens depends on the nature of the semiconductor and also on the extent to which the gas is electronegative.

An adsorbed neutral molecule or atom is bound to the crystal surface by London-Van der Waals forces. If the molecule has a dipole moment, this gives another attractive force, which is particularly strong if the crystal is also of a polar nature. For adsorbed *ions* the electrostatic Coulomb forces make an important contribution; the bond is then frequently even stronger. The first case is sometimes referred to as physical adsorption and the second as chemisorption; the distinction is fairly arbitrary, however, and there is no sharp dividing line.

The general condition to be fulfilled if electrons are to be transferred from the crystal to an adsorbed gas atom or molecule is that the resultant state should be a more stable one; in thermodynamic terms, a state of lower free energy. This will be the case when the energy level which the electron is going to occupy in the adsorbed atom is lower than the Fermi level of the semiconductor. In thermal equilibrium the state of charge of the adsorbed gas is therefore completely determined by the relative situation of both energy values.

The way in which the situation of the adsorption-desorption equilibrium is determined by the partial pressure P of the relevant gas and the electron concentration n in the semiconductor will be illustrated with the example of an electronegative, diatomic gas X_2 whose atoms are chemically monovalent. At the crystal surface the following equilibrium reactions may then take place:

$$X_{2g} \rightleftarrows X_{2a}; \quad [X_{2a}] \doteq k_1 P, \quad . \quad . \quad (34a)$$

$$X_{2a} \rightleftarrows 2X_a; \quad [X_a]^2 = k_2[X_{2a}], \quad . \quad (34b)$$

$$X_a + e^- \rightleftarrows X_a^-; \quad [X_a^-] = k_3 n[X_a]. \quad (34c)$$

(The subscript g means in the *gas*, the subscript a means *adsorbed*.) Each of these equilibrium reactions is characterized by a constant k_i, which brings the relevant concentrations into relation with one another (law of mass action, right-hand column); the constants k_i depend on the temperature. (The relation $[X_{2a}] = k_1 P$ holds only when the crystal surface is largely unoccupied.)

As can be seen from (34), k_1 and P determine the

concentration $[X_{2a}]$ of the adsorbed molecules. These quantities and k_2 determine the concentration of the adsorbed atoms (34b), which in its turn, together with the electron concentration n, determines the concentration of the adsorbed ions (34c). If n is large — N-type semiconductor, or E_F well above the middle of the forbidden zone — then the equilibrium (34c) lies well to the right. If n is small this is true when the gas is strongly electronegative, because k_3 is then large — the energy level E_a of the electron bound to an atom is well below E_c. In both cases the greater part of the adsorbed gas is present at the surface in the form of negative ions, and the electron concentration (and therefore the donor concentration or the location of the Fermi level) has a marked influence on the quantity of adsorbed gas.

Going back to the physical approach, we should like to make two further remarks. In the first place an increasing negative charge on the crystal surface causes an upward bending of the bands and hence an increase of E_a, which is linked to E_v and E_c (*fig. 8*). The energy E_a thus approaches closer to E_F, so that the occupation probability $f(E)$ given by the Fermi-Dirac equation (I,1) gradually diminishes. In the second place the electron concentration in the layer immediately below the crystal surface is strongly reduced by the increase of the negative space charge on the surface [11]. As a result the probability of an electron in the conduction band moving to the surface becomes steadily smaller. The establishment of an adsorption equilibrium is not infrequently extremely slow and can take many minutes.

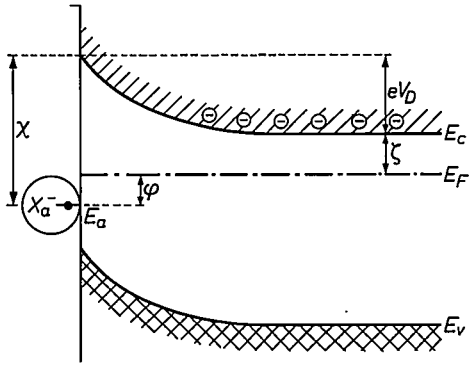


Fig. 8. As fig. 7, for the case where the negative surface charge is carried by adsorbed, strongly electronegative ions. The electrons then occupy a quantum state whose energy E_a is lower than E_f by an amount φ. The difference χ between $E_c(0)$, the value of E_c at the surface, and E_a is independent of E_c: if $E_c(0)$ rises as a result of an increase in the band curvature (eV_D greater) then E_a; also rises and φ becomes smaller.

The relation between the quantity of adsorbed gas and the pressure, called the adsorption isotherm, can be calculated as follows [12] for the case sketched above — electronegative gas which is adsorbed in the form of singly and negatively charged atoms to an N-type crystal, absence of illumination and relatively low surface occupation. For an energy level situated at a distance φ below the Fermi level (fig. 8) we find from the Fermi-Dirac distribution function (I,1) that the degree of occupation is $1/\{1 + \exp(-\varphi/kT)\}$. The ratio of $[X_a^-]$, the number of negative gas ions per unit area to $[X_a]$, the number of neutral gas atoms per unit area at the surface, given by (34c), now takes the form:

$$[X_a^-]/[X_a] = \exp(\varphi/kT). \quad \ldots\ldots\ldots\ldots \quad (35)$$

From (34a) and (34b) we have for low occupation:

$$[X_a] = (k_1 k_2 P)^{\ddagger}. \quad \ldots\ldots\ldots\ldots \quad (36)$$

Combining the last two equations we find:

$$[X_a^-] = (k_1 k_2 P)^{\ddagger} \exp(\varphi/kT). \quad \ldots\ldots \quad (37a)$$

When φ in this expression is replaced by $(\chi - \zeta - eV_D)$, see fig. 8, eq. (37a) becomes:

$$[X_a^-] = (k_1 k_2 P)^{\ddagger} \exp\{(\chi - \zeta - eV_D)/kT\}, \quad \ldots \quad (37b)$$

in which V_D is still unknown.

Now the negative charge of the gas ions on the surface is exactly compensated by the positive charge of the ionized donors in the space-charge region. This charge is eNB, where N is the donor concentration and B the effective width of the space-charge region. Substituting for B the expression (I,17a) derived in I, we find that this equality of the two charges is expressed by the following equation:

$$e[X_a^-] = \sqrt{2\varepsilon V_D N e}. \quad \ldots\ldots\ldots\ldots \quad (38)$$

Solving V_D and substituting the result in (37b) we obtain the desired relation between $[X_a^-]$ and the pressure P:

$$[X_a^-] = (k_1 k_2 P)^{\ddagger} \exp(\chi - \zeta - e^2[X_a^-]^2/2\varepsilon N)/kT. \quad \ldots \quad (39a)$$

Since in the case considered here $N = N_c \exp(-\zeta/kT)$ (see I, 2a), we can also write for this:

$$[X_a^-] = (k_1 k_2 P)^{\ddagger} (N/N_c) \exp\{(\chi - e^2[X_a^-]^2/2\varepsilon N)/kT\}. \quad (39b)$$

Although this equation does not explicitly express $[X_a^-]$ in terms of the other quantities, it is not difficult to find the relation between P and $[X_a^-]$ numerically, provided the constants k_1, k_2, N, N_c, ε and χ of the material are known. The total quantity of adsorbed gas is equal to the sum of $[X_a^-]$ and $[X_a]$, and we can therefore find an expression for it (the adsorption isotherm) by adding the quantity $(k_1 k_2 P)^{\ddagger}$ (see 36) to the right-hand side of (39b). From eq. (39b) it can be shown that the adsorption increases when the donor concentration N is increased, thus confirming earlier indications.

Effect of illumination on the gas adsorption

When additional electrons and holes are freed by illumination, the theoretical treatment given in the previous section can no longer be applied. The thermal equilibria discussed there are broken up by the illumination. The steady state which is assumed under constant illumination must be found by considering the processes of electron capture and hole capture separately and then equating the appropriate capture and formation rates.

[10] An example of this is treated by G. Brouwer, in Philips Res. Repts. **18**, 432, 1963.

[11] See A. H. Boonstra, Thesis, Eindhoven 1967 (also published as Philips Res. Repts. Suppl. 1968, No. 3). This gives a description of the effect of gas adsorption on the electrical properties of germanium and silicon.

[12] H. J. Krusemeyer and D. G. Thomas, Phys. Chem. Solids **4**, 78, 1958.

This method of treatment is completely analogous with that given in II for a recombination centre which exchanges electrons with *both* bands in the bulk of a crystal. It may be recalled that, in dealing with substances that had two kinds of such recombination centres, we encountered cases where the energy level occupation changed very markedly in a certain intensity range when the illumination was varied (see II, figs. 13 and 14). In the case we are now considering a change of this nature has various additional consequences. In the first place, owing to the localization of the centres at the surface, there is also a macroscopic spatial redistribution of the electrons when there is a change in the illumination level. This is accompanied by a change in the space-charge distribution in the boundary layer. In the second place a marked change in the state of charge of the surface states will result in a change in the number of adsorbed gas atoms, as explained above.

To illustrate these effects we shall consider two special cases in some detail.

1) Photo-desorption

As our first example we take a photoconductor which exhibits *N*-type dark conduction and in which the holes have a high mobility. We shall show that in this case the illumination has a considerable effect on the adsorption, but only through the holes.

The limited effect of the electrons is due to two circumstances. First, the illumination will cause hardly any relative change in the electron concentration, which is already high in the dark. Secondly, the field acting on the photoelectrons freed in the boundary layer causes them to move away from the surface. See *fig. 9*.

The situation for the holes is quite different. The illumination induces a very marked relative change in their concentration and moreover, under the influence of the band curvature, the holes move *towards* the surface instead of away from it (fig. 9). Many holes will therefore be captured by adsorbed ions, causing these ions to lose their charge. The concentration of neutral atoms on the surface therefore becomes higher, which leads to desorption (the equilibria (34b) and (34a) shift to the left); this process is known as *photo-desorption.*

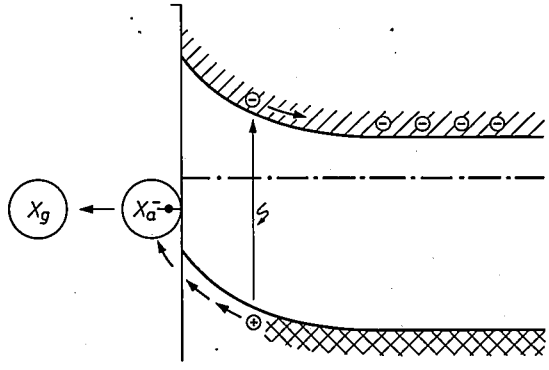


Fig. 9. Photo-desorption of negative ions (X_a^-) adsorbed on an *N*-type semiconductor. The holes freed by the light move towards the surface and discharge the atoms, which then leave the surface ($X_a^- + e^+ \rightarrow X_a \rightarrow X_g$).

The discharging of adsorbed ions by the influx of holes has a further result. Because of the reduction of the surface charge the poorly conducting depletion layer below the surface becomes thinner and the conductivity of the sample is thus increased. In crystals of moderate size this effect is relatively slight, but in thin films, or in samples consisting of powdered material, it may be significant. A considerably greater photocurrent is then found. Effects of this kind have been extensively studied on zinc oxide exposed to a gas mixture containing oxygen [13].

2) Photo-adsorption

As our second example we take a photoconductor which is an insulator in the dark and in which the holes are virtually immobile. Owing to the immobility of the holes the electronic processes at the surface will take place almost entirely by way of the conduction band. In this special case the processes can again be described in terms of equation (34), that is to say on the basis of equilibrium equations. The only difference is that, instead of the Fermi level which applies in the non-illuminated state, we must now use the quasi-Fermi level E_F' of the electrons (see I, p. 124), which depends on the illumination intensity.

Our initial assumption is that in the non-illuminated state the quasi-Fermi level lies so low that adsorbed atoms can take up hardly any electrons. In other words: E_F' then lies below the energy level E_a of electrons that are bound to a gas atom on the surface (*fig. 10a*). Little gas will therefore be absorbed. We further assume that there is no surface charge present, and therefore no band bending.

On illumination the number of conduction electrons increases, and the quasi-Fermi level of the electrons rises. If the intensity of the illumination is increased to such an extent that E_F' rises above E_a, the adsorbed gas atoms can then begin to capture electrons (fig. 10*b*). This has the two results mentioned at the beginning of this section; more gas is adsorbed (the equilibrium (34c) shifts to the right and so too therefore do (34b) and (34a)) and the negative surface charge that appears causes the bands to bend upwards (fig. 10*c*). Thus, the illumination gives rise to the adsorption of gas, called *photo-adsorption*, by a mechanism entirely analogous

[13] See G. Heiland, E. Mollwo and F. Stöckmann in Solid State Physics **8**, 191, 1959, in particular p. 268, D. A. Melnick, J. chem. Phys. **26**, 1136, 1957 or D. B. Medved, Phys. Chem. Solids **20**, 255, 1961.

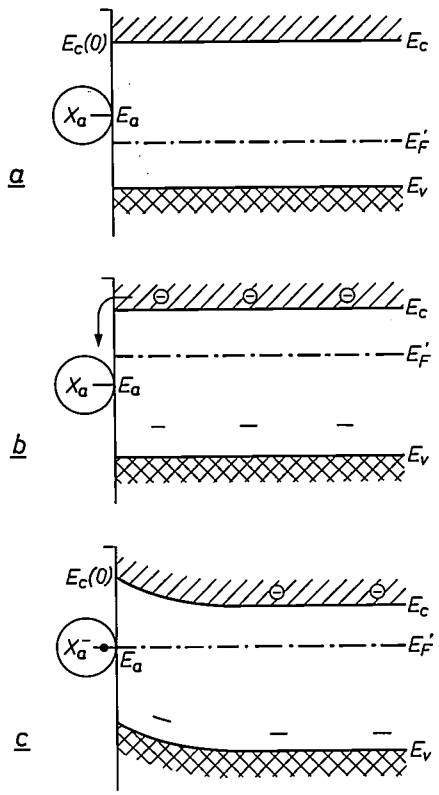


Fig. 10. Photo-adsorption. *a*) Band scheme for semiconductors containing surface-adsorbed gas atoms (X_a) which, owing to the relatively high situation of the available energy level E_a, have not taken up any electrons. (It is assumed first that no other kind of surface charge is present, so that the bands are not bent, and secondly that the holes are immobile.) *b*) After illumination strong enough to cause the quasi-Fermi level E_F' of the electrons to rise above E_a has been switched on, photoelectrons from the conduction band can occupy this level ($X_a + e^- \rightarrow X_a^-$). This gives a curvature of the energy bands (*c*), and the increase of $E_c(0)$ ultimately brings E_a near to the level of E_F', which prevents further transfer of electrons to the adsorbed atoms. Since many atoms have become ionized, fresh gas can be adsorbed.

with the adsorption that results from an increase of the Fermi level by the addition of donors.

To obtain a large readily measurable effect it is advantageous to use light that is strongly absorbed and which therefore only excites electrons in a thin layer at the surface. If the photocurrent is then measured by means of contacts arranged on the illuminated surface, the time variation of the current is found to follow the solid curve shown in *fig. 11*: after the illumination has been switched on the current rises rapidly to a maximum value and then falls relatively slowly to a final value. This happens because the new equilibrium concentration of the electrons after the light has been switched on is established much more quickly than that of the surface charge. This is illustrated schematically in *fig. 12*. Just after switching on the illumination (the variation in intensity of illumination in the crystal is shown in fig. 12*a*) the energy-band scheme is as shown in fig. 12*b*. At the surface there is a downward bending of the bands in the layer in which the light is absorbed, and the boundary layer contains a relatively large number of free electrons. At first there is hardly any change in the state of charge at the surface, but gradually the surface charge increases, and as a result the band at the surface starts to bend upwards, the region in which there are relatively many free electrons becomes narrower, and so the photocurrent decreases. The final situation is illustrated in fig. 12*c*.

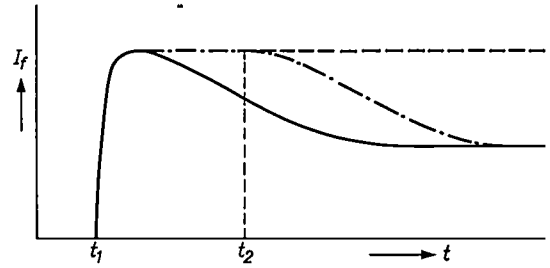


Fig. 11. Variation with time t of the photocurrent I_f between two contacts on the illuminated surface of a photoconductor that strongly absorbs the light. After the illumination has been switched on (at $t = t_1$) the current quickly reaches a maximum value (establishment of electron equilibrium) and then falls slowly to a final value (establishment of a new surface charge). If the experiment is done *in vacuo*, there is no fall in current (dashed line). The decrease does take place, however, if gas is admitted into the evacuated space at the time $t = t_2$ (chain-dashed line).

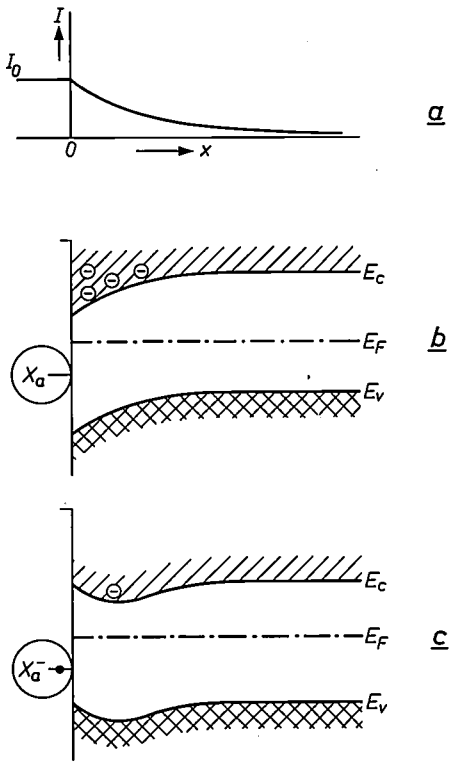


Fig. 12. Establishment of the charge equilibrium in photo-adsorption of gas on a semiconductor that strongly absorbs light. *a*) Variation of the intensity of illumination I under the surface ($x = 0$) of an absorbing substance. *b*) Shortly after the light has been switched on the energy bands bend downwards; no significant change has yet taken place in the state of charge of the adsorbed gas. *c*) After some time the surface charge has had time to adapt itself to the situation produced by switching on the illumination. Some of the additional electrons have again been driven out from a thin layer immediately under the surface, and the bands bend upwards there.

If the experiment is carried out *in vacuo*, the absence of adsorbed atoms means that no surface charges can arise and the current does not decrease (dashed line in fig. 11). If the illumination is started *in vacuo* (at t_1) and gas is suddenly admitted later (at t_2) the current follows the curve shown by the chain-dotted line, since the adsorption process can only start at t_2. The effects discussed have been found on cadmium sulphide crystals in an oxygen atmosphere [14] and also in certain sintered samples [15].

In the subjects discussed in this article we have been concerned with inhomogeneities of one sort or another: inhomogeneity in the crystal, non-uniform illumination or the inhomogeneity due to the finite dimensions of a crystal. This made the situations considerably more complicated than those which are found inside the crystal. Nevertheless, the theoretical treatment was relatively simple, owing to the use of the "ambipolar" approximation.

This approximation has a wide field of application, and is by no means limited to photoconductors, or even to the solid state. The method permits a fairly accurate quantitative treatment of all sorts of otherwise intractable problems, and it has the further advantage of providing a good insight into the many effects that can arise as a result of the interaction of the two kinds of charge carrier. Applications of the "ambipolar" method of approximation outside the field of photoconductivity are found in the theory of transistors and diodes, in the theory of the diffusion of atoms or lattice defects in solids, in the theory of the behaviour of electrolytic solutions and even in the description of processes occurring in gas discharges [16].

It is worth noting that the processes which occur at the surface of a photoconductor exposed to a gas, like those described in this final section, are closely related to effects which occur at interfaces in other fields of physics and chemistry [17]. These effects include catalytic processes in chemistry, boundary-layer effects in colloid chemistry and biochemistry, and also the thermionic emission and photoemission of electrons from metals and semiconductors used in vacuum electronics [18].

In the three parts of this series of articles we have only dealt with those aspects of photoconductivity which appear to us to be of greatest importance. Many other subjects, probably no less interesting, have had to be left out [19]. Regretfully, therefore, we have given no account of the factors determining the mobility of electrons and holes, of the structure and thermodynamics of impurity centres, of the formation of "excitons" or finally, of the interesting effects which occur when charge carriers are injected or extracted under the influence of a very strong electric field [20].

[14] P. Mark, RCA Rev. **26**, 461, 1965.
[15] H. Shear, E. A. Hilton and R. H. Bube, J. Electrochem. Soc. **112**, 997, 1965.
[16] See e.g. R. Papoular, Phénomènes électriques dans les gaz, Dunod, Paris 1963, p. 122.
[17] A survey of these effects has been given by M. J. Sparnaay in: Semiconductor surfaces and the electrical double layer, Adv. Colloid Interface Sci. **1**, 277-333, 1967 (No. 3).
[18] See J. van Laar and J. J. Scheer, Photoemission of semiconductors, Philips tech. Rev. **29**, 54-66, 1968 (No. 2).
[19] Besides the references given in I under [6], see also: S. M. Ryvkin, Photoelectric effects in semiconductors (translation from Russian), Consultants Bureau, New York 1964; T. S. Moss, Photoconductivity, Reports on Progress in Physics **28**, 15-60, 1965; and L. Heijne, Einige physikalische und chemische Aspekte der Photoleitung. Festkörperprobleme 6, 127-173, 1967.
[20] This is treated in: M. A. Lampert, Reports on Progress in Physics **27**, 329, 1964, and in F. Stöckmann, Acta phys. austriaca **20**, 71, 1965.

Summary. This third and last article in a series on photoconductivity discusses effects that are connected with non-uniform illumination, inhomogeneity of the material (*P-N* junctions, contacts, etc.) and the finite dimensions of a sample. Inhomogeneities almost invariably give rise to inhomogeneous distribution of the charge-carrier concentrations, and hence to diffusion. In this diffusion the more mobile of the two kinds of charge carrier carry those of the other kind along with them, but are themselves slowed down in the process (ambipolar diffusion). It can often be assumed to a good approximation in formal analysis that electrons and holes diffuse at the same rate, which may be defined by a single diffusion coefficient. Since their lifetimes are limited, the charge carriers diffuse over an average distance no greater than the "diffusion-recombination length". The ordinary photovoltaic effect occurs when a barrier is illuminated; in the Dember effect an e.m.f. is produced in the direction of diffusion by the difference in the diffusion rates of electrons and holes; the photomagnetoelectric effect is the Hall effect of the two diffusion currents. At a crystal surface the charge carriers may occupy quantum states which differ in terms of energy and capture probability from those in the bulk of the crystal. These surface levels, which may also be due to adsorbed gas, are usually effective recombination centres which can have a very pronounced effect on the charge-carrier concentrations in the layer beneath the surface. Conversely, a change in these concentrations, as a result of illumination, can cause an increase or decrease of the adsorption (photo-adsorption, photo-desorption).

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, Aachen laboratory, Weisshausstrasse, 51 Aachen, Germany	*A*
Philips Zentrallaboratorium GmbH, Hamburg laboratory, 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.

W. I. Acton (University of Southampton) **& M. B. Carson** (Mullard Ltd., Millbrook Works, Southampton): Auditory and subjective effects of airborne noise from industrial ultrasonic sources.
Brit. J. industr. Med. **24**, 297-304, 1967 (No. 4).

N. W. H. Addink: Our knowledge of the d.c. carbon arc. Transition probabilities.
Proc. XIV. Colloquium spectroscopicum internationale, Debrecen 1967, pp. 201-208. *E*

C. S. Aitchison: Possible Gunn-effect parametric amplifier.
Electronics Letters **4**, 15-16, 1968 (No. 1). *M*

C. S. Aitchison, R. Davies & C. D. Payne: Bandwidth of a balanced micropill-diode parametric amplifier.
IEEE Trans. **MTT-16**, 46-47, 1968 (No. 1). *M*

M. Bannink (Philips Radio, Gramophone and Television Division, Eindhoven): Theorie en procesbeheersing van het elektroforetisch verven.
Verfkroniek **41**, 41-51, 1968 (No. 2).

J. Basterfield, M. J. Prescott & B. Cockayne (Ministry of Technology, Royal Radar Establishment, Great Malvern, England): An X-ray diffraction topographic study of single crystals of melt-grown yttrium aluminium garnet.
J. Mat. Sci. **3**, 33-40, 1968 (No. 1). *M*

G. Bergmann: The electrical conductivity of $LiNbO_3$.
Solid State Comm. **6**, 77-79, 1968 (No. 2). *A*

G. Blasse & A. Bril: Investigation of some Ce^{3+}-activated phosphors.
J. chem. Phys. **47**, 5139-5145, 1967 (No. 12). *E*

G. Blasse & A. Bril: Hypersensitivity of the $^5D_0$-$^7F_2$ transition of trivalent europium in the garnet structure.
J. chem. Phys. **47**, 5442-5443, 1967 (No. 12). *E*

R. Bleekrode: Optical determination of Cs ground-state depletion in Cs-Ar low-pressure dc discharges.
J. appl. Phys. **38**, 5062-5065, 1967 (No. 13). *E*

F. J. de Boer & J. Visser: Separation of low concentrations of halogen from some luminescent materials and elemental sulphur by a modified oxygen-flask method.
Analyst **93**, 56-58, 1968 (No. 1102). *E*

R. W. Brander, D. R. Lamb & P. C. Rundle (Associated Semiconductor Manufacturers Ltd., Wembley, England): Electronic conduction in thermally grown silicon dioxide films.
Brit. J. appl. Phys. **18**, 23-27, 1967 (No. 1).

C. M. van der Burgt: Piezoelektrische Werkstoffe und Wandlertypen.
Elektro-Anzeiger, Ausg. ges. Industrie, **20**, No. 19, 24-28, 1967. *E*

J. Cornelissen, G. Piesslinger & A. M. M. de Rijk (Philips Glass Division, Eindhoven): The strengthening of glass-ceramics by ion exchange.
Proc. Symp. sur la surface du verre et ses traitements modernes, Luxembourg 1967, paper No. II-1, 19 pp.

A. J. Fox & J. R. Mansell: Light modulation at video frequencies.
Electronics Letters **4**, 12, 1968 (No. 1). *M*

P. J. van Gerwen: Efficient use of pseudo-ternary codes for data transmission.
IEEE Trans. **COM-15**, 658-660, 1967 (No. 4). *E*

J. M. Goethals & J. J. Seidel (Technische Hogeschool, Eindhoven): Orthogonal matrices with zero diagonal.
Canad. J. Math. **19**, 1001-1010, 1967 (No. 5). *B*

J. C. M. Henning, R. P. van Stapele, G. E. G. Hardeman & P. F. Bongers: Determination of exchange interactions in pairs of coupled Co^{2+} ions in Cs_3ZnCl_5 and Cs_3ZnBr_5.
Magnetic resonance and relaxation, Proc. XIVth Colloque Ampère, Ljubljana 1966, North-Holland Publ. Co., Amsterdam 1967, pp. 1203-1204. *E*

H. Jonker, H. J. Veenendaal (Philips Radio, Gramophone and Television Division, Eindhoven), **C. J. G. F. Janssen & J. O. M. van Langen** (*idem*): PD-photoplating — a new photographic additive printed circuit process.
Proc. Symp. Inst. Metal Finishing, Folkestone 1966, 5 pp.; 1967. *E*

D. R. Lamb & P. C. Rundle (Associated Semiconductor Manufacturers Ltd., Wembley, England): A non-filamentary switching action in thermally grown silicon dioxide films.
Brit. J. appl. Phys. **18**, 29-32, 1967 (No. 1).

B. Libault, R. Havot & J. Devillers: Circuits transistorisés pour oscilloscopes à grande vitesse de balayage.
Acta electronica **10**, 411-428, 1966 (No. 4). *L*

C. Loty: Les tubes à rayons cathodiques à propagation d'onde à très large bande.
Acta electronica **10**, 351-361, 1966 (No. 4). *L*

C. Loty: Distorsions non linéaires des tubes à rayons cathodiques à propagation d'onde.
Acta electronica **10**, 399-409, 1966 (No. 4). *L*

J. Neirynck: Compensation of parasitic capacitances in broadband filters.
IEEE Trans. **CT-14**, 250-259, 1967 (No. 3). *B*

W. Schilz: Experimental evidence of bulk helicon-phonon coupling in PbTe.
Phys. Rev. Letters **20**, 104-105, 1968 (No. 3). *H*

J. Schröder & F. J. Grewe: Preparation of tungsten pentafluoride.
Angew. Chemie: Int. Edit. in English **7**, 132-133, 1968 (No. 2); German Edit. **80**, 118-119, 1968 (No. 3). *A*

P. J. Severin: Superphase velocity effects caused by a single particle.
Physica **36**, 509-524, 1967 (No. 4). *E*

P. J. Severin: Superphase velocity effects caused by a single electron in a magnetoplasma.
Physica **37**, 632-652, 1967 (No. 4). *E*

R. A. Speciale (Philips Cyclotron Development Department, Eindhoven): Accelerating systems with ring-resonator configuration.
Nucl. Instr. Meth. **55**, 205-237, 1967 (No. 2).

J. S. Sussenbach: Early events in the infection process of adenovirus type 5 in HeLa cells.
Virology **33**, 567-574, 1967 (No. 4). *E*

H. J. L. Trap: Influence de certains traitements simples sur la conductivité superficielle de verres contenant de l'oxyde de plomb.
Proc. Symp. sur la surface du verre et ses traitements modernes, Luxembourg 1967, paper No. I-1, 19 pp. *E*

G. N. Wills (Associated Semiconductor Manufacturers Ltd., Wembley, England): Solid state diffusion of antimony in germanium, from the vapour phase, in a vacuum furnace.
Solid-State Electronics **10**, 1-8, 1967 (No. 1).

W. J. Witteman: High-output powers and long lifetimes of sealed-off CO_2 lasers.
Appl. Phys. Letters **11**, 337-338, 1967 (No. 11). *E*

G. Zanmarchi & P. F. Bongers: Magnon sidebands in the optical spectrum of $RbNiF_3$.
Solid State Comm. **6**, 27-31, 1968 (No. 1). *E*

Contents of Philips Telecommunication Review **27**, No. 3, 1968:

Contents of Mullard Technical Communications **10**, No. 91, 1968:

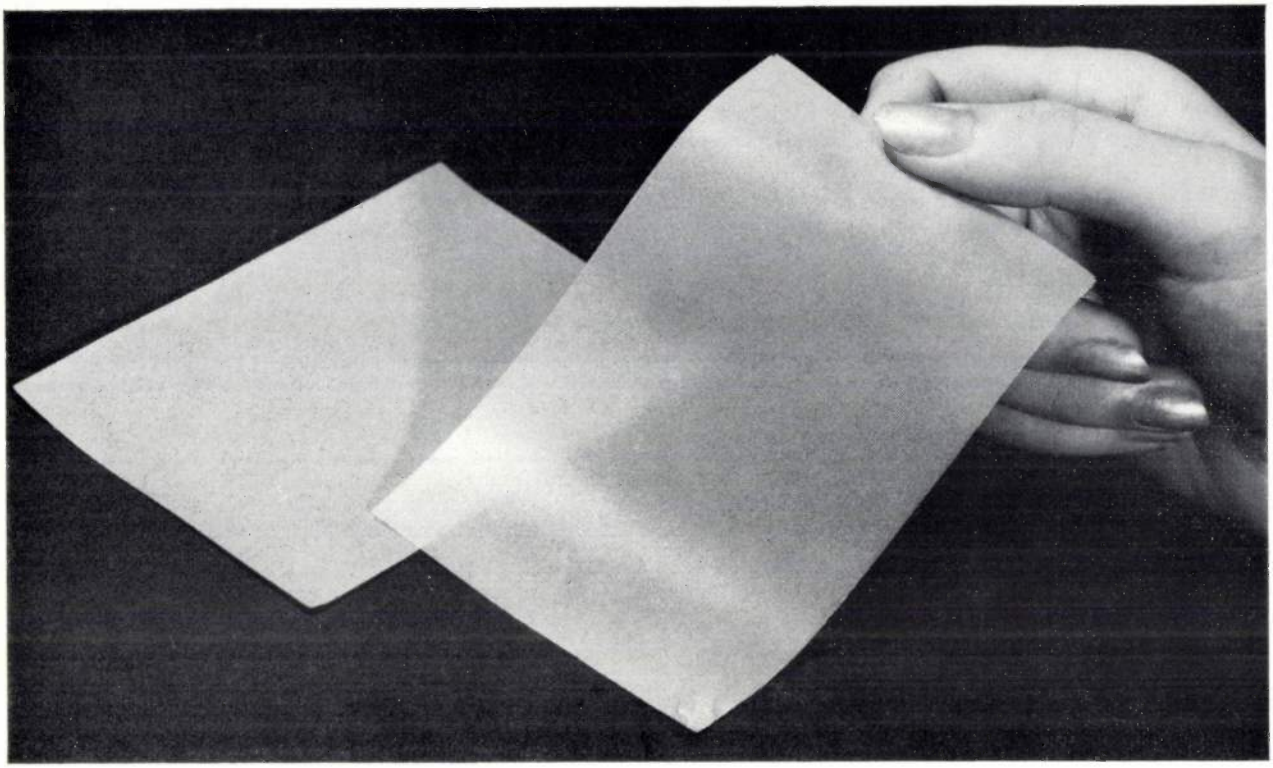
PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 8/9

The completion of a second multi-storey building, which will house about 35% of the present laboratory staff, marks an important milestone in the construction of the new complex of buildings for the Philips Research Laboratories in Eindhoven. This building, which dominates the background of the photograph above, and which, like the others, was designed, built and equipped by the Philips Building Design & Plant Engineering Bureau, was officially opened on 18th September by the Dutch Minister of Education and Sciences. The opening ceremony was attended by many governmental and municipal representatives and by leading figures from Universities and Research Institutes in the Dutch language area.

On the left of the photograph can be seen the first multi-storey laboratory block, which was completed in 1963 (the opening of this block was commemorated by a special issue of our Review at the time — Volume 24, No. 11/12). The low building in the foreground houses the workshops associated with this block; behind it are the cryogenic laboratory (on the left) and the television development laboratory with the colour-television studio. On the right of the new block the Research Laboratories' own computer centre is to be seen, and behind it part of the workshops for the new block.

As a result of the experience gained with the earlier building, the design of the new multi-storey block is a little different. Although it is of the same height, it has eight instead of seven floors above the ground floor, and the offices for the laboratory staff are not situated along one of the long sides of the block (opposite the laboratory rooms) but concentrated in the two end sections of the building. The end sections, which also contain the lift shafts, can be seen in the photograph. Easy personal contact between the staff has again been considered very important, and the design offers good access from floor to floor and from room to room. To give good access between the new and the earlier building there is a covered bridge at first-floor level.



Monograin layers

T. S. te Velde and G. W. M. T. van Helden

The article below describes a new technology for manufacturing solid-state devices in the form of very thin, flexible sheets. At the present time it is mainly being used for photoconductors and solar batteries, but it also offers prospects for the production of resistors, capacitors, diodes, etc.

The principal techniques used today for manufacturing solid-state devices are the planar technique and the thin-film or vacuum-evaporation technique [1]. In the *planar technique* large single crystals of germanium or silicon, possessing a high degree of chemical purity and physical perfection, are sawn into wafers which are subjected to sophisticated methods of oxidation, diffusion and photo-etching in order to form localized regions of *N*-type and *P*-type conductivity in the wafers. This is done simultaneously for thousands of squares measuring 1 mm^2 on each wafer, which are then sawn off and provided with contacts. The electronics industry uses this method for quantity production of diodes and transistors.

Ir. T. S. te Velde and G. W. M. T. van Helden are with Philips Research Laboratories, Eindhoven.

The *thin-film technique* can be applied to many other materials besides germanium and silicon, but the devices thus produced are usually polycrystalline. This limits the scope of the thin-film technique, because the electronic operation of quite a number of solid-state devices is at its best when monocrystalline material is used. This is the case for devices of the transistor type [2], and to a lesser extent for photoconductors, but is not usually true for resistors and capacitors. As a rule, resistors and capacitors can be made more cheaply by the thin-film technique than by the planar technique, and higher values of resistance and capacitance can also be obtained. The thin-film technique is therefore particularly important for the production of resistors and capacitors.

A method has been developed at Philips Laboratories

in Eindhoven which combines certain advantages from these two techniques — the monocrystalline nature of the one and the greater freedom in the choice of dimensions and of electrical parameters of the other. Growing large single crystals is a costly process, and not possible at all for many materials. It occurred to us that the method of making monocrystalline devices is in fact rather a roundabout way of doing things — growing a large crystal and then cutting it up into little wafers — and we came upon the idea of using material in powder form. A powder may consist of small single crystals, and it is possible to prepare the powder in such a way that the crystals are individually physically perfect. In many cases the chemical composition and the size of the powder grains can also be accurately controlled.

In our *monograin* technique a collection of such grains, all of about the same size, are embedded in a sheet of plastic, e.g. polyurethane (a thermosetting resin). The grains are close to each other in the sheet, but not above each other; the layer is nowhere more than one grain thick. It is essential for the grains to be insulated from each other by the plastic and *to protrude from one or both sides of the sheet.*

These light and flexible film-like sheets can be made by a method which is basically very simple and therefore inexpensive. The principle is illustrated in *fig. 1*. A thin layer of adhesive is applied to a rigid, flat substrate (*a*). The grains are then sprinkled over the substrate, and stick to it like flies on a flypaper (*b*). The surplus non-adhering grains are brushed off (*c*) and the substrate with powder is then dipped into a thermosetting resin, which fills up the spaces between the grains (*d*). The tops of the grains can be exposed by etching away the top layer of the resin (*e*). After drying and hardening the resultant sheet is stripped away from the substrate (*f*), and when the remaining adhesive has been removed the grains are also exposed underneath (*g*). Finally, the grains can be connected together, for example by evaporating a conducting metal film (*h*). By applying masks during the evaporation process, groups of the particles can be connected together in any desired pattern in parallel and in series.

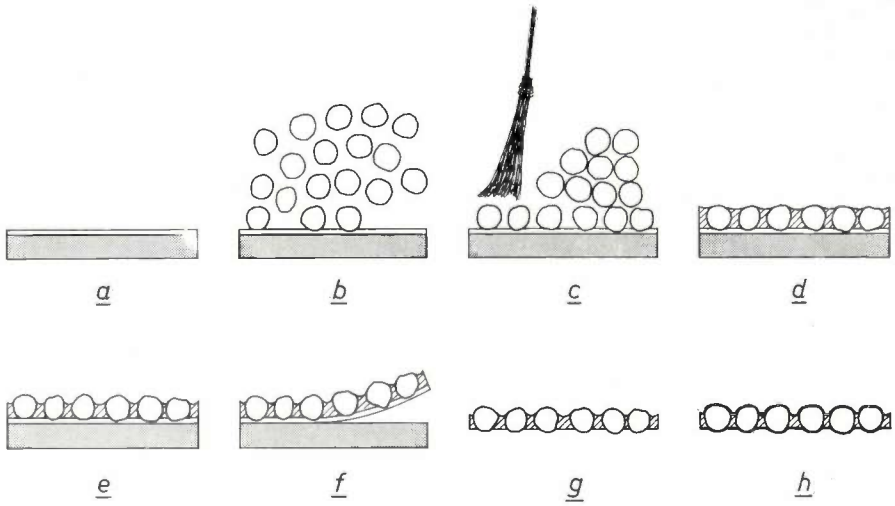


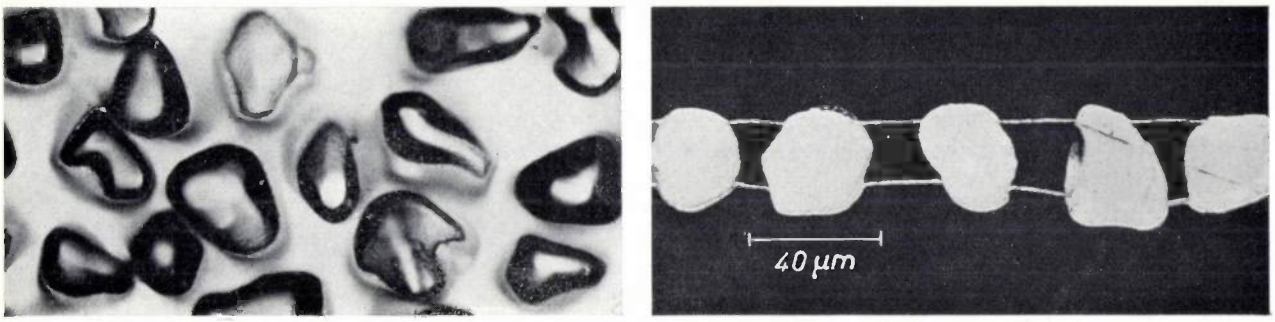
Fig. 1. Diagram showing successive stages in the manufacture of sheets containing monograin layers. *a*) An adhesive is applied to a substrate. *b*) Powder is sprinkled on to it. *c*) The loose, surplus grains are brushed off. *d*) A thermosetting resin is applied. *e*) The top layer is etched away. *f*) The resin is hardened and the resultant sheet is stripped from the substrate. *g*) The adhesive is dissolved and *h*) a thin metal film is deposited on both sides of the sheet.

The title photograph shows two photoconducting CdS sheets produced in this way. A surface view as seen through a microscope is shown in *fig. 2a*, and *fig. 2b* shows a cross-section. The conducting film can be seen clearly in the cross-section.

This new technology opens up new fields in many applications. A wide variety of circuit devices can be made on this principle, such as resistors, capacitors, diodes, photoconductors and solar batteries. In this article we shall discuss as an example the manufacture

[1] For the planar technique see A. Schmitz, Philips tech. Rev. **27**, 192, 1966, and for the thin-film technique E. C. Munk and A. Rademakers, Philips tech. Rev. **27**, 182, 1966.

[2] See H. C. de Graaff and H. Koelmans, Philips tech. Rev. **27**, 200, 1966.



a *b*

Fig. 2. *a*) Surface view and *b*) cross-section of a photoconducting CdS sheet as seen through a microscope.

and characteristics of photoconducting CdS sheets. In the last section we shall consider some of the possible applications of solar-battery sheets made by this new method.

Photoconducting CdS sheets

For making photoconducting sheets, we use cadmium-sulphide grains with a diameter of about 40 μm [3], which are doped with copper and chlorine. A rubber-based adhesive is used and the resin is a polyurethane. After the sheet has been stripped from the substrate and the adhesive removed, an electrode in the form of a gold film is deposited on the side of the sheet which will later be exposed to light; a thin gold film is chosen because it combines high conductivity with low light-absorption. The other side of the sheet is coated with a less expensive conductor such as cadmium.

Between the electrodes and the CdS grains an ohmic contact is required with the lowest possible contact resistance. This is achieved by giving the grains a strongly *N*-type surface before the electrodes are applied [4]. The usual method of doing this, by the diffusion of indium, is not practicable in this case because it requires temperatures at which the polyurethane resin breaks down. We therefore use a gas discharge at room temperature in an argon-filled space, the sheet being given a negative potential so that it can be bombarded with positive argon ions. As a result of the ion bombardment sulphur vacancies occur in the surface layer of the grains (sulphur being more volatile than cadmium, so that more of it evaporates), and this makes the surface strongly *N*-type.

To give a better idea of the characteristics of the photoelectric cell thus obtained — i.e. the *sandwich cell*, with electrodes on both sides of the photoconducting material — we shall compare it with the hitherto much more widely used *gap cell*, in which the electrodes are applied on one side of the photoconducting material, in the form of parallel strips or interlocking combs.

An important characteristic of a photoelectric cell is the radiation sensitivity, i.e. the ratio of its output photocurrent to the incident radiant flux. The radiation sensitivity of a photoelectric cell is inversely proportional to the square of the distance between the electrodes, and also depends on the voltage between the electrodes, the lifetime and mobility of the charge carriers, etc. Small electrode distances have hitherto been achieved with the least expense in photoelectric cells of the gap type. For example, in gap cells of pressed and sintered CdS [5] the interelectrode spacing can be reduced to 200 μm. Further reduction of the interelectrode spacing in this type of cell gives hardly any further increase in the radiation sensitivity, since the resistance of the electrodes themselves and their contact resistance become relatively too high. Because of the different electrode configuration, the resistance of the electrodes in our photoconducting CdS sheet is much smaller in comparable conditions. What is particularly important, however, is that their *contact resistance* has been made extremely small by using the special method of making the contacts. It has been found possible to make CdS sheets which have negligible contact resistance at an interelectrode distance of 40 μm. This means that, compared with gap cells of the same material, with the same voltage between the electrodes and the same penetration depth for the radiation, etc., the sensitivity can be made 25 times higher.

In this comparison we have not yet taken into account the fact that the CdS grains are monocrystalline, unlike the sintered cadmium sulphide of the gap-type photoconductor. Because of the monocrystalline nature of the CdS, the monograin sheets do not show the troublesome after-effects connected with the presence of grain boundaries [6].

Fig. 3 shows the spectral sensitivity curve of a typical CdS sheet.

Solar-battery sheets

Many laboratories throughout the world are working on *solar batteries* for the conversion of solar energy into electrical energy. There is considerable scope for the application of such batteries in space vehicles. The types used for this kind of application are generally made of flat wafers of monocrystalline silicon incorporating a *P-N* junction [7]. The electrons and holes released near the *P-N* junction by the incident radiation flow to opposite sides of the junction. This produces an e.m.f. which causes a current to flow when an external circuit is connected.

These monolithic solar batteries, as they are called, are rather expensive. There are various reasons for this:
1) The expense of growing large single crystals of silicon and then cutting them up into wafers. This point was mentioned in the introduction.
2) To cut down on the costs of material, the silicon wafers should be as thin as possible. But the thinner, and therefore more fragile, the wafers are made, the greater the expense of processing, making the contacts, etc. Generally speaking, therefore, the only course left open is to use silicon wafers that are thicker than is really necessary for the purpose they are meant to serve, i.e. the generation of energy by the absorption of solar radiation.

This also means that these monolithic solar batteries are *heavier* than is necessary for their actual function. This is a particular drawback for space-vehicle applications, certainly for future developments such as the moon project, where very high powers will be needed and any extra weight must be cut to an absolute minimum. One can therefore understand why efforts to make solar batteries from evaporated thin films are already under way. Some loss of efficiency due to the polycrystallinity of the evaporated material then has to be accepted. We think that monograin layers offer an attractive alternative. With this method light and flexible sheets of any required size can be produced, which can be rolled up before the space vehicle is launched, and then unrolled in orbit.

We have made sheets of CdS grains covered with an epitaxially grown layer of $Cu_{2-\delta}S$ (with small value of δ) [8]. The CdS was doped so as to make it a strong *N*-type conductor, and the $Cu_{2-\delta}S$ formed the *P* region (*fig. 4*). The solar radiation is absorbed by the $Cu_{2-\delta}S$ layer. Since this material has a particularly high coefficient of absorption, compared with silicon, the layer can be made very thin so that the sheet is extremely light. Work is now in progress with a view to producing these "solar sheets" in a practical form.

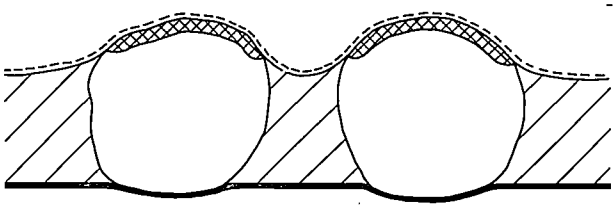


Fig. 4. After the strongly *N*-type CdS grains are embedded in the resin, a *P*-type layer of $Cu_{2-\delta}S$ is grown epitaxially on the protruding surfaces of the grains. On the side where the light will be incident (top), the grains are connected together by a transparent gold film.

If the direct conversion of solar energy into electrical energy also becomes a successful economic proposition for non-space applications, the prospects opened up will be considerable. Many desert regions remain barren because there is no cheap source of energy available for pumping existing underground water up to the surface. The abundance of solar energy for which these regions are noted makes this an obvious choice.

The economic conversion of this energy, however, is no simple matter, as illustrated by the following figures. The efficiency obtained from monolithic solar batteries of monocrystalline silicon at the present time is between 10 and 15%. Taking this efficiency and the present-day cost of material, it can be shown that it would cost several hundred thousand pounds per kW to install such a "solar generator", i.e. about a hundred times the installation costs per kW of a diesel generator.

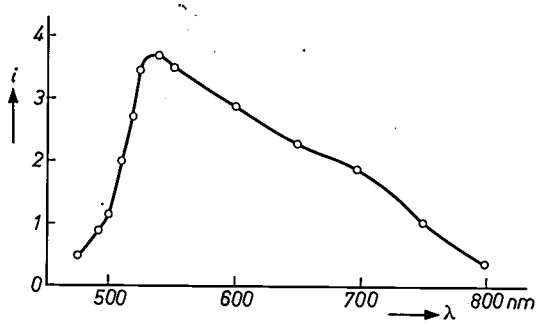


Fig. 3. Spectral sensitivity curve of a photoconducting CdS sheet. The photocurrent *i* is indicated in arbitrary units. (The measurements were made by J. H. Haanstra of these Laboratories.)

[3] Our procedure was based principally on the work of Drs. W. G. Gelling of these Laboratories.
[4] See F. A. Kröger, G. Diemer and H. A. Klasens, Phys. Rev. **103**, 279, 1956, and also L. Heijne, Philips tech. Rev. **25**, 120, 1963/64, in particular fig. 8.
[5] See N. A. de Gier, W. van Gool and J. G. van Santen, Philips tech. Rev. **20**, 277, 1958/59.
[6] This was demonstrated by Dr. P. C. Scholten of these Laboratories.
[7] See L. Heijne, Physical principles of photoconductivity, III. Inhomogeneity effects, Philips tech. Rev. **29**, 221-234, 1968 (No. 7), in particular fig. 1*c*.
[8] For research on monolithic solar batteries of $CdS/Cu_{2-\delta}S$, see F. A. Shirland, Advanced Energy Conversion **6**, 201, 1966.

Even at the experimental stage reached at present costs can be reduced considerably by using sheets instead of monolithic batteries.

To demonstrate the possibilities of the "solar sheets", we have recently built, for an exhibition, a miniature installation consisting of a generator powered by these sheets and driving a pump system. Another demonstration model is the little "cab" shown in *fig. 5*, which is driven by two electric motors powered by two solar battery sheets. The left-hand sheet supplies the power for the right-hand wheel and the right-hand sheet the power for the left-hand wheel. (With this simple device not only is the cab set in motion when it is placed in a beam of light, but it also keeps itself in the beam. If, for example, the left-hand sheet leaves the light beam as the cab moves forward, the *right-hand* wheel is driven at a lower speed and so the cab is steered in the correct direction.)

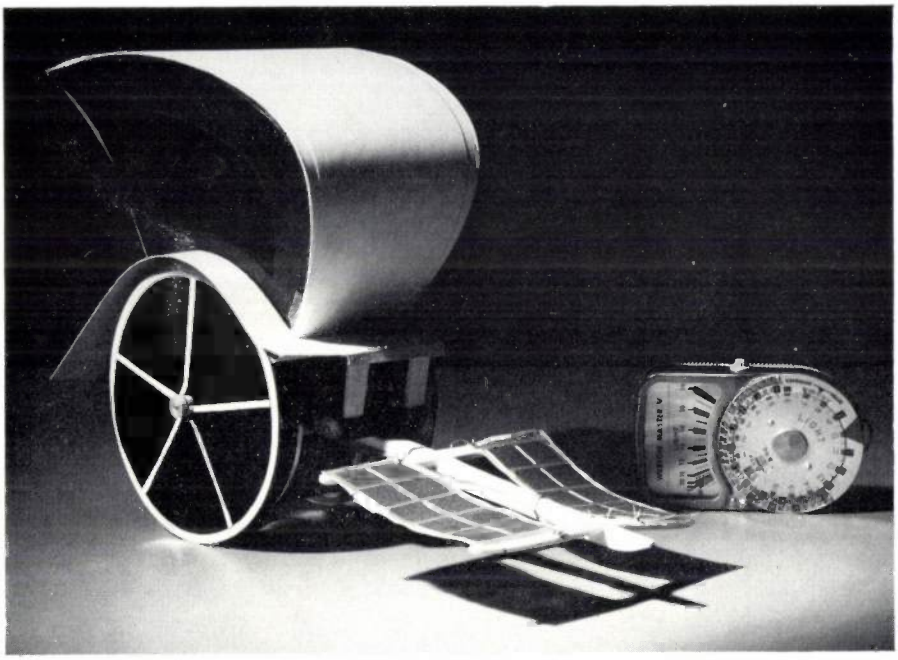


Fig. 5. Two-wheel toy "cab" driven by two miniature electric motors, fitted under the axles and powered by two "solar sheets". The left-hand sheet supplies the right-hand wheel and the right-hand sheet the left-hand wheel. The cab moves forward when placed in a beam of light (and the cross-over in the feed keeps it within the beam).

These examples demonstrate some of the possible applications of solar sheets for toys, but it looks as if space applications will remain the principal field for the time being. Whatever the developments in space travel — space stations and space laboratories in orbit, astronauts on the moon, perhaps even the exploitation of the mineral resources of the moon — there will be a growing need for high powers from equipment which does not require fuel to be carried and should be as light as possible.

Summary. Powders often consist of monocrystalline grains, which can be given a high degree of physical perfection and permit good control of chemical composition. A technique is described in which powder grains are embedded in an insulating sheet of plastic, e.g. polyurethane, with each grain protruding on either side. The protruding parts of the grains of such a monograin layer can be connected together electrically in any desired pattern, for example by evaporated metal. The method offers an inexpensive means of making monocrystalline circuit devices. Particular attention is given to photoconducting CdS films made in this way, and to solar-battery sheets, which appear very attractive for space-vehicle applications.

A delay line for PAL colour television receivers

F. Th. Backers

In colour television receivers for the PALd system the colour information is derived from the chrominance signal for each line and from the delayed signal of the preceding line of the field. If colour errors are to be avoided a delay line that gives an extremely accurate phase delay must be used. The subject of the present article is an ultrasonic delay line developed for this purpose, made from a type of glass specially developed for this delay line.

The Oslo Conference of 1966 was unfortunately unable to agree on the use of one single colour television transmission system for the whole of Europe. In the meantime most European countries have made their choice of system. The Netherlands, Germany, Great Britain, Ireland, the Scandinavian countries, Switzerland and Austria have all adopted the PAL system; France, Russia and a number of East European countries have chosen the SECAM system. These systems have previously been discussed at some length in this journal [1]. In receivers for both systems, *delay lines* are used, which delay the chrominance signal by a time virtually equal to the period of the picture lines (64 μs). The use of such a delay line is essential in a receiver for the SECAM system. A PAL receiver can work without a delay line (PALs receiver), but the PALd reception system, which does use a delay line, offers certain advantages. A discussion of this subject will be found in reference [1].

The subject of the present article is an ultrasonic delay line developed for PAL colour television receivers at Philips Research Laboratories in Eindhoven [2]. The operation of this line depends on the well-known method of propagating ultrasonic vibrations in a piece of suitably shaped material, and it is made with a type of glass specially developed for the purpose.

First of all, we shall recall the principle underlying the decoder used in a PALd receiver. *Fig. 1*, corresponding to fig. 9 in reference [1], gives the vector diagram of the chrominance signal of two successive lines of a field (assuming of course that the colour between the two lines does not change). This chrominance signal consists of two components, $0.49(B' - Y')\sin\omega_s t$ and $0.88(R' - Y')\cos\omega_s t$, each of which contains colour information. Here B' and R' are the gamma-corrected blue and red signals respectively, and Y' is a combination of these signals and the green signal G':

$$Y' = 0.30\,R' + 0.59\,G' + 0.11\,B'.$$

ω_s is the angular frequency of the subcarrier. The func-

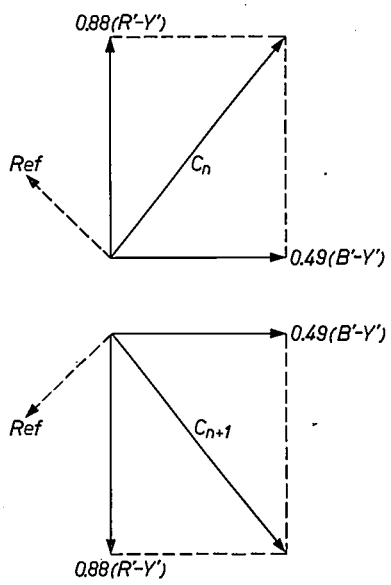


Fig. 1. Vector diagram of the chrominance signals C_n and C_{n+1} of two successive lines with the same colour, in a PAL transmission. The phase of the component $0.88(R' - Y')\cos\omega_s t$ differs in each successive line by 180°. *Ref* indicates the simultaneously transmitted burst, which is used for synchronous detection of the chrominance signals (see [1]).

Dr. Ir. F. Th. Backers, formerly with Philips Research Laboratories, is now with the Philips Radio, Gramophone and Television Division (RGT), Eindhoven.

[1] F. W. de Vrijer, Colour television transmission systems, Philips tech. Rev. **27**, 33-45, 1966. See also reference [4], p. 247 *et seq.*

[2] The delay line in question can also be used in SECAM receivers. The requirements to be met by the delay line in this system are in many respects not so strict as in the PAL system.

tion of the decoder is to split the chrominance signal into these two components. In a PALd receiver this is done by forming the sum and the difference of the chrominance signal of each line and the delayed signal from the preceding line. We shall first make the usual assumption that the sum yields the $(B' - Y')$ component, and the difference yields the $(R' - Y')$ component. Since in the PAL system the component $(R' - Y') \cos \omega_s t$ changes sign at every alternate field line, the same change of sign takes place in the corresponding signal delivered by the decoder. This change of sign is subsequently eliminated in another part of the receiver.

Requirements to be met by a PAL delay line

Phase delay and group delay

Since receivers for the PALd system derive the colour information for each line from the line itself and from the preceding one, proper colour rendering is only possible at places where the colour at corresponding points of two successive field lines is the same. This restriction is not a disadvantage in practice, and we shall therefore take it as a basis in formulating the requirements for the delay line. We shall first assume that the two lines have the same colour over the whole of their length. The chrominance signal then has a constant amplitude and phase (neglecting the brief interruptions during the line-synchronizing pulse).

Let us first assume that the phase delay is *exactly* 64 μs; the decoder is then supplied with two signals whose vector diagram is as shown in *fig. 2*. The phase relation between the delayed signal $C_{n\tau}$ and the non-delayed signal C_n in fig. 1 is found from the relation between the line frequency f_L (15 625 Hz) and the subcarrier frequency f_s. This relation in the PAL system is given by:

$$f_s = (283\tfrac{3}{4} + \tfrac{1}{625}) f_L = 4\,433\,619 \text{ Hz}.$$

The phase delay of 64 μs thus corresponds to a phase shift of $64 \times 10^{-6} \times 4\,433\,619 = 283.752$ periods, and $C_{n\tau}$ will therefore lead in phase with respect to C_n by $0.248 \approx \frac{1}{4}$ period, i.e. about 90°.

It can be seen from fig. 2 that in this case the two components of the chrominance signal are *not* obtained by adding or subtracting $C_{n\tau}$ and C_{n+1}; a phase delay of exactly 64 μs is therefore of no use for this method of signal splitting. To obtain the desired result it is necessary to choose the phase delay such that the $(B' - Y')$ component of the delayed signal $C_{n\tau}$ is *in phase* with the corresponding component of the signal C_{n+1} of the next line. This means that the chrominance signal C_n in the delay line must undergo a phase shift of an integral multiple of 2π radians. This situation arises when the phase shift is exactly 284, or $284 \pm p$, periods (p being an arbitrary integer), corresponding to a phase delay of

$$\tau = (284 \pm p)/f_s = 63.943 \pm (p \times 0.226) \text{ μs}.$$

It is not absolutely essential, however, to obtain the $(B' - Y')$ component by the *addition* of $C_{n\tau}$ and C_{n+1}, and the $(R' - Y')$ component by *subtraction*. The

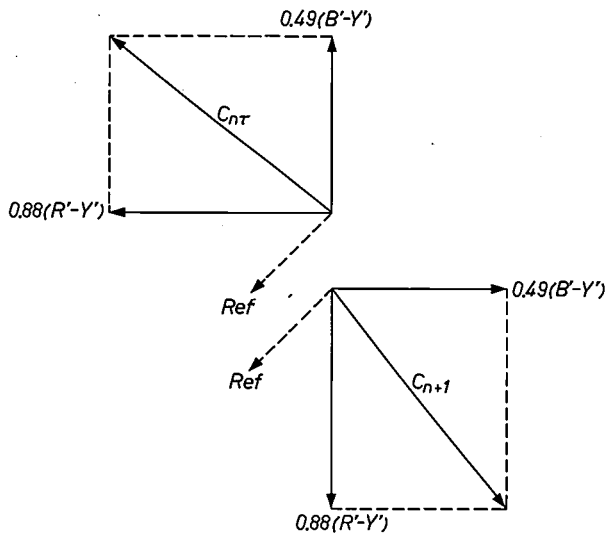


Fig. 2. The chrominance signal $C_{n\tau}$ of a line, delayed by exactly 64 μs, together with the non-delayed signal C_{n+1} of the next line. In this case the two components can *not* be obtained by simple addition and subtraction of the two signals.

signal C_n can also be given a delay such that the addition of $C_{n\tau}$ and C_{n+1} gives the $(R' - Y')$ component and subtraction gives the $(B' - Y')$ component. Obviously, the phase shift produced by the delay line must then be an odd multiple of π radians. The vector diagram of the signals for this case is shown in *fig. 3*.

In a PALd receiver a deviation from the phase delay quoted above causes the same impairment of the picture as that produced in a PALs receiver by differential phase errors, i.e. apparently moving strips, known sometimes as "Hanover bars" or "Venetian blinds". Since, where there is an error in the delay line, only the delayed signal has the wrong phase, the permissible phase error in this case is twice as great as the permissible differential phase error in the PALs system. Experiments have shown that the phase error in a PAL delay line may be allowed to reach ± 12°, corresponding to a permissible deviation in the phase delay of about 7.5 ns.

If the area being scanned is not all of the same colour, the two components of the chrominance signal are modulated in amplitude. Signal components then

occur with frequencies on either side of the subcarrier frequency. These sideband components lie approximately in the frequency range 4.43 ± 1 MHz. If two successive lines show the same colour variation, which is the case, for instance, when the picture consists of vertical coloured bars, sideband components then occur only at frequencies that are an integral multiple of the line frequency (15 625 Hz) higher or lower than the

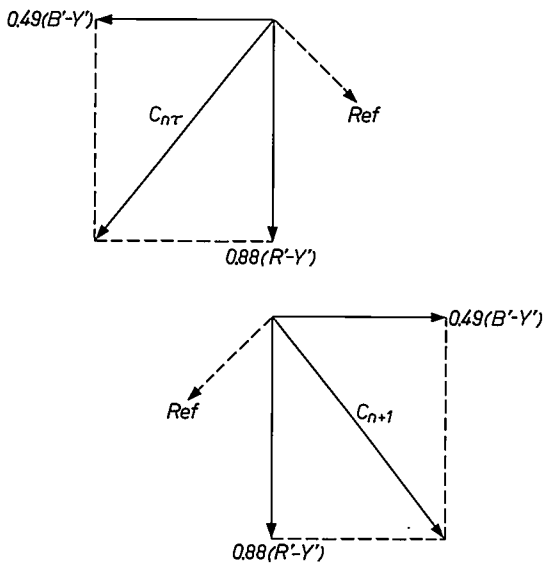


Fig. 3. When the delay line causes a phase shift of an odd multiple of π radians, the $(R' - Y')$ signal can be obtained by the addition of the $C_{n\tau}$ and C_{n+1} signals, and the $(B' - Y')$ signal by their subtraction.

subcarrier frequency. The phase shift which these sideband components undergo in the delay line is different from that of the subcarrier. The whole signal will, however, be delayed without distortion if the phase shift for all components is an integral multiple of 2π radians. This is illustrated in *fig. 4* for a special case. For simplicity it is assumed here that the colour difference signals $R' - Y'$ and $B' - Y'$ consist of a d.c. component on which a sinusoidal signal is superimposed. In this case the two components of the chrominance signal C_n each contain only two sideband components (fig. 4*a*). If we assume further that, owing to errors in the transmitter, in the transmission or in the receiver, the two sideband components have undergone a change in magnitude and in phase with respect to their carrier, the vector diagram C_n^* of the input signal of the delay line may take the form illustrated in fig. 4*b*. If all the components undergo a phase shift of an integral multiple of 2π radians, then fig. 4*b* also represents the output signal $C_{n\tau}^*$ of the delay line. The undistorted signal C_{n+1} of the next line is shown in fig. 4*c*; if this signal undergoes the same distortion as C_n, the result is then C_{n+1}^* (fig. 4*d*). This latter signal is then available simultaneously with $C_{n\tau}^*$. By the addition and subtraction of $C_{n\tau}^*$ and C_{n+1}^* we obtain once again the two components of the chrominance signal.

Fig. 4 also illustrates an incidental, though not unimportant, advantage of the PAL system. If the chrominance signal has undergone linear distortion, this need not lead to crosstalk between one component and the other (quadrature cross-colour), as it would in an NTSC decoder. The same advantage is obtained with a PALs receiver, provided at least that the eye effectively averages out the delayed and non-delayed signal. Averaging-out by means of a delay line works much better, however.

It is perhaps interesting to note here that the poor averaging-out ability of the eye, giving rise to quadrature cross-colour and "Hanover bars" in a PALs receiver, is probably largely due to

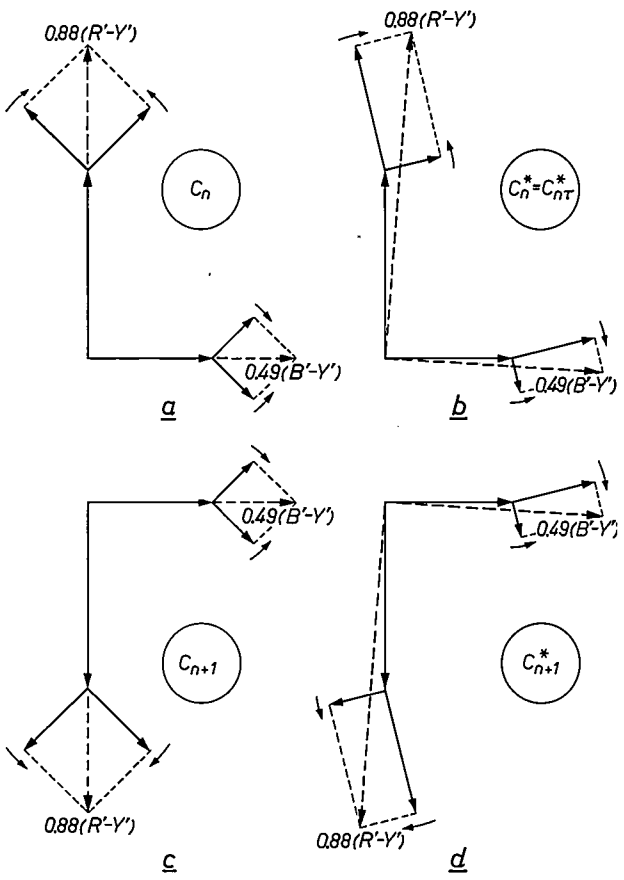


Fig. 4. *a*) Vector diagram of the two components of the chrominance signal C_n where the colour difference signals consist of a d.c. part and a purely sinusoidal part. Each component of the chrominance signal contains two sideband components.
b) The components of the chrominance signal C_n^* obtained when the sideband components show an error in magnitude and phase. If the phase shift for all components is an integral multiple of 2π rad, this figure also represents the delayed signal $C_{n\tau}^*$.
c) The undistorted chrominance signal C_{n+1} of the next line.
d) The signal C_{n+1}^* obtained from C_{n+1} when the sideband components here also show an error in magnitude and phase. This distortion does not lead to quadrature cross-colour in a PALd decoder.

the fact that, because of the interlacing, the interfering effect appears to move in a vertical direction. If the eye follows this movement, there can indeed be no question of averaging-out. This would also explain the fact that this effect is not found to be such a nuisance in the parts of the picture which are rich in detail, where the movement is constantly interrupted.

Generally speaking, the modulation of the chrominance-signal components will not be purely sinusoidal and there will therefore be more than two sideband components. A signal of this kind is delayed without distortion if the phase shift per 15 625 Hz increases by 2π radians. The phase characteristic (phase shift as a function of angular frequency) is then a straight line with a slope of 2π rad/$(2\pi \times 15\,625$ rad/s$) = 64$ μs. A second condition for the phase characteristic has already been mentioned, i.e. that at the subcarrier frequency the phase shift must be an integral multiple of π rad. *Fig. 5* shows a number of characteristics (the solid lines) which satisfy these two conditions. The

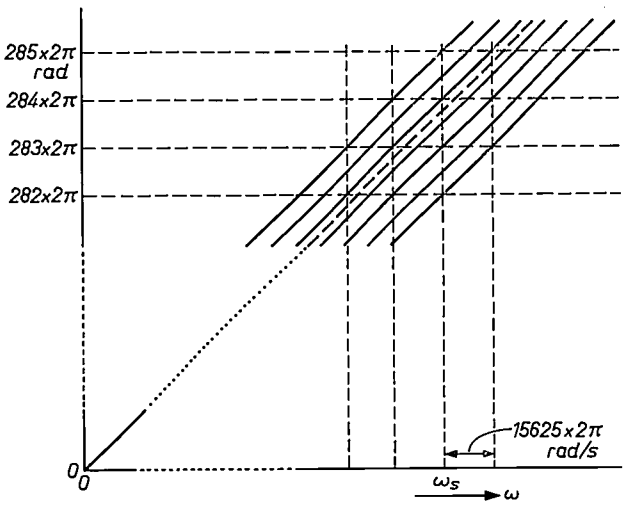


Fig. 5. Phase characteristics of the delay line. For all lines the group delay is 64 μs. The dashed line for the subcarrier does not correspond to a phase shift amounting to an integral multiple of π rad and therefore cannot be used. The solid lines satisfy the necessary conditions.

dashed line, which passes through the origin and corresponds at the subcarrier frequency to a phase delay of exactly 64 μs, does not satisfy the second condition, for although its slope has exactly the right value, the phase shift at the angular frequency ω_s is $283.752 \times 2\pi$ radians.

The slope of the phase characteristic of a network is equal to the *group delay* [3]; therefore the requirements to be met by a PAL delay line may also be defined as follows. The phase shift for the subcarrier must be an integral multiple of π radians and the group delay must be 64 μs. The accuracy of the group delay, however, is not so critical as that of the phase shift. A deviation of ± 50 ns in the group delay has been found to be permissible. The phase characteristic may also deviate slightly from a straight line.

Unwanted reflections

It is possible that signals will arrive at the output transducer of the delay line which have traversed the delay line more than once, as a result of unwanted reflections, and have thus undergone a delay several times longer than the desired delay. These reflections are referred to here as 2τ or 3τ reflections, etc. The colour information which these signals contain thus arrives one or two picture lines too late. Since in the PAL system the $(R' - Y')$ component changes sign at successive picture lines, a 2τ reflection is much more of a nuisance than a 3τ one, for a signal that originates in a 3τ reflection is of course added to the signal of a line of the same type. In equally coloured areas this effect is not visible; interference is seen only at places where vertical colour transitions are present. The signal of a 2τ reflection, however, is added to a line of the other type. This can give rise to an effect rather like "Hanover bars" in patches of certain colours. For this reason 2τ reflections need to be much more strongly suppressed than 3τ reflections.

Other requirements

Finally, we shall mention a few other requirements which have to be met by a PAL delay line.

The *amplitude characteristic* must of course be reasonably flat. A 3 dB bandwidth of 1.8 MHz has been found acceptable, the centre frequency being of course at 4.43 MHz.

The *insertion loss* must obviously be as small as possible to avoid loss of signal strength.

The *temperature* at which the delay lines in television receivers are used is between 20 and 50 °C. All the above-mentioned requirements must therefore be met within this temperature range. This is a difficult requirement, particularly for the phase delay, and it has been met by the use of a specially developed type of glass. Rather larger temperature variation, from —20 to +70 °C, should not give rise to any permanent changes in line characteristics.

Design

The design of ultrasonic delay lines and their various applications in television were dealt with in this journal a few years ago [4]. It was mentioned at that time that a piece of glass of suitable form and dimensions could be a very useful medium for delaying an electrical signal by one line period. The glass block of the delay line which we have developed measures about $80 \times 40 \times 18$

mm (*fig. 6*). The *input transducer* T_1 converts the electrical input signal into mechanical shear waves, which are propagated in the glass in the direction denoted by *A*. The waves reflected from the face *F* then move along the line *B* and so arrive at the *output transducer* T_2, where they are converted back into electrical signals.

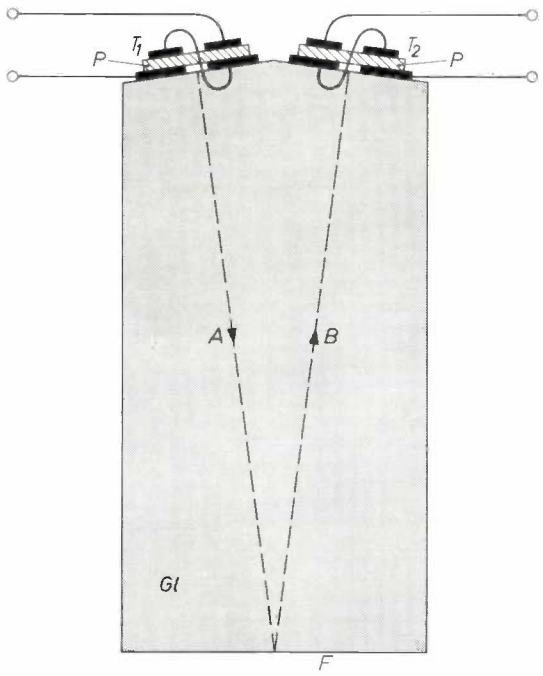


Fig. 6. Form of the delay line. *Gl* glass block. T_1 input transducer. T_2 output transducer. *P* piezoelectric plates. *A* and *B* direction of propagation of the ultrasonic shear waves. *F* reflecting face.

The transducers consist of thin plates *P* of ceramic piezoelectric material (a solid solution of lead zirconate-titanate, code name PXE 3). The large faces of the plates are provided with vacuum-deposited electrodes. To give it piezoelectric properties, the material has to be polarized. This is done by raising the transducers to a high temperature in a strong electric field which is maintained while the material is cooled down. The direction of polarization is perpendicular to the plane of fig. 6. Within a certain frequency range, determined by the thickness of the material, the transducers can then excite and receive shear waves. The material PXE 3 transmits these waves to the glass with fairly low insertion loss, permitting the desired bandwidth to be obtained.

If the large faces of a plate of the dimensions used here ($10 \times 12 \times 0.23$ mm) were to be completely covered with metal films, the impedance of the transducer would be very low. For this reason each electrode is split into two equal parts down the centre (see fig. 6), thus effectively making two transducers which form a single mechanical entity and produce coherent mechanical vibrations. The two parts are connected electrically in series, increasing the impedance by a factor of 4. The same increase of impedance could of course also have been obtained by making the area of the surface four times smaller. In that case, however, the directivity pattern of both the transmitting and the receiving transducer would have been less sharp, and the transmission consequently less efficient.

To facilitate the mounting of the transducers on the glass block, a metal coating is applied to the faces where the transducers are to be attached. This coating takes the form of a divided patch on each face, to suit the electrodes of the transducers, which are then soldered in position.

The main advantage of using a reflecting face is that it offers a simple means of meeting the very close tolerance required for the phase shift. By grinding away the reflecting face in this way the path length can be reduced after the transducers have been mounted. The phase shift can be measured during the grinding operation, which is continued until the phase shift is an integral multiple of π radians. We shall return to this point presently.

Properties of the glass

The phase delay of a delay line constructed as illustrated in fig. 6 consists of two parts. By far the greater contribution comes from the glass block, and a much smaller one from the transducers, the generator and the load to which the transducers are connected. Both parts of the phase delay vary with temperature, and here again the glass block makes the greater contribution to this variation. The phase delay in the glass depends both on the dimensions of the glass and on the velocity of propagation of sound in it. This is determined by the specific mass of the material and, if shear waves are employed, by its rigidity modulus. Since all these properties are temperature-dependent, the glass is not required to have a zero coefficient of expansion; but it should have properties such that the effect of any expansion due to temperature variations is compensated by the change in the velocity of propagation. A type of glass that meets this requirement has been developed at the Philips Glass Development Centre. It has in fact been found possible to give the glass properties which enable it to compensate for the effect of temperature on all of the contributions to the phase delay. This type of

[3] Also known as modulation phase delay, since the delay is that of the envelope of a modulation signal; see H. J. de Boer and A. van Weel, An instrument for measuring group delay, Philips tech. Rev. **15**, 307-316, 1953/54.

[4] C. F. Brockelsby and J. S. Palfreeman, Ultrasonic delay lines and their applications to television, Philips tech. Rev. **25**, 234-252, 1963/64.

glass has been used in the PAL delay line. Details of the composition and properties of this glass will be found in an article by Zijlstra and Van der Burgt [5]. We shall just mention here that in addition to SiO_2 the glass contains about 48% PbO and 3% K_2O, and that it is subjected to a special heat treatment to give it the required properties.

Owing to the elastic after-effect in the glass and the transducers, it is not possible to give a unique relation between phase delay and temperature. A better idea of the results achieved can be given by quoting the magnitude of the maximum deviation in phase delay which is found when the delay line is subjected to the kind of heating cycle that would be experienced in practice. When the delay line is raised in temperature linearly from 20 to 50 °C in three hours and then kept at 50 °C for two hours, the phase delay varies by no more than $\pm$ 5 ns compared with the value at 25 °C.

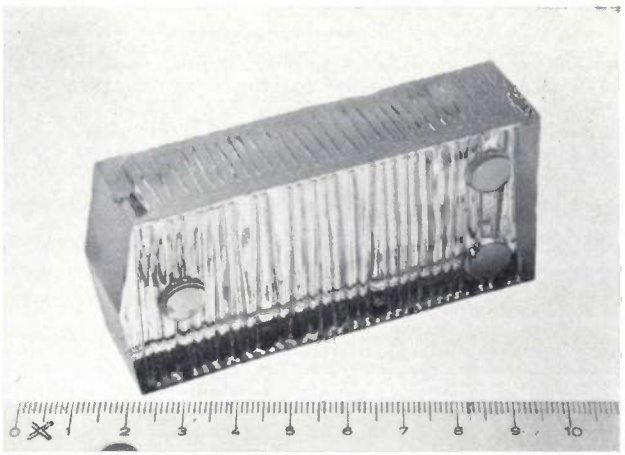


Fig. 7. The glass block of the delay line. The side walls have been "roughened" to minimize unwanted reflections of the ultrasonic waves. The three small protruding faces are used for fixing the block when the reflecting face is ground away.

Suppression of unwanted reflections

Only a negligibly small part of the acoustic energy in the glass escapes through the walls. This is due to the very great difference between the acoustic impedances (the product of density and propagation velocity) of glass and of air, and also to the fact that shear waves are used, which cannot be propagated as such in a gas. Measures have to be taken, however, to ensure that the acoustic waves, which are reflected one or more times from the side walls, do not reach the output transducer. These unwanted waves cover a distance which is longer (often very much longer) than the required path, indicated in fig. 6. This can give rise to irregularities in the amplitude characteristics and in the phase characteristic, and these irregularities can cause colour errors. This effect is opposed by making the four side walls of the glass block much "rougher" than would be necessary to make them non-reflective to light. Any acoustic waves which arrive at these walls from an incorrect path are thus scattered rather than being truly reflected. The grooves made in the glass (about 1 mm deep) can clearly be seen in the photograph in *fig. 7*. They do not follow any ordered pattern. (They have been omitted in fig. 6 for simplicity.)

In general, a certain part of the signal arriving at the output transducer will be reflected. After having covered the complete return path, this signal again arrives at the input transducer, where a further partial reflection takes place. As a result there appears at the output transducer a weak signal which has passed three times through the glass block: this is the 3τ reflection mentioned earlier, also called the "third-time-round" signal.

On page 246 we also mentioned a signal that can appear at the output transducer after having traversed the glass block *twice* ("second-time-round" signal or 2τ reflection). If it is assumed that the waves only propagate along the lines marked *A* and *B* in fig. 6, one might conclude that such a signal would arrive only at the *input* transducer. In practice, however, the wave fronts extend some way; after 2τ seconds the whole cross-section of the glass is occupied so that the *output* transducer also receives part of the energy. Fortunately in this case the transducer is not parallel to the wave front, which means that it is much less sensitive to the 2τ reflection than to the wanted signal.

To reduce the 3τ reflection to a negligible amplitude, the face of the output transducer T_2 (see fig. 6) is shifted a little from the plane perpendicular to the line *B*. If the angle of incidence of the wanted signal differs by a small amount ε from zero, the angle of incidence of the 3τ signal will be 3ε. Although this also causes slight attenuation of the wanted signal, the unwanted signal is attenuated to a much greater extent. It can in fact be eliminated completely, though only at one particular frequency (*fig. 8*).

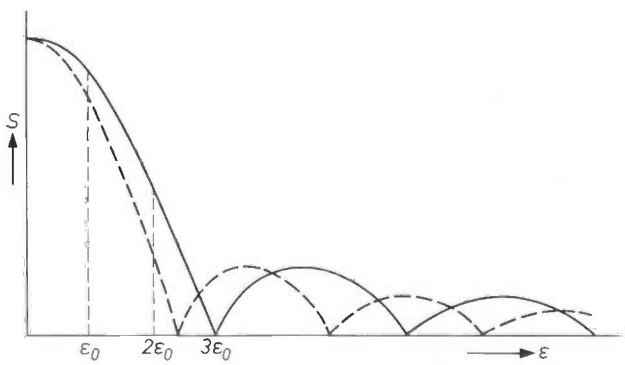


Fig. 8. Magnitude of the signal *S* delivered by a transducer as a function of the angle of incidence ε of the ultrasonic waves. At a particular angle ε_0 the angle of incidence $3\varepsilon_0$ of the 3τ reflection is such that the transducer is insensitive to it at the frequency corresponding to the solid curve. At other frequencies (dashed curve) and the same value of ε_0 this does not apply.

Measurements; grinding to size

Some circuits that can be used for measuring the chief characteristics of a delay line are shown in figs. 9 and 10. Circuits based on the same principles are also used when grinding down the reflecting face *F* in fig. 6 to bring the delay to the exact value required. First of all the reflecting surface is ground away until the group delay is a little greater than the correct value. This is most easily done if a standard delay line is available. An arrangement for comparing its group delay with that of any arbitrary line is shown in the block diagrams of *fig. 9*. The generator *Gen* delivers bursts with the frequency of the subcarrier. (The repetition frequency is fairly low.) These pass through the two delay lines, and are then detected and applied to the vertical deflection circuits of a double-beam oscilloscope. The horizontal deflection is synchronized with the repetition frequency of the bursts from the generator. The difference in group delay is found from the horizontal displacement of the two trains of detected pulses.

When the group delay has reached a value a little greater than the correct value, the next procedure is to measure the *phase shift*. The circuit used for this is shown in *fig. 10*. The generator now delivers a sinusoidal signal with a very constant frequency of 4 433 619 Hz. This signal is conducted along two paths to the amplifier *A*, one via the delay line and the other via a potentiometer *Pot* which causes no significant delay. The amplitudes of the two signals are equalized and the reflection face *F* is ground away until the amplitude of the amplifier input signal is zero. Here again, the measurements can be displayed on an oscilloscope.

The phase shift of the delay line is then an odd multiple of π radians. An additional phase shift of π radians can be obtained if necessary by switching the pairs of input or output terminals, so that the phase is shifted through an even multiple of π radians.

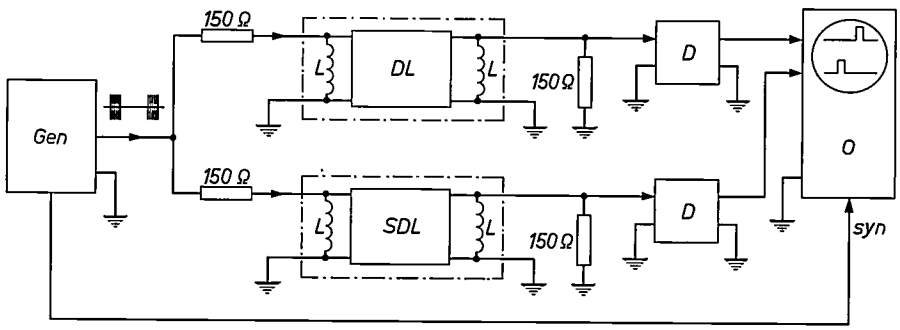


Fig. 9. Diagram for comparing the group delay of any given delay line *DL* with that of a standard delay line *SDL*. *Gen* generator delivering bursts at the frequency of the subcarrier (4 433 619 Hz). *D* detectors. *O* double-beam oscilloscope. *syn* synchronizing signal. *L* coils for compensating the input and output capacitances of the delay lines.

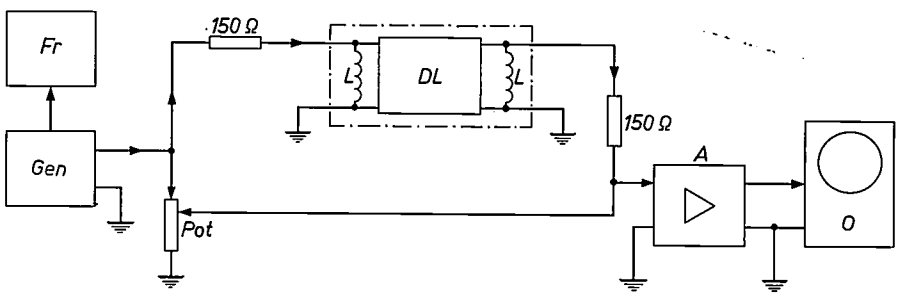


Fig. 10. Circuit that can be used when grinding off a reflecting face in a delay line for setting the correct phase shift. *A* amplifier with very low input impedance. *Pot* potentiometer. *Fr* frequency meter.

[5] A. L. Zijlstra and C. M. van der Burgt, Isopaustic glasses for ultrasonic delay lines in colour television receivers and in digital applications, Ultrasonics 5, 29-38, 1967 (Jan.).

The velocity of propagation of sound in glass is approximately 2.5 mm/μs, and since the wave front is virtually parallel to the reflecting face, about 1.25 μm of glass must be ground away to reduce the delay by 1 ns. The required accuracy of the grinding is therefore of the order of 1 μm. Fortunately the reflecting face does not have to be optically flat.

The *insertion loss* is measured by comparing the output signal of the delay line with the signal from a calibrated attenuator having matched terminations at both ends. The generator used can be a type that delivers a sinusoidal signal, the loss then being measured as a function of frequency. It is also possible, however, as in the group-delay measurement, to use bursts. The advantage of this is that the oscilloscope used as indicating instrument then shows, in addition to the wanted signal, the signals due to unwanted reflections. The 3τ reflection, for example, can then be measured at the same time.

Performance and applications

Fig. 11 shows a complete delay line with the cover removed. Inductances are connected across the input and output terminals to compensate for the capacitance of the transducers (see figs. 9 and 10). The inductances have the same values as the ones that were used when the glass was being ground. There is then no error in the phase delay, which would be the case if different coils were used. (A change of 10% in the inductance causes an error of about 5 ns.) The shape of the loss characteristic can also be changed quite appreciably by a change in the inductance.

As mentioned earlier, the phase shift is set to an integral multiple of π radians by means of grinding. Some other performance figures and tolerances are listed in *Table I*.

Table I. Performance of the delay line (at 25 °C).

Phase-delay tolerance	± 5 ns
Bandwidth (3 dB): lower limit	< 3.43 MHz
Bandwidth (3 dB): upper limit	> 5.23 MHz
Insertion loss (at 4.43 MHz)	10 ± 3 dB
Terminating resistances	150 ohms
Attenuation of 3τ reflections	> 21 dB
Attenuation of other reflections	> 26 dB

Figs. 12 and *13* present some results of measurements on a delay line chosen at random. Fig. 12 gives the insertion loss as a function of frequency, both for the wanted signal (curve *a*) and for the signal due to the 3τ reflection (curve *b*). It can be seen that the 3τ signal has been attenuated by about 30 dB with respect to the wanted signal. Fig. 13 shows, again as a function of frequency, the extent to which the phase shift deviates from the ideal value, established by the phase characteristics in fig. 5. In this particular line the maximum phase error is seen to be ± 4°.

Finally, we give two circuits for a decoder incorporating the delay line. In *fig. 14* the undelayed signal appears across the resistance R_1, whose magnitude is such that this signal is equal to half the output signal from the delay line. If R_1 is connected as shown to the centre of the $2 \times 75\ \Omega$ terminating resistance, the sum signal $B' - Y'$ and the difference signal $R' - Y'$ are obtained at the output terminals p and q respectively. With this arrangement any circuit connected to these terminals should have a fairly high input resistance (at least 1 kΩ).

In the circuit of fig. 14 the output signal has a spread equal to the spread in the insertion loss of the delay

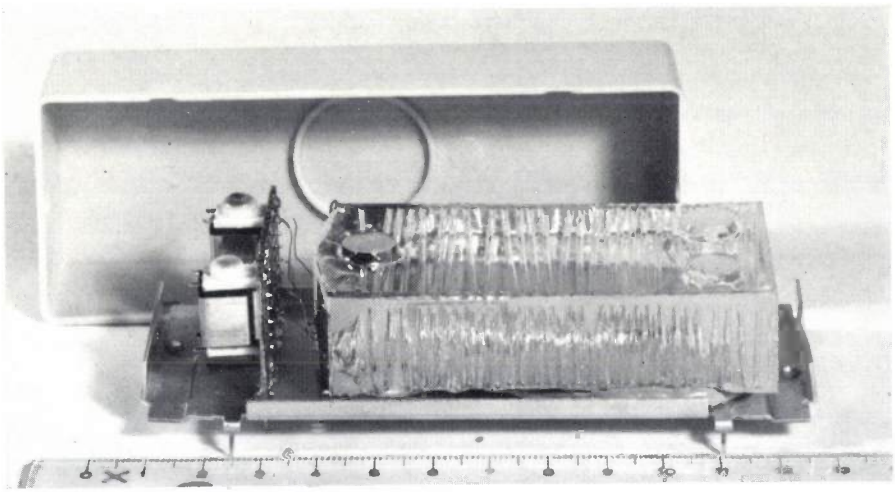


Fig. 11. Delay line with the cover removed.

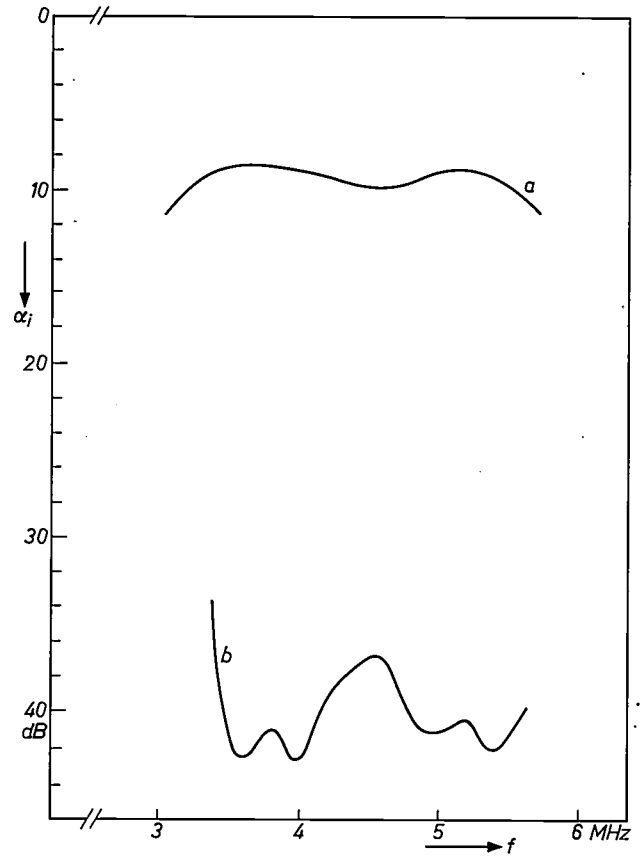


Fig. 12. Insertion loss α_i as a function of frequency f; curve a for the wanted signal, b for the unwanted signal due to the 3τ reflection.

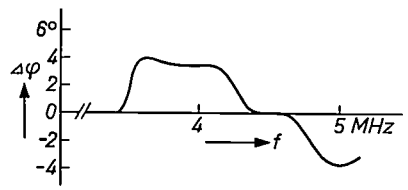


Fig. 13. Phase error $\Delta\varphi$ as a function of frequency.

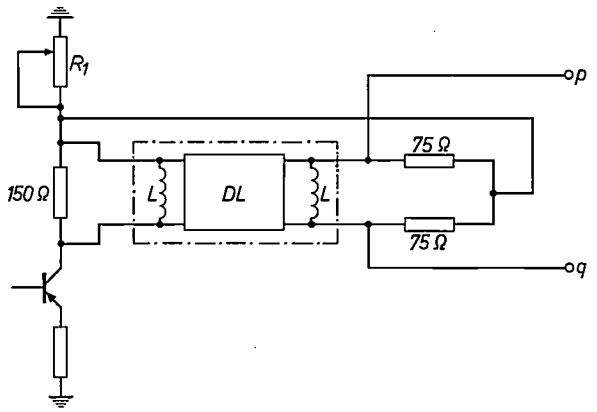


Fig. 14. Circuit of a decoder for a PALd receiver. The transistor amplifies the chrominance signal. The value of the resistor R_1 is such that the signal voltage across it is equal to half the output signal of the delay line DL, so that the $(B' - Y')$ component and the $(R' - Y')$ component of the chrominance signal appear at the output terminals p and q.

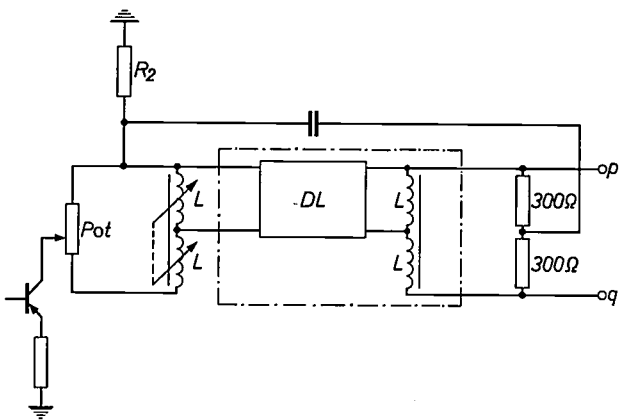


Fig. 15. The magnitude of the output signal is made independent of the spread in the insertion-loss values of the delay lines by means of the potentiometer *Pot*.

lines. This drawback is not found in the arrangement of *fig. 15*. The undelayed signal across resistance R_2 now has a fixed value which is independent of the setting of the potentiometer *Pot* used for regulating the input signal of the delay line. In this way the delayed signal can be made equal to the constant undelayed signal. Since input and output of the delay line are now provided with balanced transformers, the sum and difference signals again appear at the output terminals p and q.

Summary. In the PALd colour television receiver a delay line is used which delays the chrominance signal by one line interval. In the decoding, which is done by simple addition and subtraction of the delayed and undelayed signals, the group delay (not the phase delay) is required to be equal to the line interval (64 μs). The value of the phase delay has to be such that the phase shift for the subcarrier is an integral multiple of π radians. The permissible deviation from the required phase delay is extremely small, approximately 5 nanoseconds. The article describes a delay line which operates through the propagation of ultrasonic shear waves in a block of glass. The glass is of a specially developed type to permit the very strict tolerances to be met in the whole of the desired temperature range. The glass block has a reflecting face, which is ground away until the required phase shift is obtained. The ultrasonic waves in the glass are excited and received by transducers, which consist essentially of thin ceramic piezoelectric plates. Particular care is taken to suppress parasitic signals due to unwanted reflections. Finally, two decoder circuits are discussed which use a delay line of the type described in this article.

Crystal growth of silicon carbide (III)

Two earlier short articles in this journal [1] have discussed our work on the crystal growth of silicon carbide. This work is still in progress, and has recently yielded some interesting new results.

The earlier investigations established the growth of hair-like crystals, or whiskers, which on a microscopic scale reached an appreciable length. We are now able to report that we have succeeded in growing silicon-carbide crystals of very much greater lengths, up to about 10 cm, with diameters from 0.01 to 50 microns. Even greater lengths should be readily attainable. A few of these long whiskers can be seen in the photograph in *fig. 1*, and *fig. 2* shows some typical regular cross-sections of such whiskers; a cross-section of a human hair is reproduced for comparison.

We have also found certain unexpected relations which determine the crystal structure and growth conditions of SiC whiskers. These results will be published in more detail elsewhere, but we shall note the essential points here.

In all our experiments in this field the silicon carbide is grown by means of a chemical reaction between a silicon-containing and a carbon-containing compound, usually at temperatures above 1000 °C. The whiskers form on a substrate made from, say, electrographite. The whisker formation — an exclusive preference for a particular direction of crystal growth — is most easily obtained at relatively low concentrations of the reagents. When the concentrations are increased, particularly at higher temperatures, the whisker formation alters to the growth of "normal" SiC crystals, that is to say, crystals whose outward shapes can be understood in terms of crystal symmetry and the direction of the chemical bonds. Repeated twinning, which occurs very easily in SiC, may nevertheless give rise to crystals of surprising shape [2].

In our previous work we followed the crystal growth from the first stages, at the lowest growth temperature, with the aid of the electron microscope. We found that two basic forms occur: a) whiskers, with a more or less round or hexagonal cross-section, and b) elongated platelets. Electron-microscopic studies showed various peculiar disturbances in the crystal lattice of the whiskers. It was noted, however, that an axial crystal defect was very rare [1].

Later structure determinations were carried out by *electron diffraction*. The results were not without ambiguity: in many cases the diffraction spots do not lie on straight lines, because the whiskers are bent and twisted and therefore the lattice planes deformed. Various parts

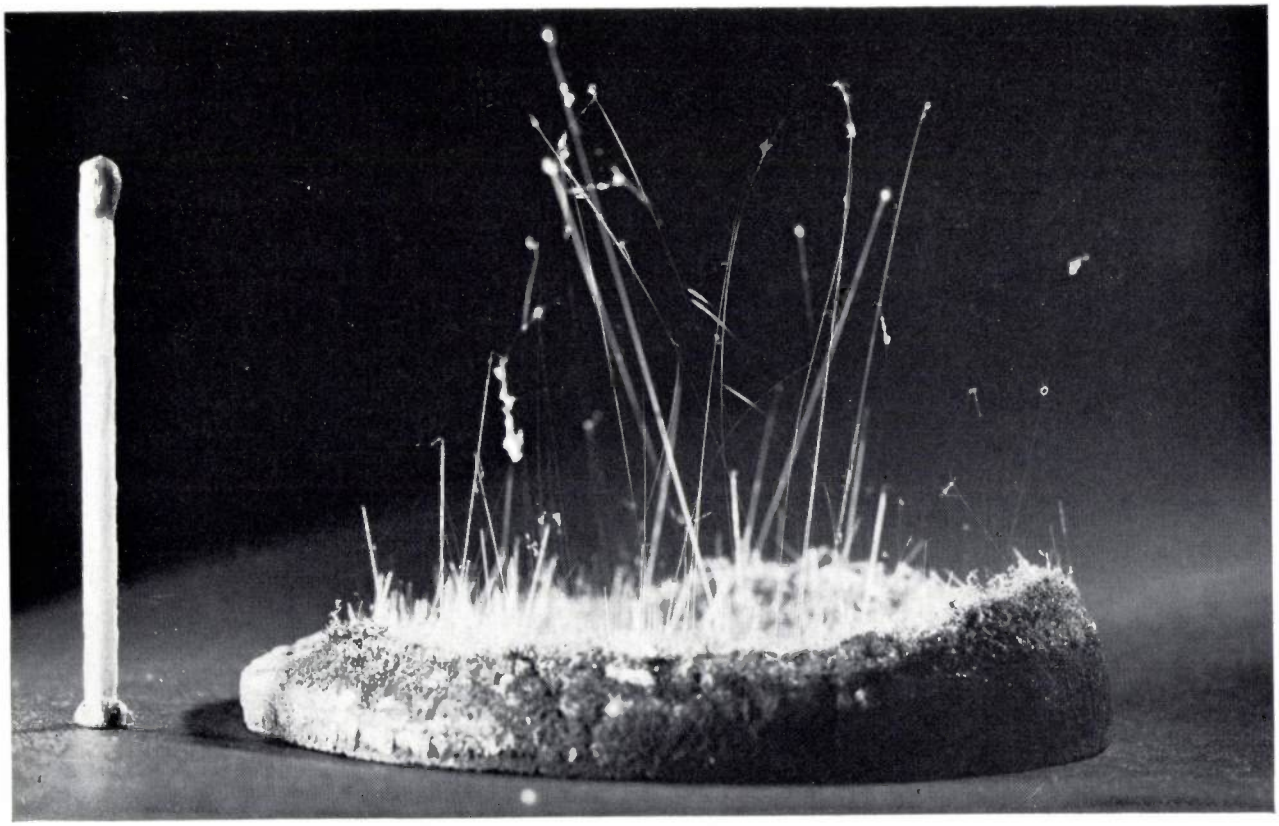


Fig. 1. Whiskers of silicon carbide, up to 6 cm long, with diameters up to 10 μm.

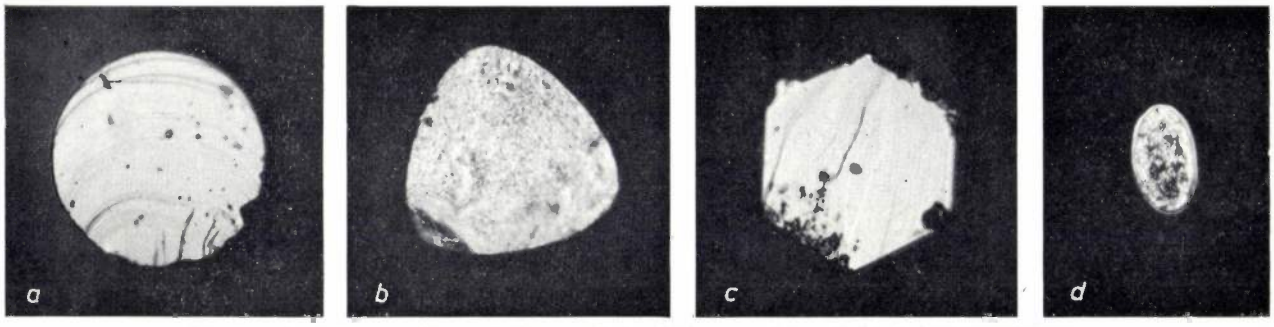


Fig. 2. *a*, *b*, *c*) Cross-sections of silicon-carbide whiskers as shown in fig. 1. For comparison, *d*) shows the cross-section of a human hair (magnification 500×).

of one and the same whisker gave different diffraction patterns, some showing cubic symmetry and others hexagonal symmetry [3].

Optical investigation, by means of polarized light, of the large elongated crystals obtained at higher temperatures had demonstrated that all samples contained *birefringent* regions, or were completely birefringent. It was assumed at the time that there would normally be intergrowth of the birefringent hexagonal wurtzite modification and the non-birefringent cubic sphalerite modification. In our recent work, however, we have found that a sharp distinction must be made here in accordance with the conditions in which the whiskers grow:

1) When whiskers are grown on an arbitrary *polycrystalline* substrate, and there are no impurities in the reaction system, only the sphalerite structure is found (at 1000-1800 °C). Nevertheless, as mentioned, there is birefringence, which gives the impression that the material is not cubic. This birefringence is due to a certain degree of disorder [4]. If impurities are present, the whisker yield is increased and, in a narrow temperature range around 1380 °C, the wurtzite modification is also formed, intergrown with the sphalerite modification. If certain closely-specified impurities are added, it is possible at about 1380 °C to obtain a crystal growth in which all whiskers possess the wurtzite structure.

When the reaction takes place at higher temperature using the same impurities, we find, in addition to the sphalerite modification, considerable quantities of whiskers with the other known hexagonal crystal structures (polytypes) of silicon carbide. At very high temperatures, above 2000 °C, the elongated platelets mentioned earlier ("ribbon-shaped whiskers") can be obtained in lengths up to 5 cm.

The addition of certain other impurities, in the temperature region of 1000-1800 °C, enables the extremely long hair-like whiskers mentioned at the outset to be grown.

2) When the whiskers are grown on large hexagonal *single* crystals of SiC of a given polytype, the whiskers generally obtained are found, surprisingly enough, to have a crystal structure of the same polytype, in other words an extreme kind of epitaxy. It was remarkable in these experiments that the whiskers on the prism and pyramid (side) faces of the lamellar SiC single crystals used were oriented along the *a*-axes of these crystals and nearly all of them acquired the structure of the substrate crystal. Whiskers that grew at right angles to the basal lattice planes (along the *c*-axis of the substrate crystal) formed rows running parallel to the *a*-axes of the substrate crystal. We always found that many of these latter whiskers possessed the sphalerite structure [5]. In this case, too, extremely long whiskers can be obtained as described above, and again these are of the polytype of the substrate.

These findings, and the better control of the growth and properties of SiC whiskers which they offer, bring the feasibility of practical applications much nearer. One application that has long been envisaged is the use of SiC whiskers for the strengthening of materials. It is known that the tensile strength of whisker-like (dislocation-free) crystals may be close to the theoretical tensile strength of a material, which is very much higher than the value found for ordinary monocrystalline or polycrystalline forms of the material. We have measured values higher than 25 000 N/mm² for our SiC whiskers. The whiskers retain a high tensile strength even at very high temperatures.

W. F. Knippenberg
G. Verspui

Dr. W. F. Knippenberg and G. Verspui are with Philips Research Laboratories, Eindhoven.

[1] W. F. Knippenberg, H. B. Haanstra and J. R. M. Dekkers, Crystal growth of silicon carbide (I), Philips tech. Rev. **24**, 181-183, 1962/63.
H. B. Haanstra and W. F. Knippenberg, Crystal growth of silicon carbide (II), Philips tech. Rev. **26**, 187-189, 1965.
[2] W. F. Knippenberg, Philips Res. Repts. **18**, 161-274, 1963, in particular p. 213 *et seq.*
[3] This has also been found by K. A. Jackson and R. S. Wagner, J. appl. Phys. **36**, 2132-2137, 1965.
[4] Disorder in SiC has a very specific significance; see reference [2], in particular p. 193 *et seq.*
[5] See also W. F. Knippenberg and G. Verspui, Philips Res. Repts. **21**, 113-121, 1966.

Flame-failure control with a UV-sensitive cold-cathode tube

I. Design of the tube and circuit

II. Statistical aspects of the detection process and choice of the alarm level

Many heating systems — from the domestic gas water heater and the central heating boiler to large industrial furnaces and steam boilers — require a simple but reliable safety device which monitors the flame to see that it is still alight. For this purpose there are particular advantages in making use of the ultra-violet radiation from the flame: the following two articles deal with this method of detection. The first describes a small gas-discharge tube which responds to UV radiation, There are statistical fluctuations in the output voltage of the tube, and therefore the maximum acceptable risk of false alarm is one of the factors determining the choice of the alarm level for the safety system. Since an exact calculation of the required alarm level is impracticable, a statistical method of approximation has been developed, and this is discussed in the second article.

I. Design of the tube and circuit

T. Poorter

If the flame of a gas-fired furnace or boiler is accidentally extinguished with the gas supply left on, there is soon sufficient unburned gas accumulated to constitute an explosion hazard. Every furnace or boiler therefore requires a safety system which monitors the flames continuously and shuts off the fuel supply and gives an alarm signal if they go out.

In oil-fired equipment, continued supply of the fuel when the flame has gone out has consequences which are perhaps less serious but are still highly undesirable, and here too it is standard practice to provide flame-failure control.

Various methods may be applied to detect the presence of a flame; they differ considerably in their speed of response and the choice between them therefore depends to a large extent on the response time required in the safety system for a particular boiler. An exceptionally short response time can be achieved with a detector that responds to the ultra-violet radiation of the flame. A further advantage of this method is that it provides a simple means of distinguishing between the flame and the hot wall of the combustion chamber. For use in this method Philips have developed a small cold-cathode gas-discharge tube which ignites when it is irradiated with ultra-violet light (wavelength between 200 and 290 nm). Combined with a simple electronic circuit, this tube gives a very reliable flame detector. Before describing the tube we shall first consider in more detail some aspects of flame-failure control in boilers and furnaces [1]. Other safety systems, such as those which respond to the gas pressure and are sometimes used in addition to flame detection, will not be considered here.

Some methods of flame-failure control

For large boilers which can deliver several hundred kW, where the explosion hazard is relatively great and the possible consequences particularly serious, the safety system is required to have a very short response time, no longer than a few seconds. For small boilers a response time of a few tens of seconds is often permitted. The flame detector can then be a simple bimetallic strip or thermocouple, and is usually placed in the pilot flame. The advantage of these detectors is that no electronic amplification for energizing a relay is required; the disadvantages are that they very quickly

Ir. T. Poorter is with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven.

[1] A more general treatment can be found in: B. Maizier, C.R. Congrès Ass. tech. Ind. Gaz **81**, 769-825, 1964.

become sooted up, which upsets the thermal contact with the flame, and the bimetallic strips have only a short life.

In furnaces the situation is similar. For domestic cookers a bimetallic strip or thermocouple is considered sufficient, but for industrial furnaces a safety system with a much shorter response time is required.

A short reaction time is obtained by using either the ionization in the flame or the radiation from the flame. In the first case several electrodes are placed in the flame; when a voltage is applied between these electrodes, a very small current (a few tens of microamperes) flows while the flame is alight. Electronic amplification has to be provided if a relay is to be energized while this ionization current flows. Disadvantages of the method are that the detected current is so small that any leakage currents due to dirt and moisture on the electrodes can cause considerable difficulties, and that the electrodes burn away slowly and therefore need regular replacement. The method is completely unsuitable for oil-fired equipment.

With a device that detects the radiation from the flame the problems of wear and leakage currents due to contamination do not arise; a difficulty here, however, is that the detector, when used in a furnace, "sees" the hot wall of the furnace. *Fig. 1* shows the spectra of the radiation emitted by an oil flame, a gas flame and the wall of a furnace at different temperatures. If the detector is to be able to tell when the flame goes out, it must be sensitive in a spectral region in which the flame is a great deal "brighter" than the furnace wall.

The ability to distinguish between flame and furnace wall can also be based on an entirely different criterion, the "flickering" of the flame. This calls for the use of a fast detector (some infrared detectors are suitable — the spectral region does not matter in this case) whose signal is applied to a selective a.c. amplifier, tuned to about 15 Hz. This method, however, is rather complicated and costly.

At wall temperatures below about 1200 °C the oil flame is brighter than the wall in the visible range (above 400 nm), and in this case cadmium-sulphide cells, which are sensitive to visible light, can therefore be used. The brightness of the gas flame in the visible range is however too low to permit safe operation with a cadmium-sulphide cell. In the ultra-violet range, on the other hand, between say 200 and 300 nm, both the oil and the gas flame are much "brighter" than even the hottest furnace wall. Ultra-violet detection thus offers a universal means of flame-failure control. Moreover, as explained below, a UV detector can be given such a high intrinsic "gain" that its output signal may if required be used directly for actuating a relay, which simplifies the system considerably. What is more, a UV detector can monitor the ignition spark as well as the flame. Since large furnaces or boilers are always automatically ignited by means of an electric spark, this indicates the possibility of monitoring and safeguarding the entire ignition programme (a spark ignites a pilot flame which ignites the main burner).

The radiation emitted by a flame becomes "redder", i.e. the spectrum shifts to longer wavelengths, as the distance from the base of the flame increases — the ultra-violet radiation is found mostly in the base of the flame. The UV detector must therefore preferably be situated in such a way that the radiation from the base of the flame is incident upon it. The best place for the detector is therefore in the air-supply opening; this has the additional advantage that the detector is cooled by the jet of air.

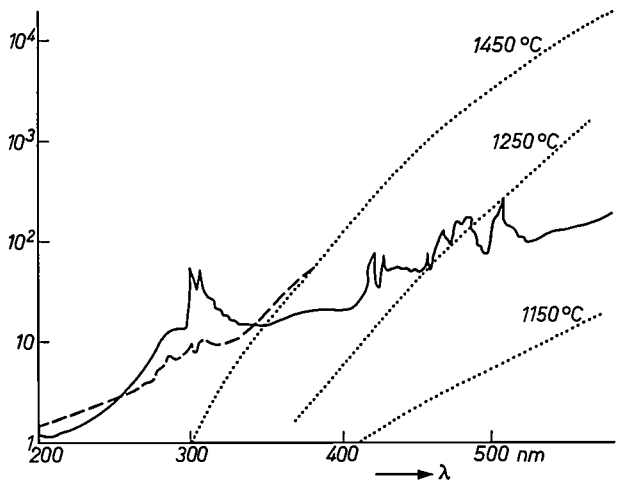


Fig. 1. Spectral intensity distribution (in arbitrary units) of the radiation emitted by a gas flame (solid line), an oil flame (dashed line) and the furnace wall (black-body radiator) at different temperatures (chain-dotted lines).

A cold-cathode tube as UV detector

If a sufficiently high voltage is applied between the cathode and anode of a gas-discharge tube (higher than the "ignition voltage"), a self-sustaining discharge occurs in the tube as soon as there are sufficient numbers of free electrons between the electrodes. These electrons may be released from the cathode by thermionic emission or by incident radiation (photoelectric effect). In the latter case, which applies for the cold-cathode tube, the type of cathode material used determines the wavelength range of the radiation to which the tube responds. The current of the self-sustaining discharge is so very much greater than that of the originally generated electrons that the tube has an extremely high "internal gain". These are the principles underlying the UV detector described below.

The tube, type 155 UG (*fig. 2*), contains two identical molybdenum electrodes. The work function of this metal (i.e. the minimum energy an electron must acquire in order to leave the solid) is 4.26 eV, corresponding

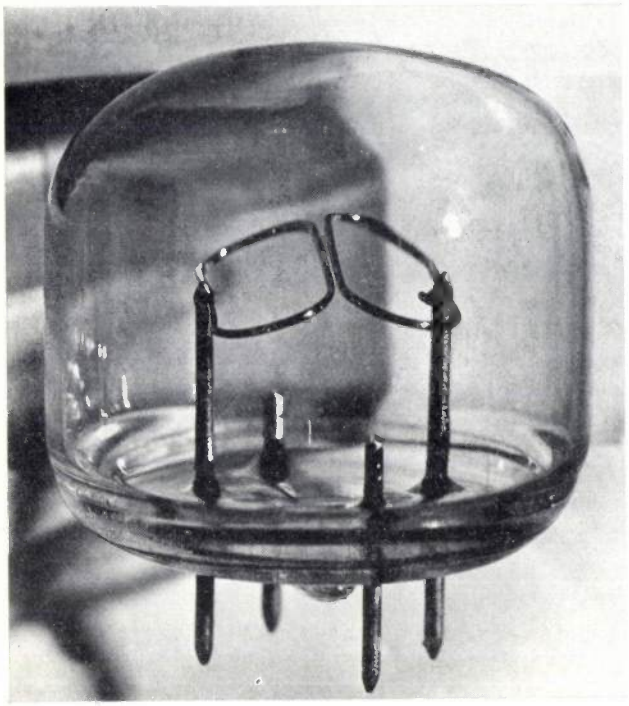


Fig. 2. The cold-cathode tube type 155 UG. If the voltage between the electrodes is sufficiently high, a self-sustaining gas discharge is produced when UV radiation is incident on the electrodes. Since molybdenum has been used as electrode material, the tube responds only to UV.

to a photoelectric threshold of about 290 nm; only photons with a wavelength below this threshold are effective. Obviously, the bulb of the tube must be made of a material that transmits UV at these wavelengths; the glass used for the bulb is a type that transmits radiation at wavelengths greater than 200 nm. The tube thus responds to radiation between 200 and 290 nm. *Fig. 3* shows the spectral distribution of the photoelectric quantum efficiency (i.e. the ratio of the number of electrons released to the total number of incident photons), the transmission characteristic of the glass, and the resultant spectral sensitivity for detection.

The tube was designed for operation at 220 V a.c. and a frequency of 50 or 60 Hz, so that it can be supplied directly from the mains. The gas filling is a mixture of hydrogen and helium, giving an ignition voltage of 200 V, i.e. about 60% of the peak voltage of the a.c. supply. The tube is operated in the glow-discharge region, in which the maintaining voltage for this tube is also about 200 V. An incident UV photon can thus initiate a gas discharge when the instantaneous value of the supply voltage is greater than 200 V; the discharge goes out immediately the voltage drops below the maintaining voltage. In order to prevent spontaneous re-ignition due to residual ionization from a preceding cycle, the de-ionization time of the gas mixture should be very short; the gas mixture which we have used meets this requirement.

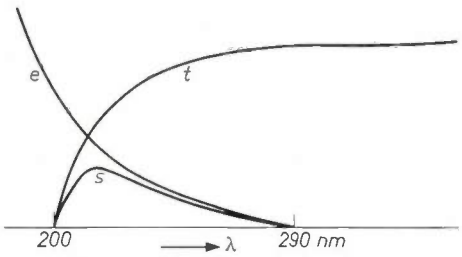


Fig. 3. General spectral distribution of the sensitivity s of the detector, derived from the quantum efficiency $e(\lambda)$ of the photoelectric effect at the electrode and the transmission $t(\lambda)$ of the glass bulb.

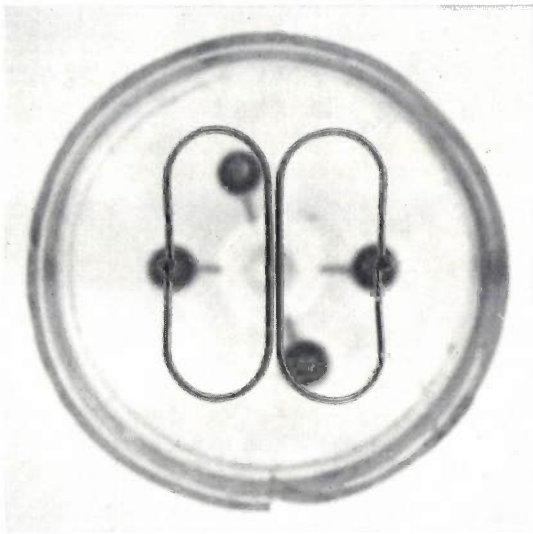


Fig. 4. Cut-away view of the 155 UG, showing the electrodes. The effective part of the electrodes consists of closely-spaced straight pieces of wire at a short distance apart. The electrodes are identical and, since the supply is a.c. they serve alternately as cathode and anode.

Fig. 5. Transverse cross-section through the effective part of the electrodes. The electrical field strength is greatest between the wires, and smallest at *A*. The probability that a UV photon will initiate a discharge therefore depends on the place where it strikes the electrode. The probability is greater at a higher voltage.

Fig. 4 illustrates the electrode configuration with which this value of the maintaining voltage is obtained. As an a.c. supply is used, the two electrodes are identical, each acting alternately as cathode and anode. The parallel parts of the wires constitute the active region of the electrodes. These wires are roughly circular in cross-section, so that the electric field strength is not the same at every point of the surface. The field strength is of course greatest *between* the wires (*fig. 5*). This

means that electrons released at different places on the electrode surface have a different probability of initiating a self-sustaining discharge. Since, when the voltage between the electrodes is increased, the area of the electrode surface for which a released electron can initiate a self-sustaining discharge becomes greater, the sensitivity of the tube (the number of ignitions per unit time for a given incident radiation) will increase with rising voltage. The 155 UG does not therefore have the plateau so characteristic of the Geiger-Müller tube, which it closely resembles in operation. For this reason the 155 UG is not so suitable for *measuring* the intensity of the radiation; this tube is primarily intended to be a very sensitive device for detecting whether or not radiation is present.

The voltage between the electrodes is subject to an upper limit. The maximum permissible value is determined by the "free-running" voltage, that is to say the voltage at which the tube no longer goes out when the irradiation is interrupted but spontaneously re-ignites in each successive cycle of the supply voltage. The cause of this effect is a declining weak ionization which persists, in spite of the short de-ionization time of the gas mixture, after the discharge has gone out. In the 155 UG the free-running voltage is considerably higher than the nominal supply voltage, so that in practice there is little danger of this effect being encountered. During the life of the tube, however, the value of the free-running voltage may decrease if the electrode surface becomes contaminated, e.g. by the growth of molybdenum-oxide crystals. Locally higher field strengths occur at the tips of such crystals, increasing the chance of re-ignition. The occurrence of such contaminations can be prevented, however, by applying a suitably shaped current waveform during the discharge. We shall now deal with this point at somewhat greater length.

Effect of current waveform on the reliability of the tube

Contamination or "poisoning" of the cathode surface is harmful not only because it lowers the free-running voltage: if the contaminants are materials of relatively low work function, like molybdenum hydride or alkali metals from the glass, the tube might begin to respond to visible light, which is the very thing that has to be avoided.

Fortunately, it so happens that the discharge itself removes the contaminants. Extensive life tests have shown that the electrode surface remains completely clean if at each discharge the glow is made to cover *the entire active area* of the electrodes by sending a short pulse of current which is sufficiently large through the tube. Since an a.c. supply is used, both electrodes are continuously decontaminated in this way. To assist our explanation of this mechanism we shall briefly consider some of the effects encountered in a cold-cathode gas-discharge tube.

During a self-sustaining discharge in such a tube the cathode surface is continuously bombarded by ions which are accelerated by the electric field in the cathode fall. The electrons which maintain the discharge are released from the cathode by the bombardment, which also releases particles of the electrode material from the surface (sputtering). These sputtered particles spread through the tube by diffusion, perhaps forming compounds with molecules in the gas, e.g. oxygen molecules, and finally settle in one form or another on the surface of the electrodes or on the glass wall. The ion bombardment thus has a decontaminating effect on the part of the cathode which is covered by the discharge (and also on the gas atmosphere), and a contaminating effect on every part that is not covered. It is evident that the contamination will be greatest on the part of the cathode surface which lies just outside the limits of the discharge.

This effect occurs in the tube discussed here, as it does in all other tubes designed for this kind of application. One of the forms which the contamination takes here is a spontaneous crystal growth of molybdenum oxide at the boundary of the discharge (*fig. 6*). The

Fig. 6. Molybdenum-oxide crystals, which can form on the electrodes of the 155 UG if the peak current through the tube is too low. Magnification 30x.

oxygen needed for this reaction comes from the glass wall. This release of oxygen from the wall is also one of the reasons for the gradual decline in the sensitivity of the tube, since the oxygen absorbs UV photons produced in the discharge and thus hinders the transition from non-self-sustaining to self-sustaining discharge, in which these photons play an important role.

Although the decline in sensitivity does not reduce the margin of safety — if the sensitivity is inadequate the tube gives the alarm while the flame is still burning (it is "fail-safe") —, this deterioration should obviously be avoided as far as possible for the sake of a reasonable tube life.

The way to stop cathode contamination by crystal growth has already been indicated: in each discharge there should be a pulse of current which is large enough to make the discharge cover the entire active surface of the cathode. During the remaining part of the half-cycle the current may not drop to such a value that the discharge contracts to a small part of the cathode surface: in this part of the cycle the current must be either zero or so high that the whole cathode remains covered. Both alternatives are possible, and the circuits which can be used will be discussed below. In the first case the mean current is very low and there is thus very little dissipation; in the second case the dissipation is relatively high.

It is interesting that this difference is reflected in all kinds of ways in the sensitivity behaviour of the tube. With *low* dissipation the glass wall remains relatively cold and gives off little oxygen. Nevertheless, oxygen pressure is gradually built up. The glass wall is also gradually darkened by the UV radiation from the discharge. The overall result is a slight decline in sensitivity, about to half the original value in 10 000 hours of operation. With *high* dissipation the wall gets much hotter and the emission of oxygen is greater. However, the oxygen pressure remains lower than it does for low dissipation, since the higher current causes greater cathode sputtering and therefore better decontamination of the gas atmosphere. The higher mean current also gives rise to greater UV irradiation of the wall, but here again there is ample compensation; since the wall temperature is higher, the glass does not darken so quickly. For one reason and another the decline in sensitivity is even smaller for high dissipation than it is for low dissipation.

What are the consequences for the practical application of the tube? Since the sensitivity of the tube gradually decreases, while the free-running voltage does *not* decrease when the tube is operated with the kind of current waveform we have described (in practice it may even rise for various reasons), the tube gives a reliable flame-failure control. The behaviour of the tube outlined above has been established from life tests of up to 25 000 hours and a total of one and a half million hours of tube life. The end of the useful life of a tube in a particular application is determined entirely by the minimum sensitivity required for that application; this depends on factors such as the distance from the flame, the size of the flame and the minimum level required to actuate the electronic circuit in which the tube is incorporated.

The circuit

The circuit in which the tube is incorporated must be designed to give the current through the tube the required waveform and produce an output signal that indicates whether or not the flame is alight. We have seen that two types of circuit are possible: one that gives a very small dissipation in the tube and one that gives a high dissipation.

In the first circuit a short high-current pulse must flow in the tube at each discharge, and the current must then drop right down to zero. This is done by making the discharge *self-quenching*. The circuit is shown in *fig. 7*. V_s is the mains voltage with an r.m.s. voltage of 220 V, the load resistance R_2 has a high value and R_1 has a very low value. We start from the situation in which the discharge is out; the full voltage then appears across the tube. When V_s increases to a value

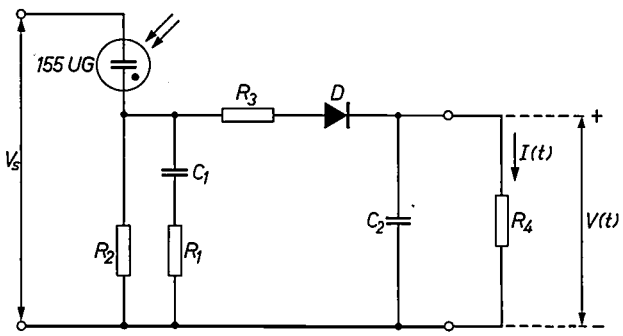


Fig. 7. Circuit for the cold-cathode tube in which the discharge is self-quenching: at each discharge a short pulse of current passes through the tube, but the average value of this current — i.e. the dissipation in the tube — is low. The supply voltage to the circuit is 220 V from the a.c. mains; the output voltage $V(t)$ is a measure of the intensity of the UV radiation incident on the tube.

greater than the ignition voltage V_i, an incident UV photon can initiate the discharge. Since R_1 is low, a high current first flows through the tube via C_1 and R_1. C_1 becomes charged, the voltage across R_2 rises and the voltage across the tube falls. The component values are so chosen that the tube voltage then falls to below the steady-state maintaining voltage V_m which corresponds to the load resistance R_2. Now when the current through C_1 and R_1 stops flowing because C_1 is charged up, the tube can no longer continue to operate and cuts out. The capacitor C_1 now discharges through R_1 and R_2, which causes the voltage across the tube to rise again. Spontaneous re-ignition is not possible because the time constant $(R_2 + R_1)C_1$ has been made so large that the voltage across the tube rises more slowly than the re-ignition voltage (which rises again rapidly because of the de-ionization). The discharge thus remains out and can only be re-ignited again when the voltage across the tube has risen again sufficiently and a new UV photon strikes the (acting) cathode. The required value of the time constant has been found to be small enough for several discharges to take place

in this way during a half-cycle of the supply voltage. Each discharge produces a voltage pulse across R_2.

The use of an a.c. supply voltage means that both positive and negative pulses appear across R_2 (in the positive and negative half-cycles) with respect to the reference potential in fig. 7. Only the positive pulses are passed via R_3 and the diode D to the capacitor C_2; this charges up and discharges again between two pulses through the load resistance R_4, giving rise to a fluctuating voltage $V(t)$ across R_4; see *fig. 8*. This output voltage $V(t)$ is a measure of the number of discharges that have taken place during a number of cycles (or, to be more exact, during their "positive" halves) before time t, and is thus a measure of the intensity of the UV radiation. The average magnitude of $V(t)$ at a particular radiation intensity is defined as the sensitivity of the tube in the circuit. The incident radiation is adjusted such that $V(t)$ is normally 10 to 15 V; if there are no discharges, then $V(t)$ is of course zero. The rate at which C_2 discharges between two pulses is determined, for a given R_4, by the capacitance of C_2; if the capacitance has a high value (i.e. the circuit has a high time constant), $V(t)$ decreases slowly, so that the average value is relatively high. A *smaller* capacitance gives a faster decrease in $V(t)$, and therefore a greater speed of response and a lower output voltage, which in turn gives a lower sensitivity.

The fluctuations of $V(t)$ have a statistical character. In the first place, the incidence of the UV photons at the electrodes is a statistical effect and in the second place not every photon incident at the negative electrode causes a discharge, the probability of this depends on the location. As already explained, this probability increases with increasing supply voltage; mains voltage fluctuations therefore affect the output voltage as well. When the circuit is used in a flame-failure control system it is necessary to take these different fluctuations into account; this point will be dealt with below.

In the circuit shown in fig. 7 the available current I in the load resistor R_4 is of the order of some tens of microamperes, so that the current has to be amplified to energize a relay in a flame-failure control system.

In the second circuit, where a high dissipation is tolerated in the tube, there is no need to amplify the output signal; a relay winding is incorporated directly in the circuit. The circuit is given in *fig. 9*; the discharge here is *not* self-quenching. V_s is the mains voltage, R_1 has a high value and R_2 a low value. The relay *Rel* is shunted across the capacitor C. The current surge begins here in the same way as in the circuit in fig. 7, but the discharge goes out only when V_s falls below the maintaining voltage. The average current flowing through the tube in this circuit is considerably higher than in fig. 7; the peak value can be as high as several hundred milliamps. The tube has been found capable of withstanding this high dissipation without suffering any damage.

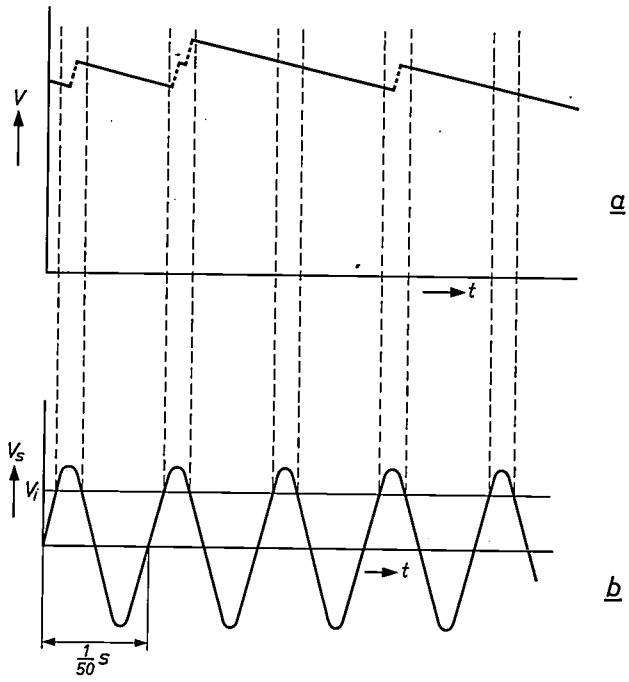


Fig. 8. Illustrating the variation of the output voltage V of the circuit in fig. 7 as a function of time (*a*). The waveform of the supply voltage $V_s(t)$ is shown in (*b*). A UV photon incident on the tube can initiate a discharge in those parts of the cycle in which V_s is greater than the ignition voltage V_i of the tube. On account of the statistical character of this process there is a fluctuation in $V(t)$.

The operation of this circuit differs from that of the circuit described earlier in that there are fewer ignitions per unit time, and because of this the tube does not discriminate so sharply between flame and background radiation. These details cannot be dealt with here. The remainder of this article will be based on the circuit of fig. 7, and the same applies for the following article.

The choice of alarm level

When the UV detection circuit in fig. 7 is used in a flame-failure control system, the output voltage $V(t)$ is

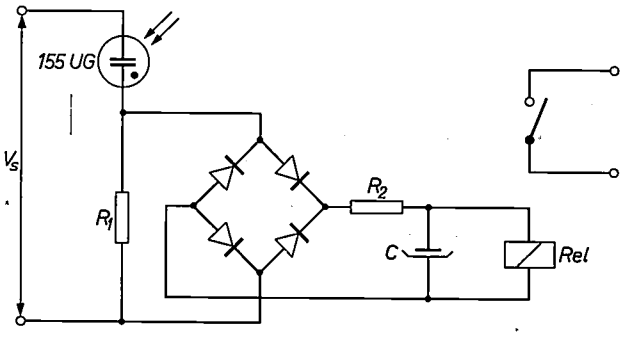


Fig. 9. Circuit in which the mean tube current is sufficiently high to energize directly a relay *Rel*. The 155 UG tube can withstand this high dissipation without damage.

applied to a circuit which closes the gas valve when the voltage $V(t)$ drops below a critical value, the *alarm level*. The gas valve is opened again when $V(t)$ rises above this level. There are some advantages in using different levels for closing and opening the gas valve, but here we shall confine ourselves to the simple case of a single level.

Even if the flames are not alight, discharges can still be caused by cosmic radiation, or by ultra-violet radiation from sunlight or a fluorescent lamp shining in through a furnace window. As a result, even when the flames are out, the voltage $V(t)$ can rise to a small positive value. The alarm level must therefore lie well above this value — the higher the better for safety. It may also happen that, while the flame is alight, there are no discharges at all during one or two positive half-cycles of the supply voltage, so that the voltage $V(t)$ drops below the alarm level and the burner is unnecessarily shut off. The chance of such a *false alarm* is of course greater the higher the alarm level has been chosen.

On the other hand, a higher alarm level implies a faster response. The correct alarm level therefore represents a compromise between the speed of response and the chance of a false alarm. The main parameters which determine the circumstances under which the compromise is chosen are the time constant of the detector circuit and the radiation level at the detector. By varying these parameters, for example in positioning the detector in relation to the source, one has to ensure that a compromise can be found which meets the minimum requirements specified by the user in both respects (speed of response and avoidance of false alarm).

Since by its nature, the output signal of the UV detector is subject to fluctuations, the relation between alarm level and the probability of a false alarm can only be investigated by means of a statistical analysis. This is the subject of the second article.

Summary. The flame-failure control system for a furnace or boiler must shut off the fuel feed the moment the flames are extinguished for one reason or another. There are various methods of detecting the presence of a flame. After an introduction to this subject, a small cold-cathode tube is discussed which is ignited by the UV radiation from a flame provided the voltage across the tube is greater than the ignition voltage. The tube operates on a 220 V a.c. supply; there are two identical molybdenum electrodes, alternately acting as cathode and anode. If steps are taken to ensure that the glow covers the entire cathode during each discharge, there is no contamination of the electrodes, which could reduce the reliability of the tube. The effects which determine the life of the tube are dealt with at some length. Two circuits are discussed which give the current through the tube the required form and deliver an output current which is a measure of the intensity of the incident UV radiation. There are statistical fluctuations in the output signal, and a result of this is that a certain chance of false alarm has to be accepted when establishing the alarm level, i.e. the level at which the fuel supply is shut off.

II. Statistical aspects of the detection process and choice of the alarm level

R. P. Adriaanse and P. van der Laan

The first of these two articles described a UV detector for use in a flame-failure control system; there are statistical fluctuations in the output voltage of the detector. In this second article the output signal will be mathematically analysed and described with the aid of a statistical model. A calculation will then be given which shows how the alarm level can be chosen in such a way that a false alarm will occur on an average no more than twice in a year of continuous operation. The problem is of a complicated nature since the statistical model describes a "Markov" process with *dependent increments*. One of the implications of this is that the fluctuations in the output voltage can only be characterized by probability distributions.

The analysis has been carried out for the circuit in fig. 7; a typical curve of the output voltage $V(t)$ from this circuit when the flame is alight is shown in fig. 8. A false alarm occurs if $V(t)$ drops below the alarm level while the flame is alight: we are therefore interested in the lowest values that $V(t)$ can reach. The various components of the circuit of course each have a spread about a nominal value. These deviations have an effect on $V(t)$; in our calculations we must always use the particular combination of values that yields the *lowest* value of $V(t)$ ("worst-case" combination). Since the calculations will be illustrated with numerical examples, *Table I* gives the nominal value of various components and the worst-case value for a particular version of the

Ir. R. P. Adriaanse, is with the Technical University of Delft, and Drs. P. van der Laan is with Philips Information Systems and Automation (Research Dept.), Eindhoven; both were formerly with Philips Research Laboratories, Eindhoven.

applied to a circuit which closes the gas valve when the voltage $V(t)$ drops below a critical value, the *alarm level*. The gas valve is opened again when $V(t)$ rises above this level. There are some advantages in using different levels for closing and opening the gas valve, but here we shall confine ourselves to the simple case of a single level.

Even if the flames are not alight, discharges can still be caused by cosmic radiation, or by ultra-violet radiation from sunlight or a fluorescent lamp shining in through a furnace window. As a result, even when the flames are out, the voltage $V(t)$ can rise to a small positive value. The alarm level must therefore lie well above this value — the higher the better for safety. It may also happen that, while the flame is alight, there are no discharges at all during one or two positive half-cycles of the supply voltage, so that the voltage $V(t)$ drops below the alarm level and the burner is unnecessarily shut off. The chance of such a *false alarm* is of course greater the higher the alarm level has been chosen.

On the other hand, a higher alarm level implies a faster response. The correct alarm level therefore represents a compromise between the speed of response and the chance of a false alarm. The main parameters which determine the circumstances under which the compromise is chosen are the time constant of the detector circuit and the radiation level at the detector. By varying these parameters, for example in positioning the detector in relation to the source, one has to ensure that a compromise can be found which meets the minimum requirements specified by the user in both respects (speed of response and avoidance of false alarm).

Since by its nature, the output signal of the UV detector is subject to fluctuations, the relation between alarm level and the probability of a false alarm can only be investigated by means of a statistical analysis. This is the subject of the second article.

Summary. The flame-failure control system for a furnace or boiler must shut off the fuel feed the moment the flames are extinguished for one reason or another. There are various methods of detecting the presence of a flame. After an introduction to this subject, a small cold-cathode tube is discussed which is ignited by the UV radiation from a flame provided the voltage across the tube is greater than the ignition voltage. The tube operates on a 220 V a.c. supply; there are two identical molybdenum electrodes, alternately acting as cathode and anode. If steps are taken to ensure that the glow covers the entire cathode during each discharge, there is no contamination of the electrodes, which could reduce the reliability of the tube. The effects which determine the life of the tube are dealt with at some length. Two circuits are discussed which give the current through the tube the required form and deliver an output current which is a measure of the intensity of the incident UV radiation. There are statistical fluctuations in the output signal, and a result of this is that a certain chance of false alarm has to be accepted when establishing the alarm level, i.e. the level at which the fuel supply is shut off.

II. Statistical aspects of the detection process and choice of the alarm level

R. P. Adriaanse and P. van der Laan

The first of these two articles described a UV detector for use in a flame-failure control system; there are statistical fluctuations in the output voltage of the detector. In this second article the output signal will be mathematically analysed and described with the aid of a statistical model. A calculation will then be given which shows how the alarm level can be chosen in such a way that a false alarm will occur on an average no more than twice in a year of continuous operation. The problem is of a complicated nature since the statistical model describes a "Markov" process with *dependent increments*. One of the implications of this is that the fluctuations in the output voltage can only be characterized by probability distributions.

The analysis has been carried out for the circuit in fig. 7; a typical curve of the output voltage $V(t)$ from this circuit when the flame is alight is shown in fig. 8. A false alarm occurs if $V(t)$ drops below the alarm level while the flame is alight: we are therefore interested in the lowest values that $V(t)$ can reach. The various components of the circuit of course each have a spread about a nominal value. These deviations have an effect on $V(t)$; in our calculations we must always use the particular combination of values that yields the *lowest* value of $V(t)$ ("worst-case" combination). Since the calculations will be illustrated with numerical examples, *Table I* gives the nominal value of various components and the worst-case value for a particular version of the

Ir. R. P. Adriaanse, is with the Technical University of Delft, and Drs. P. van der Laan is with Philips Information Systems and Automation (Research Dept.), Eindhoven; both were formerly with Philips Research Laboratories, Eindhoven.

Table I. Nominal values and worst-case values for the components of the circuit in fig. 7.

	Nominal	Worst case
R_1	100 Ω	90 Ω
R_2	330 kΩ	300 kΩ
R_3	150 kΩ	160 kΩ
R_4	470 kΩ	440 kΩ
C_1	12 nF	10 nF

circuit of fig. 7. Four different values for the capacitor C_2 are considered:

$$C_2 = 1;\ 2.2;\ 4;\ \text{and}\ 6\ \mu\text{F}.$$

As already explained in part I, the speed of response is greater for a lower value of C_2 but the sensitivity is lower.

The supply voltage for the detector is the normal mains voltage (r.m.s. value 220 V), but since this voltage fluctuates and can incidentally drop to 200 V, we take the supply voltage to be $V_s(t) = \hat{V}_s \cos 2\pi t/T$, with $T = 1/50$ s, and the worst-case value:

$$V_{s\,\text{eff}} = 200\ \text{V, so that}\ \hat{V}_s = 283\ \text{V}.$$

The ignition voltage V_i and the maintaining voltage V_m are taken equal to the highest values which can occur with a tube of the type used. For the 155 UG cold-cathode tube considered here, these values are $V_i = V_m = 220$ V.

The average value of the output voltage $V(t)$ measured under these worst-case conditions (with $C_2 = 2.2$ μF) was 5 to 15 V, depending on the intensity of the radiation, and the lowest instantaneous value was between 1 and 10 V. Since a small probability of false alarm is allowed when the alarm level is established, we may expect that the alarm level will have a value of the same order of magnitude.

The variation of the output voltage

In describing the fluctuations in the output voltage $V(t)$ we adopt the following assumptions:

1) There is never more than one discharge in a "positive" half-cycle of the supply voltage.
2) When the tube ignites during the positive half-cycle it always does so at the peak value $\hat{V}_s$ of the supply voltage.

One or two comments should be made about these assumptions. The first can be considered as a worst-case assumption, since several ignitions will often occur in a positive half-cycle, resulting in a higher value of $V(t)$. The second assumption does not lead to the lowest possible value of $V(t)$, but some simplification was essential in order to make the process susceptible to analysis. In view of the ignition pattern of the tube, however, it seems reasonable to suppose that the effects of the two assumptions will roughly balance out.

The supply voltage $V_s(t)$ reaches the peak value $\hat{V}_s$ at the times:

$$t = 0,\ T,\ 2T,\ \ldots \quad (T = 1/50\ \text{s}).$$

From assumption (2) these are also the times at which a discharge may occur. We define further:

$$x_n = V(nT),\ \text{where}\ n = 0, 1, 2, \ldots, \qquad (1)$$

so that x_n is the output voltage at the time $t = nT$. There are now two possibilities for the relation between x_n and x_{n+1}. If a discharge occurs at the time $t = nT$, the output voltage rises and we may write as an approximation:

$$x_{n+1} = \lambda_1 x_n + c, \quad (0 < \lambda_1 < 1, \quad c > 0). \qquad (2)$$

If no discharge occurs at the time $t = nT$, the output voltage decreases and x_{n+1} is given by:

$$x_{n+1} = \lambda_2 x_n \quad (0 < \lambda_2 < 1). \qquad (3)$$

The quantities λ_1, λ_2 and c may be shown to be related to the circuit constants as follows:

$$\left.\begin{aligned} \lambda_1 &= \left(1 - \frac{1}{\dfrac{R_3 R_4}{R_3 + R_4} C_2 \beta_2}\right) e^{-T/R_4 C_2}, \\ c &= \frac{\hat{E}}{R_3 C_2 \beta_2} e^{-T/R_4 C_2}, \\ \lambda_2 &= e^{-T/R_4 C_2}. \end{aligned}\right\} \qquad (4)$$

Here $\hat{E}$ is equal to the peak value of the voltage across C_1 which is reached at the moment of the ignition, so that $\hat{E} = \hat{V}_s - V_m = 63$ V. *Fig. 10* illustrates the kind of curve for $V(t)$ that would be obtained with this model.

The statements made above can be derived in the following way. If a discharge occurs at the time $t = 0$, then provided diode D in fig. 7 is conducting (i.e. only in the "positive" half-

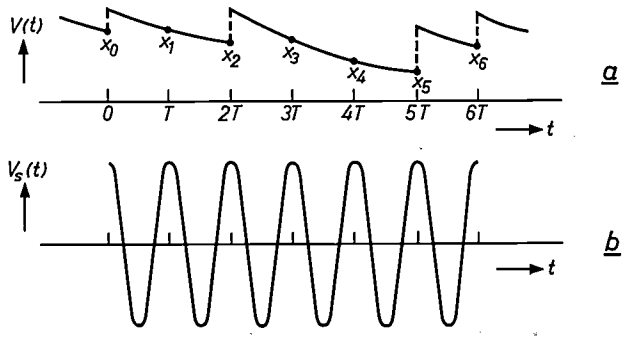


Fig. 10. *a*) Typical curve for the output voltage $V(t)$: it is assumed that no more than one discharge can take place in each positive half-cycle of the supply voltage $V_s(t)$ shown in (*b*), and that the discharge always takes place at the peak value of the supply voltage, i.e. at the moment $t = nT$. The values of $V(t)$ at these times are x_0, x_1, x_2, etc.

cycles of the supply voltage) the supply voltage $V(t)$ satisfies the following differential equation:

$$\ddot{V} + a\dot{V} + bV = 0,$$

with $$a = \frac{1}{R_2C_1} + \frac{1}{R_4C_2} + \frac{1}{R_3C_2} + \frac{1}{R_3C_1}$$

and $$b = \frac{R_2 + R_3 + R_4}{R_2R_3R_4C_2C_1}.$$

It is assumed here that R_1 is negligible as compared with R_2 and R_3. The general solution of this differential equation is:

$$\left.\begin{aligned} V(t) &= \alpha_1 e^{-\beta_1 t} + \alpha_2 e^{-\beta_2 t}, \\ \text{where} \quad \beta_{1,2} &= \tfrac{1}{2}a \mp \sqrt{\tfrac{1}{4}a^2 - b}. \end{aligned}\right\} \quad \ldots\ldots \quad (5)$$

The constants α_1 and α_2 can be determined from the initial values of V and $\dot{V}$. These initial values are determined by what has happened beforehand, that is to say by the number of discharges that have already taken place and the time between these discharges.

If no discharge occurs at the time $t = nT$, the output voltage falls at a rate which is determined by the time constant R_4C_2; we then have:

$$x_{n+1} = x_n e^{-T/R_4C_2}.$$

This is the equation which we have called (3) and the expression for λ_2 can be read off directly.

If a discharge does occur at the time $t = nT$, a voltage increase is obtained which can be calculated from (5), and then there is once again a voltage decrease, given by R_4C_2. It can be shown that the relation given by (2) between x_{n+1} and x_n now applies to a good approximation, giving the expressions in (4) for λ_1 and c.

Now we come to the problem of the irregular succession of increases and decreases in $V(t)$. We shall assume that, when the flame is alight, the probability of an ignition is equally great for all times $t = nT$, and we shall call this probability p. The magnitude of p depends on the sensitivity of the tube, the supply voltage and the intensity of the radiation. In practice p can be estimated by determining the fraction of positive half-cycles in which one or more discharges occur. The probability of ignition being p, and the probability of non-ignition thus being $1 - p$, then with the aid of (2) and (3) we can set up the following statistical model for the output signal:

$$x_{n+1} = \left\{\begin{aligned} &\lambda_1 x_n + c && \text{with probability } p, \\ &\lambda_2 x_n && \text{with probability } (1-p), \end{aligned}\right\} \quad (6)$$

where $0 < \lambda_1, \lambda_2 < 1$ and $c > 0$.

When the system is switched on, the output signal begins at about the zero level. The maximum value x_{max} that the output voltage can reach can be calculated in a simple way from the following equation:

$$x_{max} = \lambda_1 x_{max} + c.$$

This leads to:

$$x_{max} = \frac{c}{1-\lambda_1}. \quad \ldots\ldots \quad (7)$$

Table II gives the values of c, λ_1, λ_2 and x_{max} for our four values of the output capacitance C_2[2].

Table II. The values of c, λ_1, λ_2 and x_{max} from equations (6) and (7) for four different values of C_2.

C_2	c	λ_1	λ_2	x_{max}
1 μF	0.391	0.947	0.956	7.39
2.2 μF	0.183	0,976	0.980	7.48
4 μF	0.101	0.986	0.989	7.52
6 μF	0.068	0.991	0.992	7.53

Choice of alarm level

We now wish to determine an alarm level d for the output voltage $V(t)$ at which with p constant, a false alarm due to the output voltage dropping below the value d while the flame is alight will occur on an average no more than twice in a year of continuous operation. Since the initial value of the output voltage x_0 can be zero, it will be necessary, in order to avoid a false alarm at the beginning of the process, to make allowance for a certain starting period of n_0T seconds. For an exact statement of the above condition to be met by the alarm level, we make use of *distribution functions*:

$$F_n(x) = P[\underline{x}_n \leqq x] \quad \text{for } n = 0, 1, 2, \ldots \quad \ldots \quad (8)$$

Expressed in words, $F_n(x)$ is the probability that $\underline{x}_n$, the output voltage at the time $t = nT$, will be smaller than or equal to x volts. The functions $F_0(x)$, $F_1(x)$, $F_2(x)$, etc., each represent a probability distribution, hence their name. Quantities such as $\underline{x}_0$, $\underline{x}_1$, $\underline{x}_2$ etc., which may assume different values in accordance with a particular probability distribution, are called *stochastic* quantities, and are denoted by underlining the symbol.

Determining the alarm level d which meets the conditions named above now amounts to finding the highest value of x for which:

$$F_n(x) \leqq 10^{-9} \quad \text{for each } n \geqq n_0. \quad \ldots \quad (9)$$

This follows since there are about 1.6×10^9 cycles in a year if the supply frequency is 50 Hz, so that if we choose for d the highest x value defined by (9), the number of false alarms expected in a year is equal to:

$$\sum_{n=n_0}^{n_0+1.6\times10^9} P[\underline{x}_n \leqq d] = \sum_{n=n_0}^{n_0+1.6\times10^9} F_n(d) \leqq 1.6\times10^9 \times 10^{-9} < 2. \quad \ldots\ldots \quad (10)$$

Here we have made use of the theorem that the expectation of the sum is equal to the sum of the expectations, even if the events are dependent.

[2] More decimal places of c and λ_1 than are quoted in Table II were used in the calculations of x_{max}.

The distribution functions $F_0(x)$, $F_1(x)$, $F_2(x)$, etc., are monotonic increasing functions, for which:

$$\left.\begin{aligned} &F_n(x) = 0 \quad \text{for} \quad x < 0, \\ 0 \leqq F_n(x) \leqq 1 \quad &\text{for} \quad 0 \leqq x < \frac{c}{1-\lambda_1}, \\ &F_n(x) = 1 \quad \text{for} \quad \frac{c}{1-\lambda_1} \leqq x. \end{aligned}\right\} \quad (n = 0, 1, 2, \ldots) \quad . \; . \quad (11)$$

To explain in words: the output voltage x cannot be negative, so the probability that x_n will be less than x for $x < 0$ is therefore equal to zero; also x is in every case smaller than $x_{\max} = c/(1-\lambda_1)$, so that for $x \geqq c/(1-\lambda_1)$ this probability is equal to 1. The curves of the functions $F_{n_0}(x)$, $F_{n_0+1}(x)$, $F_{n_0+2}(x)$, . . ., which are required for determining d with the aid of equation (9), may for example take the form shown in *fig. 11*. If these curves are known, the required alarm level d can then be simply determined in the way indicated schematically in this figure.

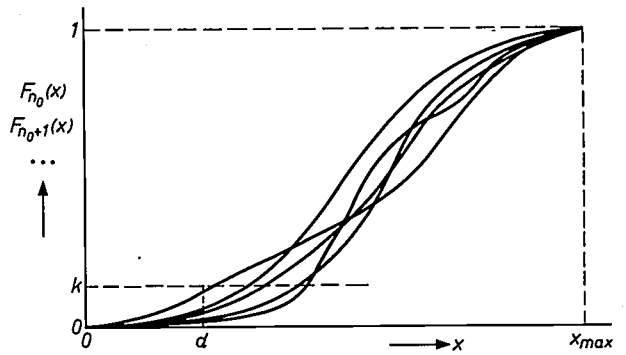


Fig. 11. Possible form of some distribution functions $F_n(x)$ for $n \geqq n_0$, and the determination of d as the highest value of x where $F_n(x) \leqq k$ for all $n \geqq n_0$.

Starting from the function $F_0(x)$, representing the initial distribution with which the process begins (i.e. at $t = 0$), we shall now calculate the functions $F_1(x)$, $F_2(x)$, etc., up to $F_m(x)$ where m is of arbitrary magnitude (in ten years m would have reached $m = 1.6 \times 10^{10}$) in order to be able to determine the highest value of x for which (9) is in fact valid. This we do by determining the relation between $F_{n+1}(x)$ and $F_n(x)$.

From equation (6) we have:

$$P[\underline{x}_{n+1} \leqq x] = pP[(\lambda_1 \underline{x}_n + c) \leqq x] + (1-p)P[\lambda_2 \underline{x}_n \leqq x]$$
$$= pP\left[\underline{x}_n \leqq \frac{x-c}{\lambda_1}\right] + (1-p)P\left[\underline{x}_n \leqq \frac{x}{\lambda_2}\right] \quad . \; . \; . \quad (12)$$

With the aid of definition (8) this gives the following relation between the distribution functions:

$$F_{n+1}(x) = \left\{\begin{aligned} &0 && \text{for } x < 0, \\ &pF_n\left(\frac{x-c}{\lambda_1}\right) + (1-p)F_n\left(\frac{x}{\lambda_2}\right) && \text{for } 0 \leqq x < \frac{c}{1-\lambda_1}, \\ &1 && \text{for } \frac{c}{1-\lambda_1} \leqq x. \end{aligned}\right. \quad . \; . \; . \quad (13)$$

This equation gives the relation between the functions $F_{n+1}(x)$ and $F_n(x)$ for every value of x; an equation of this kind is known as a *functional equation.*

Proceeding from an initial distribution $F_0(x)$, it is now possible in principle to calculate the functions $F_1(x)$, $F_2(x)$, etc., with the aid of equation (13) for a finite number of values of x by the repeated application of the functional equation (13). As an example *fig. 12* gives the calculated distribution functions $F_1(x)$, $F_2(x)$, $F_3(x)$ and $F_4(x)$ for a fictitious case in which $\lambda_1 = 0.855$, $\lambda_2 = 0.916$, $c = 2.58$ and $p = 0.7$. This case

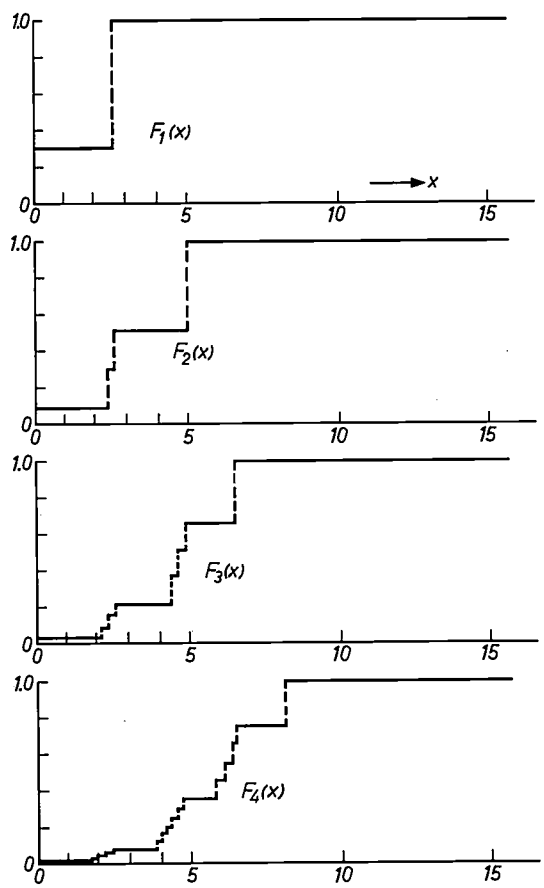


Fig. 12. Starting from the initial distribution $F_0(x)$ given in (14), the distribution functions $F_1(x)$, $F_2(x)$, $F_3(x)$ and $F_4(x)$ have been calculated for specific values of λ_1, λ_2, p and c by means of the functional equation (13). It can be seen that $F_4(x)$ is already fairly complicated; it is therefore not practicable to use this method for calculating all the functions needed for determining d.

was based on the initial value $x_0 = 0$, corresponding to the initial distribution:

$$F_0(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } 0 \leqq x. \end{cases} \quad \ldots\ldots \quad (14)$$

We can see from fig. 12 that even if we start with a very simple F_0, the form of the distribution function F_4 is already quite complicated.

Although it is thus theoretically possible to calculate the series $F_1(x)$, $F_2(x)$, ..., $F_n(x)$ at any desired point with the aid of the functional equation, the calculations required become impossibly involved when n is no longer small. Evidently, if we wish to determine F_n at a point x_1, we shall need to know F_{n-1} at *two* points, i.e. at:

$$x_{11} = \frac{x_1 - c}{\lambda_1} \quad \text{and} \quad x_{12} = \frac{x_1}{\lambda_2}.$$

In order to determine F_{n-1} at these points, we need to know F_{n-2} at *four* points, i.e. at:

$$x_{21} = \frac{x_{11} - c}{\lambda_1}, \; x_{22} = \frac{x_{11}}{\lambda_2}, \; x_{23} = \frac{x_{12} - c}{\lambda_1}, \; x_{24} = \frac{x_{12}}{\lambda_2}.$$

And so on. It is easily seen that the total number of calculations necessary in this iterative procedure increases very rapidly. A direct numerical solution of the functional equation is therefore not possible, even with a large computer. Since equation (13) cannot be solved analytically either, the only hope left is to try and find a method of approximation. We have found such a solution for this problem in the form of an iterative procedure which can be used to determine closely adjacent *upper* and *lower limits* for the desired alarm level.

Determining a lower and an upper limit for the alarm level

The procedure for approximating to the alarm level involves the determination of two distribution functions $G_{n_0}(x)$ and $H_{n_0}(x)$ [3] in such a way that for all x and for all $n \geqq n_0$ we have:

$$H_{n_0}(x) \leqq F_n(x) \leqq G_{n_0}(x). \quad \ldots\ldots \quad (15)$$

This situation is shown schematically in *fig. 13*.

With the aid of the functions $G_{n_0}(x)$ and $H_{n_0}(x)$ we can now find a lower limit and an upper limit for the alarm level d. For if y' is the greatest value of x at which $G_{n_0}(x) \leqq 10^{-9}$, and y'' is the greatest value of x at which $H_{n_0}(x) \leqq 10^{-9}$, then it follows from (9) and (15) that:

$$y' \leqq d \leqq y''. \quad \ldots\ldots \quad (16)$$

The various curves are shown in *fig. 14*. If the two functions $G_{n_0}(x)$ and $H_{n_0}(x)$ can now be chosen such that y' and y'' differ only very slightly, then y' is a very good approximation to the alarm level d, because for all $n \geqq n_0$ we now have:

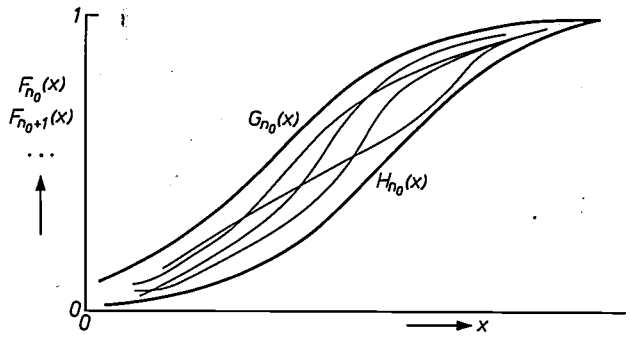


Fig. 13. To approximate to the alarm level, two distribution functions $G_{n_0}(x)$ and $H_{n_0}(x)$ are determined which, for every value of x, are respectively larger and smaller than all functions $F_n(x)$ for $n \geqq n_0$.

$$F_n(y') \leqq G_{n_0}(y') \leqq 10^{-9}. \quad \ldots \quad (17)$$

The restriction $n \geqq n_0$ means in practice that when y' is used as the alarm level it is necessary to make allowance for a starting period of $n_0 T$ seconds.

We shall indicate briefly the method by which the function $G_{n_0}(x)$ can be constructed. $G_{n_0}(x)$ is the function of rank n_0 from a series of step functions $G_0(x)$, $G_1(x)$, $G_2(x)$, ... Here $G_0(x)$ is a given initial distribution, defined as follows:

$$G_0(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } 0 \leqq x. \end{cases} \quad \ldots\ldots\ldots \quad (18)$$

For determining the other functions in this series, the interval between $x = 0$ and $x = x_{\max} = c/(1 - \lambda_1)$ is divided by means of the points $\xi_0 = 0 < \xi_1 < \xi_2 < \ldots < \xi_{N-1} < \xi_N = c/(1-\lambda_1)$ into N equal sub-intervals. The function $G_1(x)$ is now determined at the points $\xi_0, \xi_1, \ldots, \xi_N$ by applying the functional equation (13). This gives:

$$G_1(\xi_i) = pG_0\left(\frac{\xi_i - c}{\lambda_1}\right) + (1-p)G_0\left(\frac{\xi_i}{\lambda_2}\right) \quad (i = 0, 1, 2, \ldots N). \quad \ldots \quad (19)$$

At the intermediate points, however, $G_1(x)$ is differently defined, as follows:

$$G_1(x) = \begin{cases} 0 & \text{for } x < \xi_0, \\ G_1(\xi_i) & \text{for } \xi_{i-1} < x < \xi_i, \\ 1 & \text{for } \xi_N < x. \end{cases} \quad \ldots \quad (20)$$

At such a point the function therefore acquires the value which from equation (13) corresponds to the upper limit of the sub-interval in which the point lies.

In this way a step function $G_1(x)$ is obtained which is a distribution function and for which, at every value of x, we can write:

$$G_1(x) \geqq F_1(x).$$

Here $F_1(x)$ is a distribution function determined by the functional equation (13) for $n = 0$ with an arbitrary initial distribution $F_0(x)$.

In the same way, if $G_{n-1}(x)$ is known, we can calculate the function $G_n(x)$ at the points $\xi_0, \xi_1, \ldots \xi_N$ using the functional equation (13), so that

$$G_n(\xi_i) = pG_{n-1}\left(\frac{\xi_i - c}{\lambda_1}\right) + (1-p)G_{n-1}\left(\frac{\xi_i}{\lambda_2}\right) \quad (i = 0, 1, 2, \ldots N), \quad \ldots \quad (21)$$

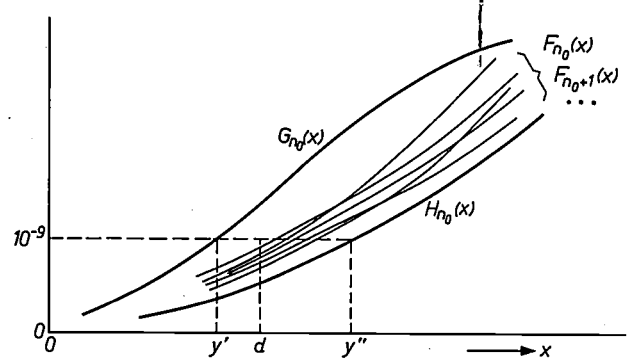


Fig. 14. Determination of the lower limit y' and the upper limit y'' for the alarm level d.

while at the intermediate points the function $G_n(x)$ is defined as follows:

$$G_n(x) = \left\{ \begin{array}{ll} 0 & \text{for } x < \xi_0, \\ G_n(\xi_t) & \text{for } \xi_{t-1} < x < \xi_t, \\ 1 & \text{for } \xi_N < x. \end{array} \right\} . \quad . \quad (22)$$

The step function $G_n(x)$, which is easily constructed in this way, is a distribution function for which it can be proved that, at every value of x,

$$G_n(x) \geqq F_n(x).$$

Here $F_n(x)$ is again a distribution function obtained by the iterative application of the functional equation (13), starting from an arbitrary initial distribution $F_0(x)$.

By induction on n we can prove for the series of step functions $G_n(x)$ ($n = 1, 2, \ldots$) that:

$$G_n(x) \geqq G_{n+1}(x).$$

It follows from the last two inequalities that, for any given n_0, we can write:

$$G_{n_0}(x) \geqq F_n(x) \quad \text{for each} \quad n \geqq n_0, \quad . \quad . \quad . \quad . \quad . \quad (23)$$

which satisfies the first inequality of equation (15).

In roughly the same way, starting from the initial distribution:

$$H_0(x) = \left\{ \begin{array}{ll} 0 & \text{for} \quad x < \dfrac{c}{1-\lambda_1}, \\ 1 & \text{for} \quad \dfrac{c}{1-\lambda_1} \leqq x, \end{array} \right\} \quad . \quad . \quad . \quad (24)$$

we can construct a series of step functions $H_n(x)$ such that, for any given n_0, we have:

$$F_n(x) \geqq H_{n_0}(x) \quad \text{for each} \quad n \geqq n_0. \quad . \quad . \quad . \quad . \quad (25)$$

The second inequality of equation (15) is then satisfied.

To calculate the function $G_{n_0}(x)$ we must apply equation (21) n_0 times. A similar operation is required for $H_{n_0}(x)$. When these functions have been calculated, the lower limit y' and the upper limit y'' can easily be determined. The iterative procedure for calculating $G_{n_0}(x)$ and $H_{n_0}(x)$ can be carried out on a digital computer for given values of λ_1, λ_2, c and p. The number of interval points N and the number of steps n_0 can still be freely chosen in this procedure. After comparing the results of calculations for $N = 500$ and $N = 1000$, we decided to take $N = 1000$ in the case described. The value of n_0 was determined from case to case with the aid of a fixed criterion.

[3] A distribution function is here understood to be a monotonic increasing function which satisfies equation (11).

[4] The computer programme was written by S. Jonkmans, who with J. Prakken also performed many of the calculations.

After the functions $G_{n_0}(x)$ and $H_{n_0}(x)$ had been calculated for different values of the ignition probability p, the lower limit y' and the upper limit y'' were in this way plotted as a function of p (*fig. 15*) for the four different values of the capacitance C_2 in fig. 7.

As already pointed out, the results are applicable after a starting period of n_0T seconds. This means that the lower limits derived can be used as the alarm level with a continuously burning furnace flame provided the alarm mechanism is set so that it will not respond until n_0T seconds after the UV detector has been put in operation. For the case $C_2 = 1$ μF, this starting period is five seconds. In the other three cases the starting period is ten seconds. Since the upper and lower limits are close together, the lower limits are very suitable for use as alarm level. The lower limits for all four cases are shown in *fig. 16*.

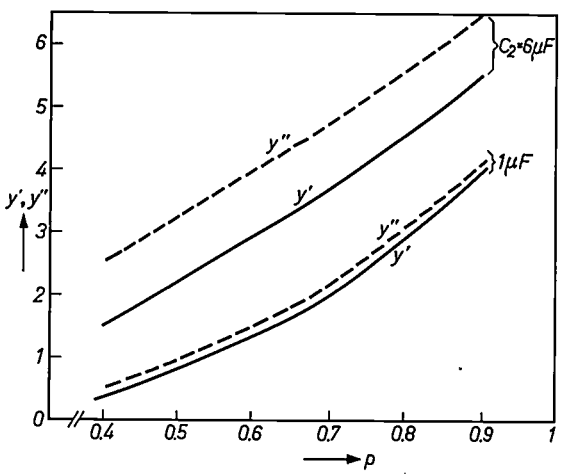


Fig. 15. The lower limit y' and the upper limit y'' for the alarm level as a function of the ignition probability p. The limits have been calculated from the functions $G_{n_0}(x)$ and $H_{n_0}(x)$ used for approximating the distribution functions $F_n(x)$. The calculations were carried out for the circuit in fig. 7, with the four values given for C_2 at the beginning of the article (1 μF, 2.2 μF, 4 μF and 6 μF). Only the results for 1 μF and 6 μF are plotted here.

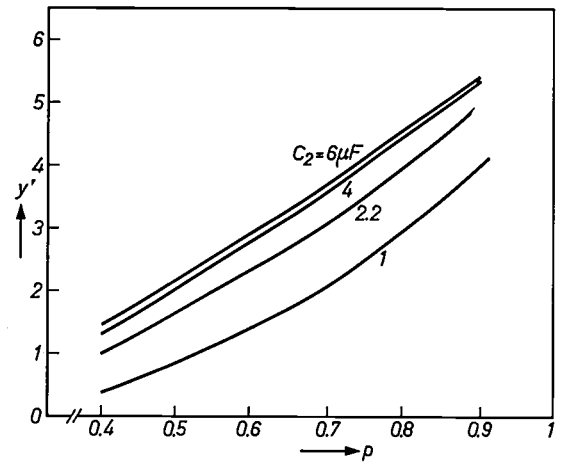


Fig. 16. Since the lower and upper limits for every value of C_2 are found to lie close together (see fig. 15), the lower limits can very suitably be used as an alarm level. These lower limits y' are shown here for the four values of C_2 as a function of p.

Conclusion

The results of the calculation of the lower and upper limits for the alarm level, represented in fig. 15, show that in the cases considered these limits lie fairly close together. This is a remarkable result for what is essentially an unorthodox method. It should moreover be noted that the limits are applicable after a starting period of known length. We have also seen that the starting periods for the various values of the output capacity are short, being no more than ten seconds. From all this it may be concluded that the lower limits can very appropriately be used as an alarm level.

Since we are concerned here with an iterative procedure, in which the same operation is applied a large number of times, the method described can most efficiently be carried out with the aid of a computer. The method described in this article in fact gives general lower and upper limits for distribution functions $F_n(x)$ $(n \geqq n_0)$, and appears therefore to be suitable for many other applications apart from the one described here.

Summary. In a flame-failure control system using the UV-sensitive cold-cathode tube described in part I of this article, the statistical nature of the ignition process gives random fluctuations in the output signal. These fluctuations have to be taken into account when choosing an alarm level. This part of the article describes a method which permits the alarm level to be chosen in such a way that a false alarm will occur no more than twice in a year of continuous operation. First, a statistical model is derived for the output voltage of the detection circuit with the flame burning. The probability that this output voltage will drop at a certain moment below a specific value is given by a distribution function. Since, in the case described, exact calculation of these distribution functions is impossible, even with the aid of a large computer, an approximative method was developed in which these functions are bounded by two simpler distribution functions. A lower and an upper limit for the alarm level can thus be determined, and the lower one can serve as the alarm level. A numerical example for a particular case is given.

Design and characteristics of present-day photomultipliers

G. Piétri and J. Nussli

The photomultiplier is a particularly useful instrument for the study of weak or rapidly occurring effects in which light is produced, and it has proved indispensable for many investigations in the field of nuclear physics. The progressive improvement of this device during the last twenty years has been accompanied by the development of an astonishing diversity of types to suit the various different applications; the Philips catalogue today lists about a hundred types. The article below gives an account of the general considerations underlying the development of photomultipliers and also discusses the design of photomultipliers which have an extremely rapid response or other unusual features.

A photomultiplier is a photoemissive cell associated with an electron multiplier (*fig. 1*), forming an amplifier which, for most of the time, possesses virtually negligible inherent noise. Another characteristic feature, at first sight rather paradoxical, is that no noise temperature can be assigned to it, since its noise is independent of temperature. Moreover, the two fundamental effects on which it depends — photoemission and secondary emission — are extremely fast. The time resolution is limited only by the random variations of the initial velocities at which the emitted electrons leave the electrodes.

These features, combined with its very high sensitivity, make the photomultiplier a particularly useful instrument for studying effects in which light is produced. In certain conditions its sensitivity is in fact immeasurably superior to that of the photographic film and exceeds that of the retina of the eye, considering the amount of information gathered in a time say of the order of one second.

The photomultiplier has proved to be indispensable for the observation and counting of scintillations produced by the passage of nuclear radiation through certain materials (scintillators) [1]. The study of these scintillations makes it possible to determine: a) Their frequency, by counting the individual pulses; rates of up to 10^7 counts per second are possible. The activity of the source can be deduced from this information. b) The energy of the incident particle, the number of photons produced being, within certain limits, proportional to this energy. c) The time at which the particle appears, with an accuracy approaching 10^{-10} second.

G. Piétri, Ing. E.S.E. Radio, L. ès Sciences (deputy director) and J. Nussli, L. ès Sciences, are with Laboratoires d'Electronique et de Physique Appliquée (L.E.P.), Limeil-Brévannes (Val-de-Marne), France.

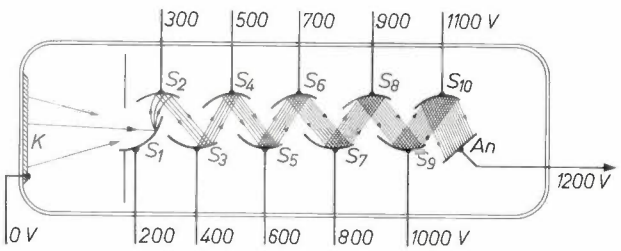


Fig. 1. Principle of a photomultiplier. The photocathode *K*, illuminated by the radiation to be detected, emits photoelectrons which are multiplied by secondary emission from electrodes S_1 to S_{10} as they travel towards the anode *An*. The voltages indicated at the electrodes are only intended to give an idea of the kind of magnitudes to be expected.

d) The variation of the luminous flux with time during the scintillation, with an accuracy approaching 10^{-9} second.

The design of present-day photomultipliers has been much influenced by their widespread application in nuclear techniques, but these applications have in turn been a great stimulus in the steady improvement in performance of these tubes. Photomultipliers are now available for all applications where weak or very fast luminous effects have to be observed.

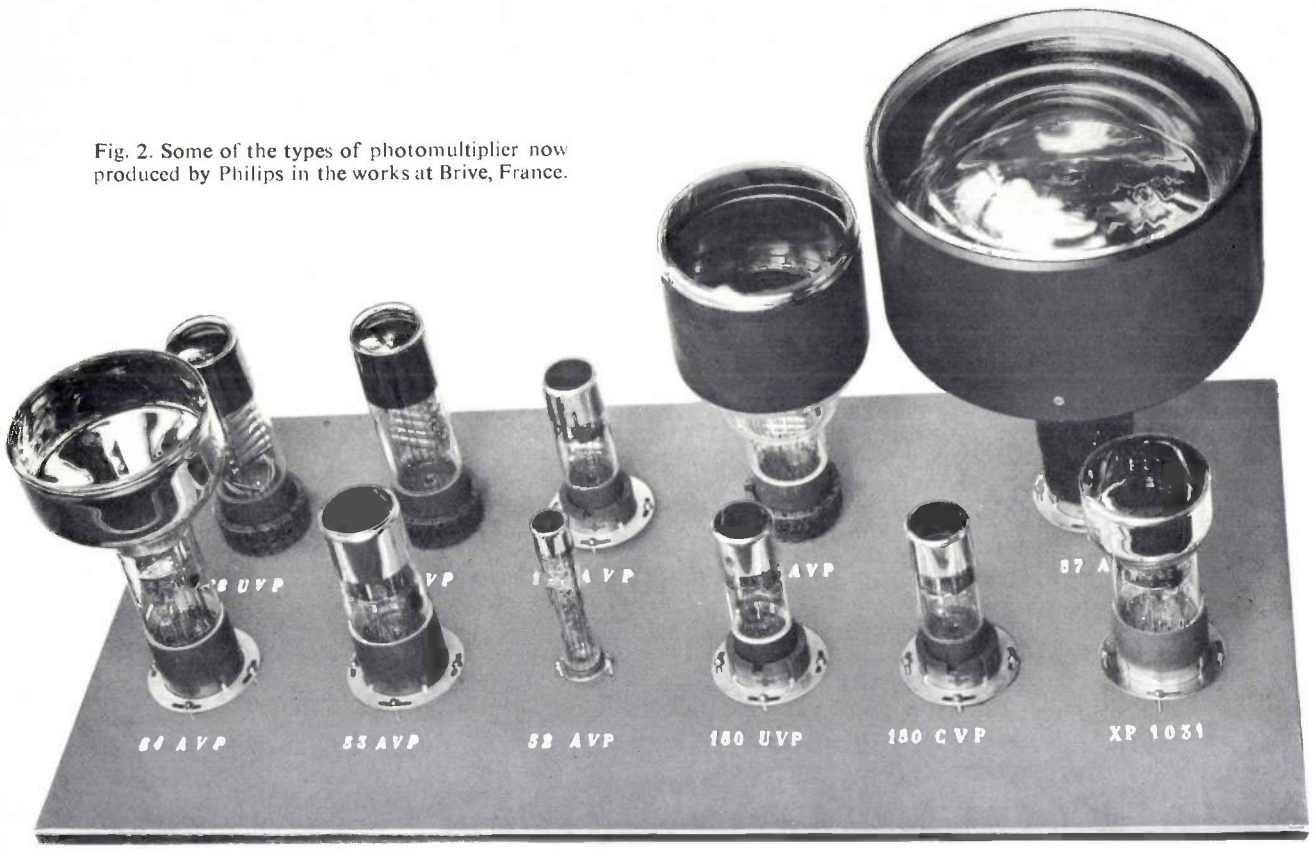
The development of Philips photomultipliers was started in 1952 at Laboratoires d'Electronique et de Physique Appliquée (L.E.P.) near Paris. In 1954 large-scale production was started by Philips in the works at Brive-la-Gaillarde with the type 150 AVP, while L.E.P., after their work on the 53 AVP, started on the development of fast photomultipliers, in particular the

[1] See for example: G. A. Morton, The scintillation counter, Adv. in Electronics **4**, 69-107, 1952. J. B. Birks, The theory and practice of scintillation counting, Pergamon Press, London 1964. R. Champeix, H. Dormont and E. Morilleau, A photomultiplier tube for scintillation counting, Philips tech. Rev. **16**, 250-257, 1954/55.

56 AVP, and special photocathodes. In about 1960 a development laboratory was set up at Brive. One of the developments from this laboratory has been that of very small photomultipliers with cup-shaped dynodes, such as the types 152 AVP and PM 400 [2]. The current Philips catalogue lists a very wide range of photomultipliers, consisting of about a hundred different types (*fig. 2*).

Before describing the work that led up to the production of these tubes, with special emphasis on the high-performance types, we should discuss the physical effects which are applied in the photomultiplier, the general principles of the design and construction, and the essential characteristics.

Fig. 2. Some of the types of photomultiplier now produced by Philips in the works at Brive, France.



Principles of the photomultiplier

The operation of a photomultiplier depends on two fundamental types of electronic emission: photoelectric emission and secondary emission.

Photoelectric emission by any material is the result of three successive processes:

1) Absorption of the incident photon by the transfer of its energy $h\nu$ to an electron. For any material a potential, known as the photoelectric work function Φ_{ph}, can be defined which is equal to the energy difference between the highest level from which photoelectrons can originate and the vacuum level. If an electron thus excited is to be able to leave the material, Einstein's relation must be satisfied:

$$h\nu \geqq e\Phi_{ph}.$$

2) Diffusion of the electron towards the surface, with energy loss due to collisions. The energy which the electron retains when it escapes from the surface is higher the weaker the absorption of the material for low-energy electrons, and the nearer the electron is to the surface when it is excited.

3) Transmission through the surface: wave mechanics shows that the probability of this is not equal to unity immediately the energy of the electron becomes greater than the work function, but tends towards unity at an energy which has at least twice that value.

Since the materials used nowadays are almost invariably semiconductors, we should draw attention to an important difference between photoelectric emission from semiconductors and that from metals (*fig. 3*). In *metals* the conduction band is partly filled up to the Fermi level. This is therefore an occupied level, that is to say electrons, when excited by photons of corresponding energy, may be emitted from this level to a vacuum. Since this level is also the highest level from which electrons may be emitted by the thermionic effect, the photoelectric work function Φ_{ph} is identical with the thermionic work function Φ_{th}.

In a *semiconductor* the Fermi level is in the forbidden zone (in

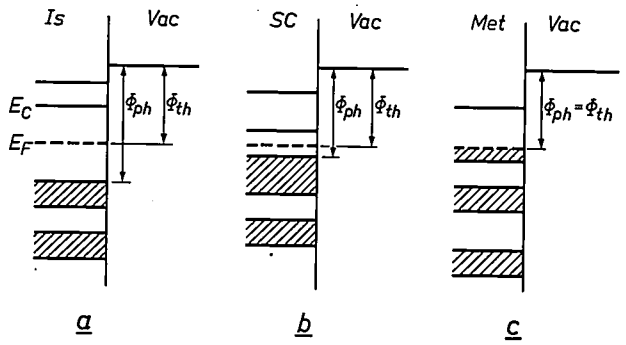


Fig. 3. Model of energy bands in an insulator (*Is*), an intrinsic semiconductor (*SC*) and a metal (*Met*), all with the same thermionic work function Φ_{th} but with different photoelectric work functions Φ_{ph}. The occupied energy zones are shown hatched.

the middle of it for an intrinsic semiconductor). This level again defines the thermionic work function, since it is the mean level of origin for thermionically emitted electrons (equilibrium exists between the population at the top of the valence band and at the bottom of the conduction band). In general, however, the occupation of the conduction band is much too small to give rise to any appreciable photoelectric effect; photoelectric emission therefore originates only from the top of the valence band. This shows that in semiconductors the photoelectric work function Φ_{ph} is always higher than the thermionic work function Φ_{th} [3].

A consequence of the processes (2) and (3) is that appreciable photoelectric emission can only take place if the energy of the photons substantially exceeds the threshold energy. The photoelectrons then emitted will have a surplus energy which appears in kinetic form. For example, in the case in which we are interested their initial velocity, which will vary statistically in magnitude and direction, will lie within a range corresponding to energies between about 0 and 1 eV.

The conditions for emission mentioned above are satisfied reasonably well in several semiconductors that can be produced in the form of thin films (of thickness less than 0.1 μm) and with a large surface area by the successive evaporation of various layers under vacuum. However, these films — at any rate the ones which are sensitive to visible light — are all alkaline (usually compounds of caesium with antimony or with silver and oxygen which are unstable at temperatures higher than 150 °C and are rapidly destroyed by the least excess of oxygen).

Another factor which has to be taken into account in the preparation of the films is that they are almost invariably semi-transparent: the light penetrates from the side opposite to the one from which the electrons are emitted. The film is deposited directly on to the glass of the tube, or, to be more exact, on to a window situated at the end of a cylinder (see fig. 1). With this arrangement, the light excited in a scintillator covering the whole of the outside of the window is received over the largest possible useful angle. The thickness of the film is however very critical in this case, because the light has to be absorbed, but not too far from the emissive surface. Incidentally, the arrangement described permits much easier location of the electrodes used for focusing the electrons on to the input of the multiplier unit.

An electron multiplier consists of a succession of a variable number of targets called dynodes. When their surface is struck by electrons of appropriate energy (primary electrons) these dynodes emit secondary electrons. The geometry of the dynodes and any following electrodes, and of the applied electric (or perhaps magnetic) fields, is such that the electrons emitted by each dynode are accelerated towards the next one and strike it with an energy between 100 and 500 eV. There are several well-known arrangements which satisfy these conditions and for all types now in production we have used the system proposed by Rajchman (see fig. 1 and, for example, fig. 14), which is a linearly focused multiplier of a purely electrostatic type. This system is fairly universally applied since it can generally be expected to give the most rapid response.

Secondary emission obeys laws which are even more complex and less well-known than the laws of photoelectric emission. Metals are very poor secondary emitters. Good secondary emitters are either insulators (MgO, BeO and the alkali halides), or semiconductors (those which are also photoemissive). One advantage of insulators is that they can be prepared outside the tube before assembly, since they are relatively insensitive to air, but they have to be used in the form of extremely thin films if surface-charge effects are to be avoided. At present we are mainly using Ag-MgO-Cs, obtained by heat treatment of an Ag-Mg alloy in an oxidizing atmosphere, followed by treatment in caesium vapour and oxidation. The resultant material has a stable secondary-emission coefficient of between 3 and 8 for primary electrons with an energy of between 100 and 300 eV (*fig. 4*). The value of this coefficient is practically independent of the primary current up to values of the order of 100 μA/cm² or more, thus meeting the requirements for linearity.

These dynodes have a fairly small useful surface area for their volume. This makes it necessary to devise a more or less complex focusing electrode system (the "input optical system") between the photocathode and the first dynode. It will be noted, moreover, that the

[2] E. Morilleau, Production industrielle des photomultiplicateurs (technologie et mesures de contrôle), Acta electronica **5**, 39-51, 1961.

[3] An extensive treatment of the photoemissive effect has recently been given in this journal: J. van Laar and J. J. Scheer, Photoemission of semiconductors, Philips tech. Rev. **29**, 54-66, 1968 (No. 2). The energies $e\Phi_{ph}$ and $e\Phi_{th}$ are referred to in this article as E_d and φ ($= E_d - \delta$) respectively.

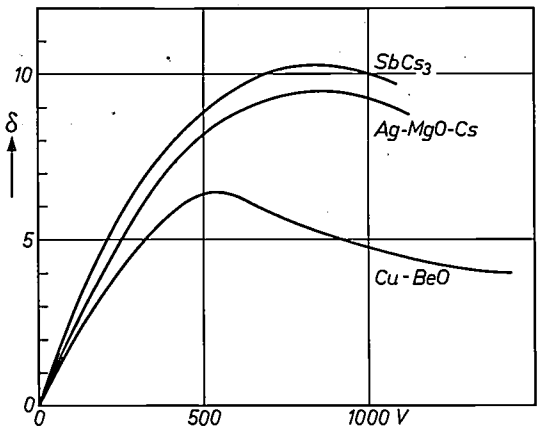


Fig. 4. Variation of the coefficient of secondary emission δ as a function of the energy of the primary electrons, for various materials.

form and arrangement of the first two or three dynodes are somewhat different. This region, referred to as the coupling region between the input optics and the last part of the multiplier, which consists of a succession of identical stages, is rather critical on account of the marked curvature of the paths between the first and the second dynode.

The secondary electrons also leave the material at random initial velocities, corresponding to energies about ten times greater than those of the photoelectrons. With velocities of this order it is difficult to focus the secondary electrons, and in fact some slight loss always has to be accepted (of the order of 10%).

The electrons emitted by the last dynode are collected by the anode. This usually consists of a grid situated at a short distance from the last dynode and is traversed by the electrons travelling to that dynode.

Finally, a word or two about certain special problems encountered in the manufacture of this type of electron tube. It follows from what has been said that the photocathode cannot withstand high temperatures or exposure to air. It therefore has to be prepared inside the tube itself, on the vacuum pump, before sealing-off. The inside of the tube must therefore contain everything that is needed for this preparation — antimony to be evaporated, as well as a mixture of alkaline salts and a reducing metal for the preparation of the alkali "*in situ*". All these materials and the dynodes cannot be degassed at the high temperatures normally used in degassing the electrodes of ordinary electronic tubes. The only possible course of action is to pump the tube for a very long time (12 hours or more) at less than 400 °C. The forming of the cathode then begins by the cold evaporation of the antimony layer (with a thickness of the order of 10 nm). The caesium for the $SbCs_3$ is next prepared in excess by reduction of caesium chromate; since the reaction with antimony is sufficiently fast only above 100 °C, the tube is raised to a temperature of about 120 °C. The vapour pressure of the caesium is then extremely high and a monolayer settles on all the surfaces. This monolayer cannot be removed, since further pumping is prohibited by the presence of the cathode. (With no further pumping it is also impossible to eliminate all the gases liberated during the preparation of the caesium, so that after the tube is sealed off the pressure is not likely to be much lower than 10^{-6} torr, in spite of the presence of a getter.) The caesium deposited on the dynodes considerably reduces their work function. As we shall see later, this is of paramount importance for the dark current. The caesium which is not absorbed is evacuated in the pump trap.

One final remark will illustrate the variety of problems encountered. The glass used for the envelope has to meet the following specific requirements: a) It must not devitrify nor go black during the prolonged and delicate operation of sealing the end window. b) It must be resistant to attack by caesium at the temperature at which the cathode is formed. c) It must contain a minimum of potassium, because of the constant presence of the radioactive isotope ^{40}K. These requirements are to all intents and purposes only met by the borosilicates (the third condition is never entirely satisfied), but in this case the metal seals have to be either tungsten or an alloy of the Fe-Ni-Co type. These materials are difficult to use and necessitate stringent and repeated quality control, in particular to ensure a vacuum-tight seal.

Characteristics of a photomultiplier

We shall now recall some of the chief quantities that characterize a photomultiplier.

Photocathode sensitivity

The sensitivity σ_K of the photocathode is a function of the wavelength λ of the incident light. It is expressed either in amperes per watt of incident radiation or by means of the quantum efficiency η_q, a quantity expressing the relation between the number of electrons emitted and the number of photons incident on the window of the tube ($\eta_q < 1$). The quantities η_q and σ_K are related by the expression:

$$\eta_q = \sigma_K \frac{hc}{e\lambda}, \quad \ldots\ldots (1)$$

where h is Planck's constant, c the velocity of light in vacuum, and e the electron charge.

For greater convenience another sensitivity has been defined which broadly characterizes the photoelectric properties of the cathode, namely the sensitivity to white light emitted by a tungsten filament lamp at a

colour temperature of 2850 °K. It is usually expressed in μA/lumen.

The highest quantum efficiencies, obtained for wavelengths of the order of 0.4 μm, are about 30%, corresponding to about 100 mA/watt. The colour response of several different types of photocathode is shown by the spectral characteristics in *fig. 5*. Each cathode is characterized by a particular position of the photoelectric threshold. The S11 cathode is well adapted to the blue light emitted by scintillators. Since it is the least troublesome to prepare, this cathode is the one most widely used. A typical value for its overall sensitivity is 60 μA/lumen.

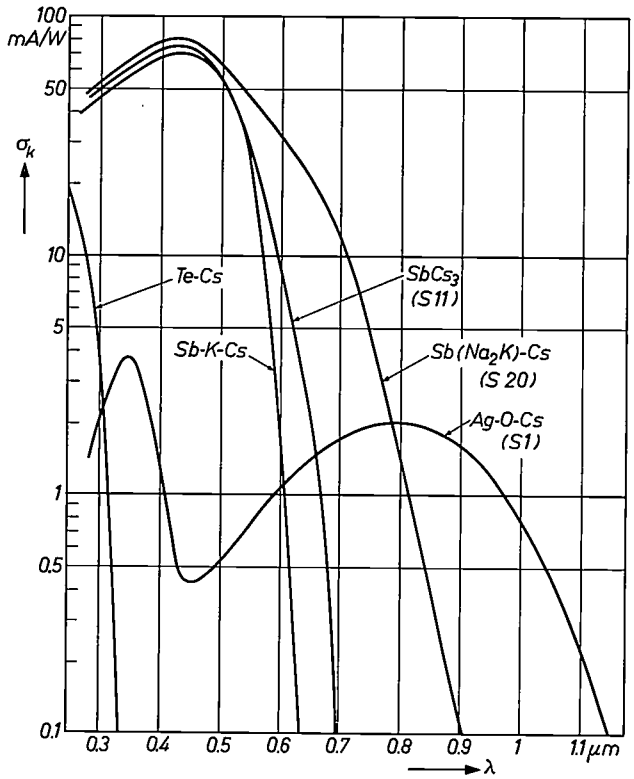


Fig. 5. Spectral sensitivity σ_k of photocathodes in current use.

It is important to note that, like all semiconductor characteristics, this sensitivity varies considerably from one tube to another. The sensitivity of the S11 may permissibly deviate by a factor greater than 2 (i.e. from 40 to 90 μA/lm). Again, the sensitivity is never perfectly constant at all points of the surface of a photocathode. A useful measure of the homogeneity of the sensitivity is given by the ratio $\sigma_{min}/\sigma_{max}$ [4] (in terms of white or monochromatic light; in the latter case $\sigma_{min}/\sigma_{max}$ may be a function of the wavelength). As a rule, however, the characteristic usually quoted is the mean overall sensitivity.

These different quantities are to a slight extent dependent on temperature [5].

Gain; collection efficiency

The gain is defined as the ratio of the anode current to the photocathode current. It is given by:

$$G = \eta_c \delta^n, \quad \ldots\ldots\ldots \quad (2)$$

where η_c is the collection efficiency of the input optical system and δ is the mean secondary-emission coefficient of the n dynodes. The gain of the multiplier proper may also be used; this is δ^n.

The collection efficiency is in general a function of the co-ordinates of the point of the cathode illuminated, and also of the wavelength, since, as we shall show later, this efficiency is affected by the initial velocities of the photoelectrons.

We may also define an anode sensitivity σ_a, equal to the product of G and the sensitivity σ_k (in white or in monochromatic light).

The coefficient of secondary emission varies almost linearly with the energy of the primary electrons in the range 100-200 eV, which is the one ordinarily used. The gain as a function of the tube supply voltage V thus varies approximately as V^n (*fig. 6*), which means that a very well stabilized power supply must be used. The highest gains achieved are of the order of 10^9.

The gain varies slightly with temperature [5] (a few tenths of a per cent per degree), but it is particularly subject to variations in the mean output current [6].

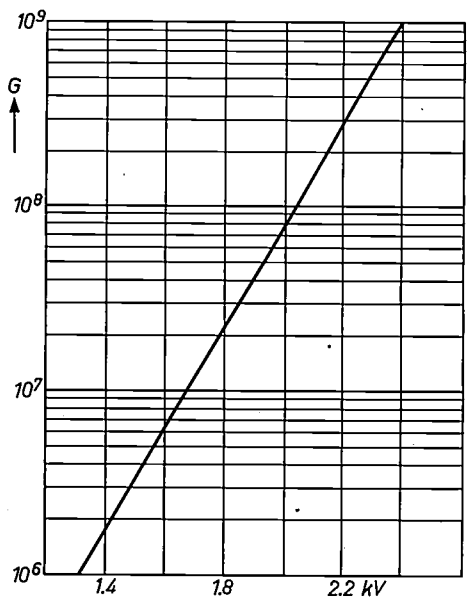


Fig. 6. Variation of the gain G of a multiplier with 14 dynodes of Ag-MgO-Cs as a function of the high voltage.

[4] G. Piétri and C. Arvin, J. Phys. Radium **19**, Suppl., 154A-159A, 1958.

[5] G. Laustriat and A. Coche, J. Phys. Radium **19**, 927-929, 1958, and **20**, 719-720, 1959. A. T. Young, Appl. Optics **2**, 51-60, 1963.

[6] I. Cantarell, IEEE Trans. on nuclear science **NS-11**, No. 3, 152-159, 1964.

Depending on the tube, reversible variations (positive or negative) in gain of several per cent are observed when the anode current varies from 0 to 100 μA. Currents higher than 1 mA usually cause an irreversible and progressive decrease of gain.

Dark current

The dark current is the current I_0 observed at the anode when no light falls on it. This current is most frequently initiated by electrons emitted from the cathode, either through thermionic emission or by excitation due to luminescent effects inside the tube. Instead of I_0, it is therefore more convenient to consider the ratio I_0/G, which varies less than I_0 with the high voltage.

Pulse-amplitude variations

If we consider the photomultiplier in association with the electronic circuits designed to process the information supplied by the photomultiplier, we may assign to it a resolving time τ. Let N be the number of photoelectrons emitted by the cathode during this time. This number will fluctuate from one measurement to another; the standard deviation of the probability distribution curve will be $\sqrt{N}$ and the width of the curve at half-height will be $2.36\sqrt{N}$ (if the distribution is Gaussian). At the first dynode, with secondary-emission coefficient δ_1, the number of secondary electrons will also fluctuate, but the relative fluctuation will be reduced by a factor of $\sqrt{\delta_1}$.

This effect can appear in various ways:

a) If we investigate the amplitude V of the pulses, due say to scintillations, and if we consider as a function of V the probability that the amplitude will be between V and $V + \mathrm{d}V$, we may define a resolution R which is the ratio $\Delta V/V_{\mathrm{peak}}$, ΔV being the width at half-height of the probability distribution curve, and V_{peak} the most probable amplitude. To a first approximation, R will be of the order of $2.36/\sqrt{N}$, if N is the mean number of photoelectrons per scintillation.

b) If we investigate the amplitude of the pulses produced by the photons arriving at a fairly low rate, so that all the pulses are well separated in time, we obtain a distribution curve of the amplitudes, which is called the *single-electron spectrum* (SES). Its form is determined by the fluctuations of the secondary emission of the first dynodes and its width will be of the order of $\sqrt{\delta_1}$. This spectrum is characteristic of the photomultiplier itself.

The "mean gain" δ^n of the multiplier system (cf. eq. 2) can be calculated from the mean amplitude of the single-electron spectrum, i.e. the amplitude at its "centre of gravity".

c) If we investigate the amplitude of the pulses in the total absence of light, we obtain the *dark-pulse spectrum* (DPS). This spectrum would be identical with the SES if all the pulses came from the cathode. This is never the case and the DPS is therefore very useful for studying the dark current.

d) If we observe a luminous effect modulated in amplitude, we have the well-known Schottky relation:

$$\overline{\Delta I_{\mathrm{k}}} = \sqrt{2e I_{\mathrm{k}} \Delta f} \text{ at the cathode . . .} \quad (3a)$$

and

$$\overline{\Delta I_{\mathrm{a}}} = G\sqrt{2e I_{\mathrm{k}} \Delta f p} \text{ at the anode, . .} \quad (3b)$$

where Δf is the bandwidth of the system and p is a factor slightly greater than unity, related to the fluctuations in the secondary emission. This is the Schottky (shot) noise which is associated with the signal and increases with the square root of the signal. It is independent of frequency and is therefore referred to as white noise.

Transit time; speed of response

To establish the speed of response of the photomultiplier it is useful to define the following quantities:

a) Pulse response.

If the cathode is exposed to an infinitely short pulse of light, the shape of the resultant electrical pulse characterizes the speed of the photomultiplier. In general this shape varies considerably from one type of tube to another and is very rarely Gaussian. It can include peaks later than the main one ("after-pulses"). The speed of response may be expressed in terms of the width of the pulse at half peak height and in terms of its rise time.

b) Mean transit time.

Again in the case of a very short pulse of light a certain time elapses between the moment at which the cathode is illuminated and the arrival of the resultant avalanche at the anode. The moment of arrival of this pulse can for example be taken as the moment of arrival of its "centre of gravity" (or of the point half-way up the rising edge of the pulse). The time elapsed will fluctuate from one measurement to another and its mean value is defined as the mean transit time. This is in general less than 5×10^{-8} s, and varies with the voltage V as $1/\sqrt{V}$.

c) Transit time fluctuations.

Since the instant at which the anode pulse appears (defined as indicated above) fluctuates from one measurement to another, it is more aptly described by a probability distribution. If each pulse is produced by a single photoelectron, the width at half-height of this distribution gives a much better measure of the accuracy of the photomultiplier for measuring very short times — in particular the time of flight

of particles — than is given by the "speed of response".

If there are N photoelectrons in each pulse, this fluctuation will decrease as N increases.

d) Frequency response.

Experience shows that when a photomultiplier detects a light signal modulated in amplitude at a frequency f, the output signal is a decreasing function of f. A 3 dB (or 6 dB) bandwidth can be defined, corresponding to a decrease of the output signal by a factor of $\sqrt{2}$ (or 2). The lower limit of the passband is of course $f = 0$, since the photomultiplier can amplify a direct current.

Linearity

With a constant light flux Φ the anode current I_a is at first proportional to Φ, i.e. $I_a = k\Phi$, up to the value $I_a = 100$ μA. With short pulses, however, I_a may safely exceed 100 μA by a considerable margin. From a certain value onwards, the current I_a does not increase as fast as the flux Φ. The limit of linearity is defined as the value of I_a at which the $I_a = f(\Phi)$ curve, expressed in dB, falls 1 dB below the line $k\Phi$.

Improvement of the characteristics of the photomultiplier

We shall now review some of the work done in our laboratories during recent years with the object of gaining a better understanding of the physics of the operation of the photomultiplier, and of improving their characteristics or adapting them to special applications.

Photocathode

We have already noted (and we shall discuss the subject in more detail later) that the fluctuation effects are closely related to the number of photoelectrons delivered by the cathode. Any improvement here therefore comes from an improvement of the sensitivity of the cathode. For example, the sensitivity of the S11 cathode ($SbCs_3$), which has been known for more than 30 years, has been doubled in the last ten years, and now approaches 100 μA/lumen [7]. This advance has been made possible through the constant improvement of the vacuum obtained with modern diffusion pumps, the progress made in the methods of cleaning the windows and materials used in the tubes, and the greater purity of the products used in the preparation of the cathodes. The deposition of a substrate of MnO_2 on the window has also been found to have a favourable effect.

The S1 cathode (Ag-O-Cs) has also been known for some considerable time. It is the only one which has appreciable sensitivity in the infra-red (1.2 μm), but its quantum efficiency remains poor (0.1 to 1%) in the whole of the useful wavelength range. The S20 cathode ($Sb(Na_2K)$-Cs) is at present the cathode with the greatest efficiency between 0.5 and 0.8 μm; in white light its sensitivity is as high as 250 μA/lm. But maintaining its four constituents in the correct ratio over the whole surface of the cathode presents such problems that the tubes which have this cathode are three times as expensive as equivalent tubes with the S11 cathode. These problems are chiefly related to the presence of the focusing electrodes in the input optical system, which obstruct the deposition of the alkalis on the cathode.

A new cathode is now being developed which looks as if it will supersede the S11 in photomultipliers used for scintillation counters, at least in certain cases. This is the Sb-K-Cs cathode, whose quantum efficiency may be better than 25% at 0.4 μm.

The extension of the sensitivity range of photomultipliers towards the *ultra-violet* has been and still is the subject of a great deal of investigation. The sensitivity range is in fact limited by the transparency of the window (*fig. 7*). Borosilicate glasses are transparent up to about 0.3 μm, provided they are thin enough.

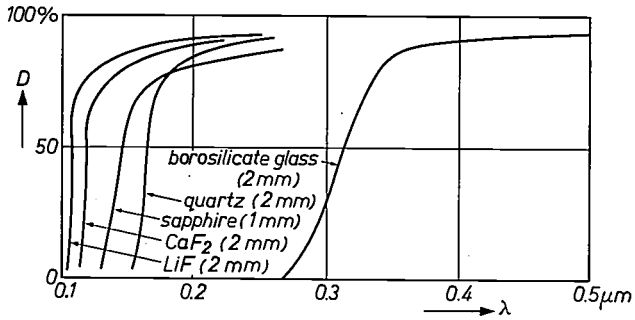


Fig. 7. Optical transmission D of various windows used in photomultipliers, as a function of wavelength.

Beyond this limit it is necessary to use quartz (down to 0.16 μm), sapphire (0.15 μm), CaF_2 (0.12 μm) or LiF (0.105 μm). Generally speaking it is impossible, both for economic and purely technical reasons, to make complete tubes with these materials; they are used for the window only. Since these materials have widely different coefficients of thermal expansion from those of the types of glass used, it has been necessary to develop complex sealing techniques based in each case on the use of a number of materials with an intermediate coefficient of expansion.

All cathodes sensitive in the visible range have good quantum efficiencies in the whole of the ultra-violet [8] and may therefore be used in photomultipliers. There

[7] V. A. Stanley, IEEE Trans. NS-13, No. 3, 63-71, 1966.
[8] L. Dunkelman, Ultraviolet photodetectors, J. quant. Spectr. rad. Transfer 2, 533-544, 1962.

is however a new cathode which is of particular interest for certain applications because of its selectivity. This is the Te-Cs cathode, which has a quantum efficiency of up to 10% at 0.25 μm and has a sensitivity which is virtually zero beyond 0.32 μm. This matches exactly the wavelength range of the solar radiation which is transmitted through the atmosphere. This "solar blind" photocathode is therefore particularly useful for photometric measurements of rays which are not transmitted by the Earth's atmosphere.

Finally, no window transparent to wavelengths less than 0.105 μm exists (incidentally, air itself is opaque to wavelengths of 0.18 μm and below). For these wavelengths we have developed windowless photomultipliers in which the cathode (which is no longer semi-transparent) consists simply of a metal, e.g. nickel, or an insulator (BeO, KCl, etc.) [9]. No material obstacle must be interposed between this cathode and the radiation source.

Dark current

The dark current obviously sets a lower limit to the level of light signal that can be detected, and it is therefore a significant factor in the quality of the tube. To give some idea of magnitudes, the luminous flux equivalent to the dark current in a photomultiplier with a normal S11 cathode is of the order of 10^{-11} lumen.

Investigation of the amplitude distribution of the pulses which form the dark current can give considerable information about the origins of this current, particularly if multichannel pulse-amplitude analysers (kicksorters) are used [10] [11]. The amplitude of a pulse is in fact a function of the point of emission of the electron, or electrons, which initiated the avalanche. This point may be on the cathode, but it may also be on the focusing electrodes or on the dynodes. The interpretation is complicated, however, by the fluctuations in the pulse amplitude.

The most obvious source of dark current is thermionic emission from the cathode, on account of the low thermionic work function Φ_{th} of the layers. (We should recall here that Φ_{th} is lower than Φ_{ph} in the materials which are of interest; see page 269.) Thermionic emission obeys the law:

$$I_{th} = AT^2 \exp\left(-\frac{e\Phi_{th}}{kT}\right).$$

The variation of the dark current with temperature should therefore make it possible to minimize this current by simply reducing the temperature. However, if we measure the number of stray electrons N emitted per unit time and draw curves of $\log N = f(1/T)$ (*fig. 8*), we see that this is not true: although a slope corresponding to an energy nearly equal to Φ_{th} is in

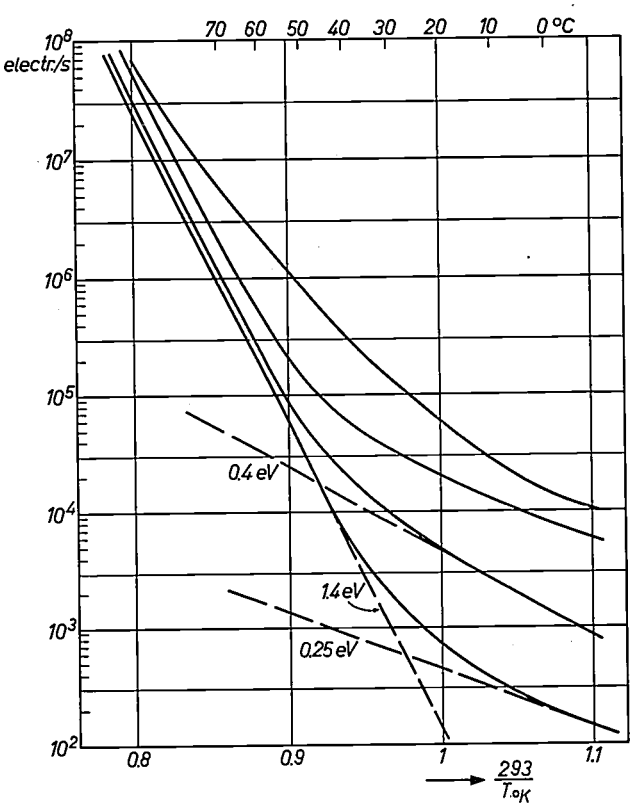


Fig. 8. Variation of the dark current (number of electrons per second) as a function of temperature T for various photomultipliers. The slope at low temperatures corresponds to a much lower work function (shown in eV) than that at high temperatures.

fact found at a fairly high temperature, a much smaller slope is always found at normal or low temperatures. Under normal conditions, therefore, the thermionic emission calculated by extrapolation of the steeply sloping section of the curve is always much smaller than the total current: it only represents a small part of it. It is also found that the dark-pulse spectrum (*fig. 9*) changes its shape with temperature: the relative number of pulses with amplitudes corresponding to electrons emitted from the cathode decreases with decreasing temperature. If the temperature is reduced sufficiently for the thermionic emission to be negligible and the cathode is then disconnected, thus suppressing its emission, it is sometimes possible to observe a further decrease in the number of pulses in the corresponding region of amplitudes. It may therefore be assumed that these pulses were due to photoelectrons initiated by luminescences inside the tube. The number of these pulses is in fact frequently observed to increase rapidly with the high voltage.

To limit this contribution to the dark current, we first concentrated on obtaining the required gain while using the lowest possible high voltage — which amounts to improving the secondary emission — and on determining the optimum number of dynodes. This proved

to be 14 for a gain of 10^8; the interdynode voltages are then of the order of 130 to 140 V.

This choice, however, impairs the speed of response, as will be shown later. It was therefore necessary to find other preventive measures and to identify the cause of the luminescences. We shall mention just a few results of these investigations.

It is difficult to produce electrodes and connecting leads which have no surface irregularities. A deposit of caesium at such irregularities lowers the work function and thus encourages cold emission. Moreover, the material Ag-MgO-Cs, which is a finely grained insulator, is in itself an important source of cold emission. The electrons which are emitted are accelerated by the electric fields throughout the tube and collide either with insulators (mica or ceramic spacers, the glass wall, and the inside or outside surfaces of the dynodes coated with MgO) where they give rise to scintillations, or with metal surfaces, where they produce soft X-rays, which in turn can excite considerable photoelectric emission. All of these radiations excite either the cathode or the dynodes, particularly the first dynodes. It should be noted that these effects have a fairly high "efficiency". The ratio of photoelectric emission to cold emission could be as high as 0.1 if all the photons were able to reach the cathode directly.

We shall now indicate some of the preventive measures used. To prevent the electrons from reaching the wall of the tube, it is sufficient in principle to bring this wall to the most negative potential in the tube, that of the cathode. Certain tubes are therefore given a conducting outer coating (colloidal carbon) held at cathode potential. (Incidentally, it has also been found that the lowest and most stable dark currents are often obtained when the tube is operated with the cathode earthed and the anode at positive potential.) The luminescences must be counteracted in other ways. They often occur in the base of the tube, since the leads carrying all the potentials are close to one another here, and also at the dynodes, between which, at certain points, very high fields exist. To prevent these photons from reaching the cathode, a system of shields is used — in fact the shielding system is formed by the electrodes of the input optics — and black glass is used to insulate the dynodes and the various electrodes [10]. All metal parts are carefully polished and their shapes carefully chosen so that sharp edges (e.g. at the dynodes) giving rise to high electric fields do not occur. Various precautions are taken to reduce the amount of alkali deposited on the surfaces during the activation of the cathode. In cooperation with the glass manufacturers we are carrying out investigations to see if the percentage of potassium contained in the types of glass we use can be reduced. Thermionic emission itself can be reduced by the introduction of the Sb-K-Cs cathode which has a higher work function. The result of all these measures is that in some of the tubes the dark current is less than 1000 electrons per second (at the cathode).

A unique solution consists in the use of the disc-seal technique. This has been adopted, for example, in the XP 1210, which will be discussed presently (*fig. 10*). In this arrangement each dynode is connected to a metal ring, sealed in the envelope, by means of a disc which occupies practically the whole diameter of the tube. The tube is thus subdivided by a large number of screens, and moreover the potentials are distributed uniformly over the length of the tube. With this structure interdynode potentials of the order of 300 V may be applied without greatly increasing the dark current. This is a fairly expensive method, however, and is only considered for special cases (e.g. where extremely fast response is required).

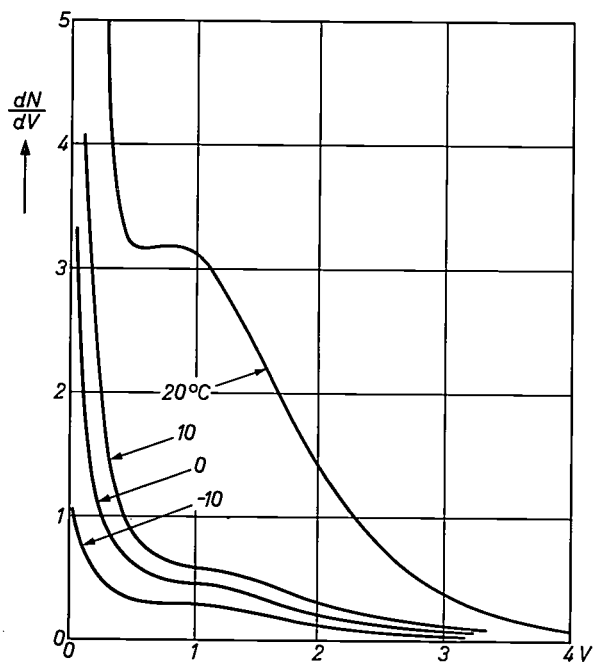


Fig. 9. Pulse-amplitude distribution of the dark current in a photomultiplier, for various temperatures. The probability density dN/dV (in arbitrary units) is shown as a function of pulse amplitude V. (The pulse amplitudes have been normalized to the mean amplitude of the pulses for single photoelectrons, as determined by the measurement of the single-electron spectrum, see p. 272.)

[9] M. Brunet, C. Jehanno, C. Julliot and A. Tarrius, Photomultiplicateurs sans fenêtre—étude, réalisation, Onde électr. **42**, 746-753, 1962.

[10] G. Piétri, Present state of research and new developments at the Laboratoires d'Electronique et de Physique Appliquée (L.E.P.) in the field of photomultiplier tubes, IEEE Trans. **NS-11**, No. 3, 76-92, 1964.

[11] M. Duquesne, I. Tatischeff and N. Recouvreur, Nucl. Instr. Meth. **41**, 13-29, 1966. E. Morilleau and R. Petit (Brive), Photomultiplicateur rapide à faible bruit de fond, Onde électr. **46**, 111-119, 1966.

To conclude this section we shall consider the part played by the residual pressure. By themselves the atoms of the residual gas cannot liberate a single electron. Below 10^{-4} torr the mean free path of the electrons is so long that these atoms have hardly any effect on the multiplication. If the gain is fairly high, however, the electrons, which are very numerous in the final stages, produce ions, and these ions may travel to the cathode and produce secondary electrons there, or they up to *several minutes* after the extinction of an intense illumination.

At the values of gain normally used (10^8), this kind of behaviour is fortunately rare in the photomultipliers now in production.

Amplitude fluctuations

Even without extensive theoretical analysis [12] it will be clear that the fluctuations in amplitude originate

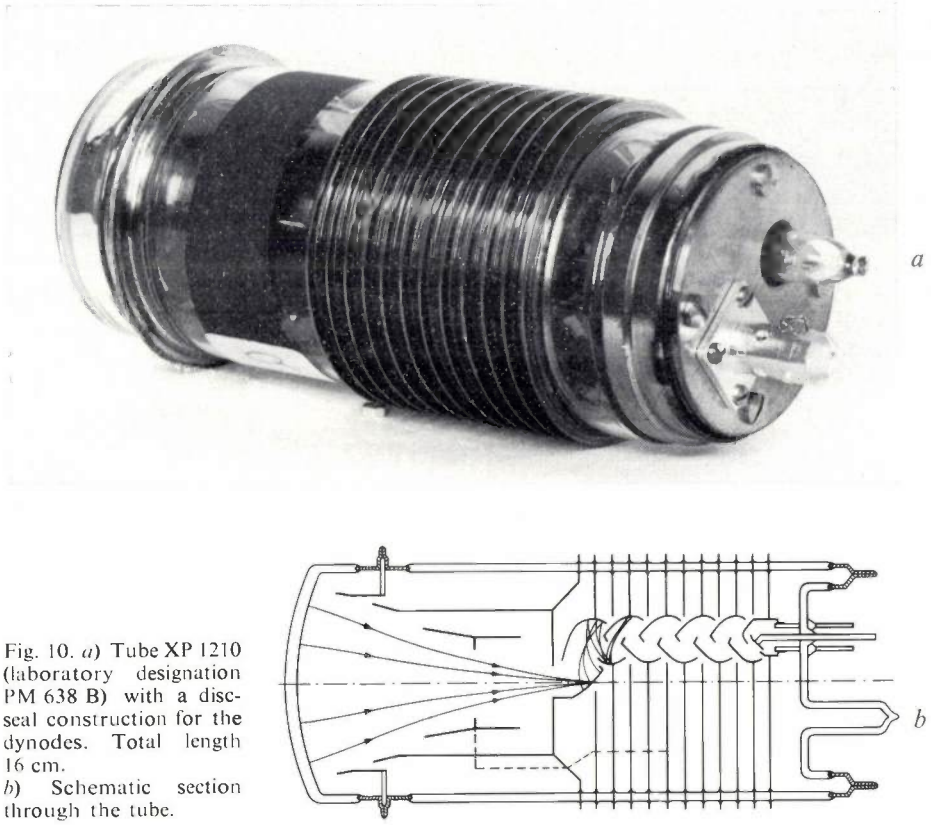


Fig. 10. *a*) Tube XP 1210 (laboratory designation PM 638 B) with a disc-seal construction for the dynodes. Total length 16 cm.
b) Schematic section through the tube.

may recombine, with the emission of light. There is thus an interaction between the output and the input which could introduce instability. In reality, these effects are not instantaneous and thus appear in the form of after-pulses, delayed by a time almost equal to the transit time in the photomultiplier if the interaction is due to optical coupling, or by a much longer time (several microseconds) if the interaction occurs via ions travelling to the cathode. In the extreme cases these after-pulses may be regarded as dark-current pulses since they usually appear when there is no longer any illumination. Such an interpretation is even more tempting in the case of atoms excited at metastable levels: although these atoms are much fewer in number than the ions, they are very much more mobile since they are not attracted by the electrodes of the tube, and are thus able to cause secondary emission from the cathode from the stages where the number of electrons conveying information is lowest; thus, the photocathode K is the most important source, followed by the first, second and third dynodes, $S_{1,2,3}$, as progressively less important sources. The contributions from the following stages are negligible. Every attempt is therefore made to develop cathodes which have the highest possible sensitivity, and to achieve high coefficients of secondary emission, particularly for the first dynode. Normally the second condition is fulfilled by making

[12] E. Breitenberger, Scintillation spectrometer statistics, Progr. nucl. Phys. **4**, 56-94, 1955. G. Piétri, Les photomultiplicateurs, instruments de physique expérimentale, Acta electronica **5**, 7-30, 1961.

[13] M. Brault and C. Gazier, C.R. Acad. Sci. Paris **256**, 1241-1244, 1963. R. F. Tusting, Q. A. Kerns and H. K. Knudsen, IRE Trans. **NS-9**, No. 3, 118-123, 1962. C. Gazier, Contribution à l'étude des propriétés statistiques du gain des photomultiplicateurs, Thèse d'ingénieur du CNAM, 1963.

the potential between K and S_1 two or three times greater than that between the other stages, thus assuring a value of δ which may be greater than 8. A high collection efficiency is also highly desirable [13], but the uniformity of the collection efficiency between K and S_1 and between S_1 and S_2 is equally important. This may be illustrated with an example.

Suppose that the cathode is divided into two equal parts whose sensitivities are in the ratio of 2 : 1 (which is effectively the same as a non-uniformity of 2:1 in the collection efficiency between cathode and S_1). Suppose further that scintillations appear near to the cathode (which therefore each only illuminate a small area) and produce on average $n = 30$ electrons in the most sensitive zone and $n = 15$ in the other, with a probability of $\frac{1}{2}$ of appearing in each zone. The probability distribution obtained $P(n)$ will be the sum of two distributions centred on 30 and on 15 with corresponding widths at half-height of $2.36\sqrt{30} = 13$ and $2.36\sqrt{15} = 9$. If the cathode were uniform and its sensitivity was equal to the average of that of these two zones, there would only be one distribution with the width $2.36\sqrt{22.5} = 11$

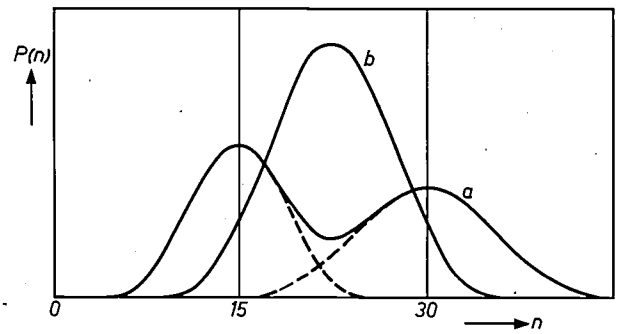


Fig. 11. Illustrating the effect of non-uniformity in the sensitivity of the photocathode on the pulse-height resolution of a photomultiplier.

(*fig. 11*). We have exactly the same situation with the first dynode if the collection $S_1 \rightarrow S_2$ is not uniform over the whole area of S_1 capable of receiving the electrons emitted from the cathode. This area depends on the focusing quality of the input optics and on the initial transverse velocities of the photoelectrons, and it has a typical diameter of a few millimetres in tubes of conventional dimensions (*fig. 12a, b*). Defects in the

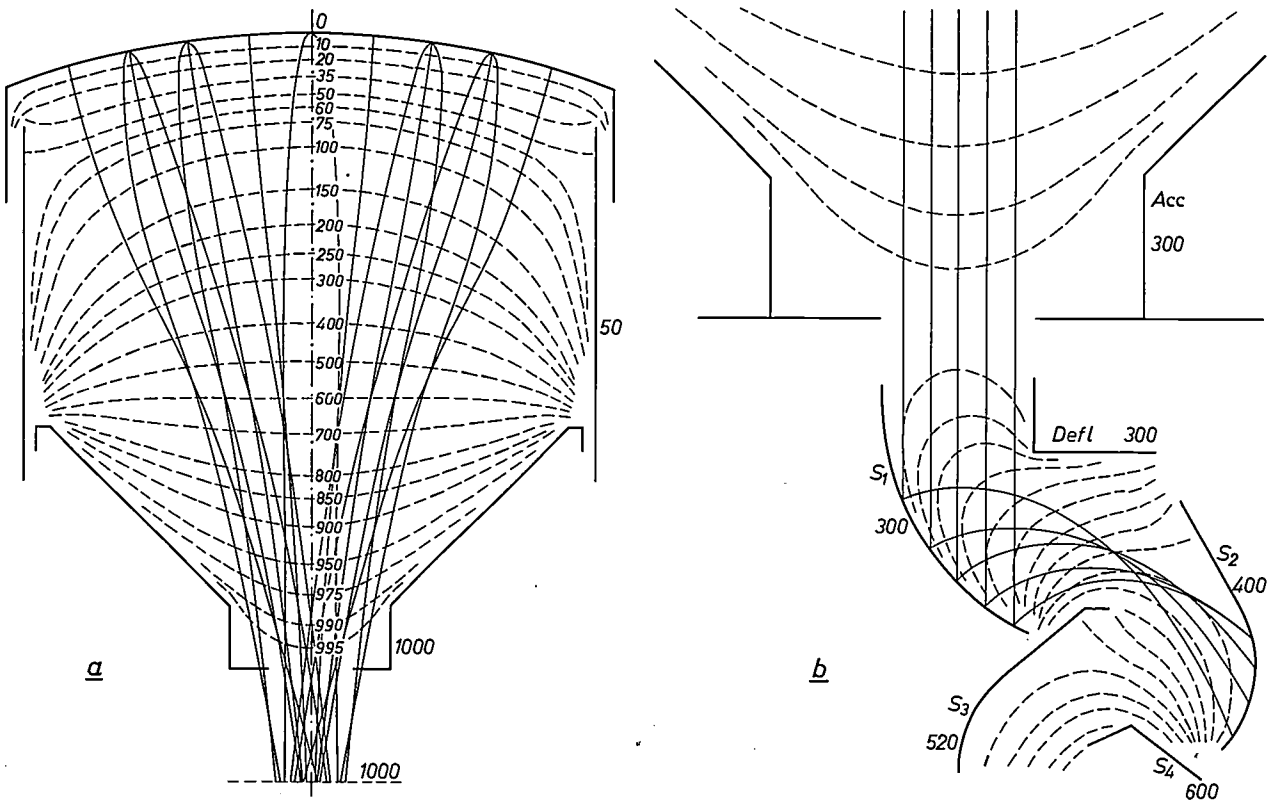


Fig. 12. *a*) Equipotential surfaces of the input optical system in the 56 AVP. Some paths are shown for electrons coming from the cathode, some without initial velocity and others with an initial transverse velocity corresponding to an energy of 0.3 eV. The diagram gives an indication of the area of the first dynode within which photoelectrons can arrive. (The potential values quoted are relative.)

b) Equipotential surfaces of the region of the first dynodes. The vertical lines are possible paths for electrons originating from the cathode. These connect with the paths of secondary electrons (zero initial velocities) which should all reach the second dynode. All the paths will almost inevitably have different lengths since the electron beams never arrive normal to the target surfaces.

uniformity of collection will affect more especially the single-electron spectrum, whose form is in fact determined by the secondary-emission fluctuations [13]. *Fig. 13* shows the effect on the SES of deliberate defocusing by altering the potentials of the input optics in the 56 AVP tube. The effect of the wavelength of the light (this comes into it since it affects the initial velocities) is represented by an analogous family of curves.

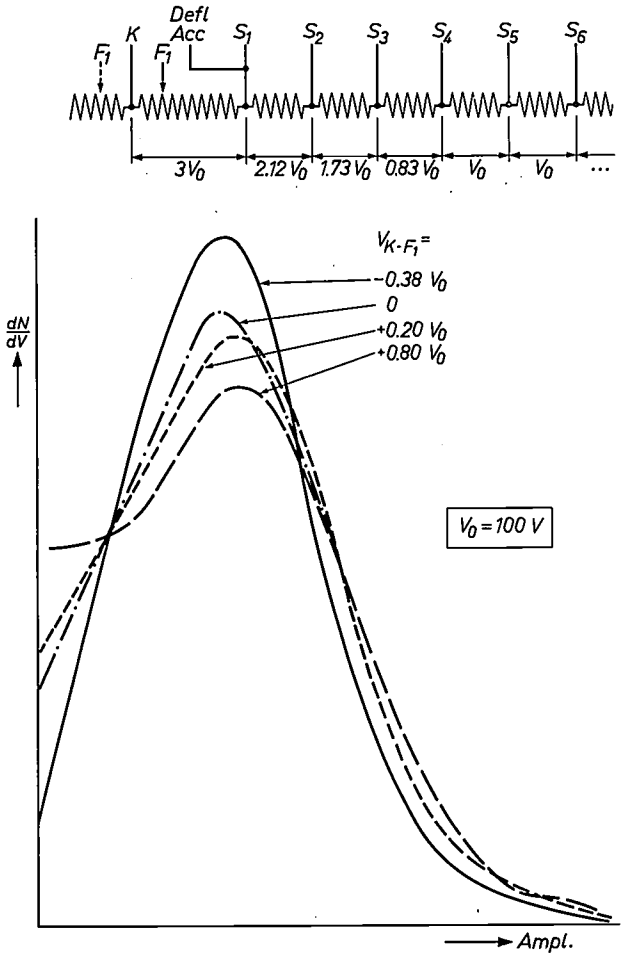


Fig. 13. Effect on the single-electron spectrum of deliberate defocusing by altering the potentials of the input optical system in the 56 AVP. The diagram at the top shows the distribution of the potentials on the different electrodes. Beginning with S_4, the voltage between successive dynodes is equal to V_0. The curves *1* to *4* were obtained for four different values of the voltage between photocathode K and focusing electrode F_1.

The improvement of the collection efficiency, both in value and in uniformity, is a problem in electron ballistics which can be solved by progressive modification of the configuration of the electrodes in an electrolytic tank. It is advisable to avoid the use of accessory focusing electrodes where possible, at least if the user has to adjust their potential himself. Nevertheless, in high-performance tubes (e.g. types 56 AVP and XP 1020) voltage adjustments have been accepted to compensate for the effect of the relative inaccuracies in the geometry and assembly of the electrodes. However, it is preferable to use arrangements in which the auxiliary electrodes can be connected internally to one of the dynodes. For example, the 56 AVP has a triode optical system (cathode, focusing electrode and accelerator) in which only the focusing electrode has a variable potential, the accelerating electrode being connected to S_1. The XP 1020 has an extra electrode connected to S_9. *Fig. 14* gives a comparison of the geometries of these two types, and *fig. 15* shows the variation of their collection efficiencies as a function of wavelength. The collection efficiency of the input optics is defined here as the ratio of the number of pulses with a height greater than a certain threshold value to the number of photoelectrons emitted during the same time. A description of the method of measuring this quantity would be beyond the scope of this article [10].

The shape of the single-electron spectrum and the collection efficiency are of great importance for the detection of single photoelectrons: since it is always necessary to count only the pulses above a certain threshold, it is important that this spectrum should contain only relatively few pulses of small amplitude.

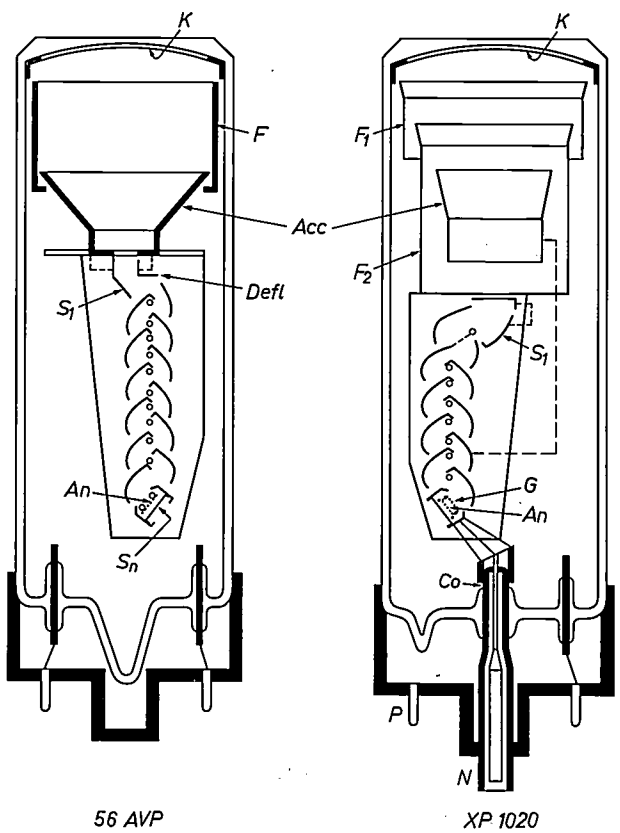


Fig. 14. Section through the 56 AVP and the XP 1020. K photocathode. $F_{(1,2)}$ focusing electrodes. *Acc* accelerating electrode. *Defl* deflector. $S_{(1\ldots n)}$ dynodes. G screen grid. *An* anode. *Co* coaxial feed. N connection to standard coaxial cable. P pins.

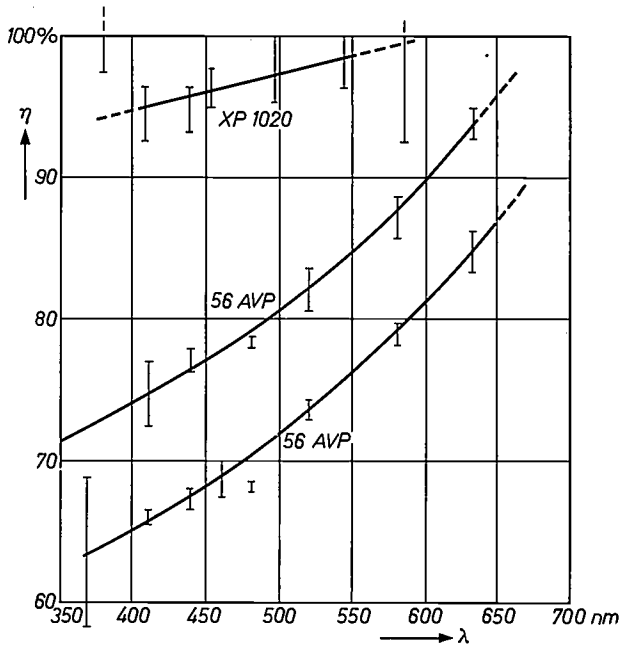


Fig. 15. Variation of collection efficiency η of the 56 AVP (two samples) and the XP 1020 as a function of wavelength.

On the other hand, if the number of photoelectrons in the resolving time is greater than 1, then the fluctuation of the number of electrons predominates. Schottky's equation then applies for the detection of modulated light, the fluctuations of the secondary emission being represented by the correcting factor p in equation (3b).

If we neglect non-uniformity in the secondary-emission coefficients δ_1 and δ of the first and the following dynodes, respectively, the factor p may be written:

$$p = 1 + \frac{\delta}{\delta_1(\delta - 1)}. \quad \ldots\ldots \quad (4)$$

For the typical values $\delta_1 = 7$ and $\delta = 3.5$, the value of p is 1.2, but non-uniformity may increase this value considerably.

The validity of Schottky's equation in the detection of modulated light has been verified experimentally for all cathodes of the usual types between 75 Hz and 100 kHz: deviations from the predicted values were 20 % at the most and the noise was in fact found to be "white". In all cases involving the detection of modulated light the value of the ratio S/N can therefore be calculated previously, provided that the dark current is negligible compared with the d.c. component of the signal. The value I_k of this d.c. component must in fact be introduced in Schottky's equation to find the noise.

[14] G. Piétri, Glimpse on some problems encountered in the study and construction of photomultipliers with fast response, Le Vide **15**, 120-131, 1960.

If m is the depth of modulation, the ratio S/N is given by:

$$\frac{S}{N} = \frac{m\eta_c^2 \bar{I}_k^2}{4ep\Delta f \eta_c \bar{I}_k} = \frac{m\eta_c \bar{I}_k}{4ep\Delta f}.$$

It can be seen that the methods of improving this ratio S/N are always the same: improvement of the collection efficiency of the input optics and the first stages (which increases δ and reduces p).

When modulated light is detected in the presence of a *non*-negligible dark current, Schottky's equation, even when corrected by the substitution of $I_k + I_0$ for I_k, is only valid to a first approximation. In fact, if the dark current is not mainly thermionic in origin, its fluctuations do not obey this law; the noise produced by the dark current in a circuit of bandwidth Δf may depend on frequency, and it may be larger or smaller than the value predicted from Schottky's equation, depending on whether the dark current is due to scintillations (ambient radioactivity) or to ohmic leakage.

Finally, we should note that in the detection of modulated light it is always desirable to derive as much benefit as possible from the amplification of the multiplier by making the gain sufficiently high that the Schottky noise at the anode is much greater than the thermal noise in the load resistance R. This condition may be expressed as:

$$G^2 \gg \frac{2kT}{ep} \frac{1}{RI_k}. \quad \ldots\ldots \quad (5)$$

The proper value of the load R is usually determined by the required bandwidth and the stray capacitances.

Speed of response

We shall now present a more detailed account of the physical effects which determine the speed of response of photomultipliers, and we shall then give a discussion of recent improvements in this field.

1) Effects limiting the speed of response [14].

The effects to be taken into account are the initial velocities of the electrons and the induced currents caused by the movement of the electrons through the tube.

Let W_n be the energy corresponding to the component of the initial velocity of an electron (of mass m) normal to the emissive surface. This component will have the effect of reducing the transit time of the electron with respect to the transit time of an electron emitted at zero initial velocity. It can be shown that this reduction of time is given by:

$$\Delta t = \sqrt{\frac{2m}{e}} \frac{\sqrt{W_n}}{E_0}, \quad \ldots\ldots \quad (6)$$

where E_0 is the electric field at the emissive surface. Since W_n fluctuates from one electron to another with a mean value $\overline{W}_n$, this expression with $W_n = \overline{W}_n$ will give the order of magnitude of the fluctuation in the transit time of the electron. For a given geometry this effect will thus vary in inverse proportion to the applied voltage.

Let us now consider the transverse velocity component (perpendicular to the mean electron path). This component has the effect of spreading out the electron beam over the surface of the first dynode S_1. In practice, the cross-section of the beam at this location is arranged to be no larger than the useful surface of the dynode, that is to say a surface such that the electrons emitted from any point of it will have a reasonable chance of reaching the next dynode. The effect of the transverse velocities may now be seen from fig. 12*a* and *b*: the asymmetry gives a variation in the length of the path with the direction and magnitude of the initial transverse velocity. On the other hand, the electrons emitted from different points of S_1 also take different times to travel from S_1 to S_2, again because of the asymmetry. There is thus a correlation between the spread in the successive stages which makes any accurate calculation of the fluctuation observed at the anode impossible.

The difference in transit time for paths starting from two different points A and B varies with $1/\sqrt{V}$. If these two points are the points of arrival of two electrons emitted from the same point on the preceding surface with two different values of initial transverse velocity, their distance apart will also vary approximately as $1/\sqrt{V}$. If the spread in the transit times of electrons emitted from points A and B varies linearly with the distance AB, we may predict that the final spread will vary approximately as $1/V$.

The difference in the average transit times to the first dynode of electrons starting from the centre and the edge of the cathode (times which can easily be measured and calculated) is referred to more simply as the "centre-to-edge spread".

The induced currents arise when the electrons travelling to the anode arrive in its vicinity, and they last for as long as the movement of these electrons [15]. It is desirable by suitably shielding the anode to try and suppress the effect of electrons not travelling to the anode and to reduce the effect of the "useful" electrons which are travelling towards it. The geometry of S_n in the 56 AVP meets this requirement. However, the electrons moving towards S_n pass through the anode, which takes the form of a grid, and these electrons already induce an unwanted signal. Moreover, the electrons emitted by S_n are not all captured immediately when they arrive at the anode. Some pass through it and oscillate for some time around the wires of the grid, inducing high-frequency currents in the wires. These currents become even larger if their frequency happens to coincide with the resonant frequency associated with the capacitances and inductances of the anode system.

The signal is taken off between S_n and the anode: the connections to these two electrodes form the beginning of the line along which the signal will be transmitted. The impedance of this line is in general very different from that of the standard coaxial cables and there are discontinuities in it which give rise to various perturbations in the shape of the pulse.

Finally, the inductances of the dynode connections, which carry rapidly varying currents, give variations in the dynode potentials, and hence gain variations, which also have the effect of distorting the pulse.

2) Improvement of the pulse response.

All the effects considered above tend in a varying degree to lengthen the pulse. It can be said, however, that in all commercial photomultipliers so far in use the main cause of the longer pulse response is to be found in the initial velocities; the after-pulses observed in the 56 AVP (*fig. 16*) are about the only exception since they are due to oscillations of the electrons around the anode. We have seen that it is difficult to calculate the overall response of a photomultiplier from the response of the individual stages because of the correlation between the stages. Useful results can be obtained, how-

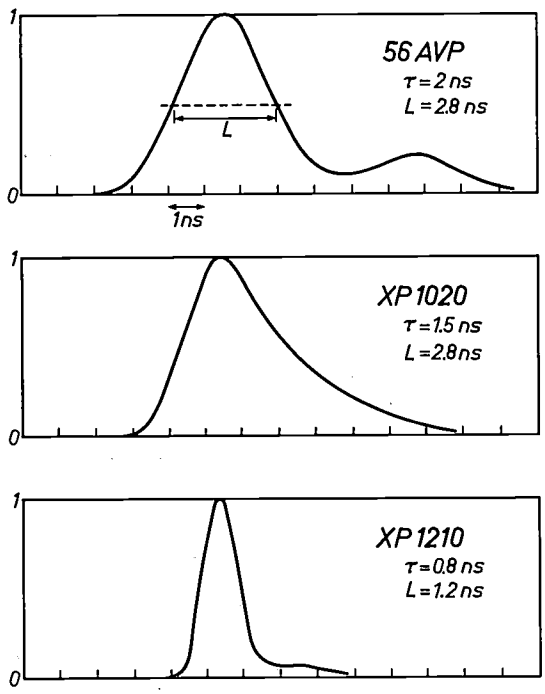


Fig. 16. Pulse response of three fast photomultipliers. In each case the rise time τ is indicated and the width L at half-height of the pulse.

ever, if it is assumed that the responses of all the stages add as the square. Let σ_{KS1}, σ_{S1S2} and $\sigma_{S(n-1)Sn}$ be the standard deviations corresponding to electron transit-time fluctuations in the first, second and nth stages. The responses of the stages beyond the second may usually be taken as equal, so that we may write:

$$\lambda^2 = \sigma^2_{KS1} + \sigma^2_{S1S2} + (n-1)\sigma^2_{S(n-1)Sn}.$$

The width at half-height of the pulse response, L, is related to the standard deviation λ by: $L = 2.36\,\lambda$. In the 150 AVP, for example, the most important term is σ^2_{KS1}, and for practical purposes we have: $L = 2.36\,\sigma_{KS1} \approx 10$ ns. In the development of the 56 AVP, whose speed of response was to be higher, the main effort was therefore concentrated on the input optics. The pulse responses of all the stages have widths of the order of 0.8 ns (except for stage S_1-S_2, where the response is about twice as long). With 14 dynodes, i.e. 15 stages including the stage from S_n to the anode, this results in an overall pulse response of about $0.8 \times \sqrt{15} = 3$ ns. (The contribution from the S_n-anode stage is due to the induced currents.)

A further substantial reduction in the length of the pulse response can only be obtained by simultaneously reducing the spread in *all* the stages. This has been done with the XP 1210, now under development, by increasing the voltages applied between the dynodes by a factor of 2 or 3, and by modifying the input optics and the stage S_1-S_2. This results in a reduced centre-to-edge spread and a 7 times higher electric field at the cathode. (These modifications had in fact already been made for the XP 1020, as we shall see later.)

We have already touched on the difficulties caused by increasing the applied voltages. The anode here is a shielded disc and the signal is collected on a matched transmission line. Finally, the inductance of the dynode connections has been considerably reduced by using the disc construction, whose relatively large capacitance also appreciably improves the dynamic stability of the potentials. The pulse response obtained has a width of 1.1 ns and a rise time of 0.8 ns.

If the full benefit is to be derived from a rise time of the order of a nanosecond, the fast photomultipliers should have a very high gain ($> 10^8$). This can be seen as follows: the electron tube systems which have to be used for processing the time information obtained require pulse levels of the order of 10 to 20 V (3 to 5 V for semiconductor systems) across an impedance of 50-125 ohms. Amplification is hardly possible, since conventional amplifiers of suitable bandwidth have only recently become available and they are still expensive. A burst of say 10 photoelectrons must therefore produce a signal of 10 V across an impedance of 50 ohms, which means that the gain has to be 10^8. To provide this gain, 14 stages are necessary in the 56 AVP, in which the interdynode potentials are of the order of 120 V; however, 10 stages are sufficient in the XP 1210. This helps to improve the speed of the device and moreover the total high voltage does not have to be increased along with the interdynode potentials.

3) Reduction of the spread in the overall transit time [16].

Fast photomultipliers are very often used not only for determining directly the shape of a light pulse but for determining as accurately as possible the instant of appearance of a short flash of light of low intensity, producing very few photoelectrons (sometimes only one). An example is the pulse produced by the Cerenkov effect in a transparent medium by the passage of a high-energy particle. This pulse, which produces 100 electrons at most, lasts about 10^{-11} second, and it is necessary to determine the instant at which it appears with an error of less than 10^{-10} second, since the velocity (and energy) of the particle can be determined from the time the particle takes to cover the distance between two photomultipliers provided with Cerenkov radiators.

The accuracy of such measurements is limited by the spread in the total transit time in the photomultiplier. This spread is of course related to the spread in the different stages, but we shall see that the contributions from the different stages do not have the same relative importance as in the calculation of the pulse response. What we therefore have to do is to examine the spread in the time taken by the avalanche initiated by the emission of a single electron by the cathode to traverse the photomultiplier. The transit time of the electron between K and S_1 varies with the point of emission and the initial velocity, with a standard deviation of σ_{KS1} (see above). If N electrons are emitted at the same time, it may be assumed that the position of the centre of gravity of the resultant pulse fluctuates with a standard deviation of $\sigma_{KS1}/\sqrt{N}$. Similarly, when δ secondary electrons are emitted at the same time by the dynode S_1, the fluctuation of the position of their centre of gravity is $\sigma_{S1S2}/\sqrt{\delta}$ (as a first approximation; the fact that δ fluctuates must also be taken into account). The standard deviation σ_t of the total fluctuation for the case of a single electron is found by summing the squares, taking care to avoid the difficulties which arise because of the correlation between the paths in successive stages. Introducing σ as the standard deviation of the total fluctuation for the case of N electrons emitted simultaneously, we arrive at the relation:

[15] The theory of the operation of high-frequency diodes and triodes will be found in the standard works on electronics.

[16] E. Gatti and V. Svelto, Nucl. Instr. Meth. **4**, 189-201, 1959. E. Gatti and V. Svelto, Nucl. Instr. Meth. **30**, 213-223, 1964.

$$\sigma^2 N = \sigma_t^2 = \sigma_{KS1}^2 + \frac{\sigma_{S1S2}^2}{\delta_1}\left(1+\frac{1}{\delta_1}\right)+\frac{\sigma_{SS}^2}{\delta_1(\delta-1)}\left(1+\frac{1}{\delta}\right), \quad \ldots \quad (7)$$

assuming that only the first stage gives values of σ_{S1S2} and δ_1 different from those of the other stages, σ_{SS} and δ.

It can be seen that the contributions from the input optics and from the stage S_1-S_2 are by far the most important, even though all the σ values are of the same order of magnitude. A substantial reduction of σ_{KS1} and of σ_{S1S2} would thus greatly assist in reducing the spread in transit time while the effect of this on the pulse response would be insignificant. In view of this the input optics and the stage S_1-S_2 in the XP 1020 were redesigned. In the XP 1210 the increase of the operating voltages also resulted in a substantial reduction in σ_{S1S2}. The table below shows the contributions from the various stages, expressed in nanoseconds, for three types of fast photomultiplier. It can be seen that, for the last two types, the dominant term is the one due to the stage S_1-S_2; this is accounted for by the special configuration of this stage. To give a numerical example, the probable error in the measurement of the instant at which a single photoelectron is emitted from the cathode of an XP 1210 is $2.36 \times 0.13 = 0.31$ ns, while the measurement of the instant of arrival at the cathode of a Cerenkov pulse giving 100 photoelectrons is subject to an error of $0.31/\sqrt{100}$ ns, i.e. 3.1×10^{-11} s.

	56 AVP	XP 1020	XP 1210
σ_{KS1}	0.42	0.07	0.07
$\sigma_{S1S2}\sqrt{\frac{1+1/\delta_1}{\delta_1}}$	0.41	0.20	0.10
$\sigma_{SS}\sqrt{\frac{1+1/\delta_1}{\delta_1(\delta-1)}}$	0.06	0.04	0.03
$\sigma/\sqrt{N} = \sigma_t$	0.59	0.22	0.13

As an illustration, we shall describe a particularly useful application of this knowledge of the spread in transit times: the measurement of the time constant of the decay of short scintillations, where the duration of this time constant is less than or equal to the pulse response of the photomultiplier (of the order of 10^{-9} s).

Let us assume that the pulse of light can be repeated as often as required, and that the resolving time is sufficiently short (10^{-10} s for example). We assume that the scintillation causes the emission from the cathode of an average of 100 photoelectrons. If we attenuate this flux by a factor of 1000, each scintillation will then give rise to an average emission of $\bar{n} = 0.1$ photoelectron. This means that, over a large number of measurements, there will be an average emission of 1 photoelectron for every 10 scintillations. To be more exact, there is a certain probability that a scintillation will cause the emission of a burst of more than 1 photoelectron, but by taking a fairly small value for $\bar{n}$ it can be shown that this probability is small compared with the probability of the emission of a single electron. In fact, when Poisson's law is applied, the probability for the emission of n photoelectrons can be written:

$$P_n = \frac{\bar{n}^n}{n!}\exp(-\bar{n}).$$

The probability that a burst of *at least* 1 electron will be emitted, that is to say the probability of a signal appearing at the anode, is:

$$\sum_{n=1}^{\infty} P_n = 1 - P_0 = 1 - \exp(-\bar{n}), \quad \ldots \quad (8)$$

while the probability of obtaining just 1 electron is:

$$P_1 = \bar{n}\exp(-\bar{n}). \quad \ldots\ldots \quad (9)$$

It can easily be verified that if $\bar{n}$ is small compared with unity, the ratio $P_1/(1-P_0)$ of these probabilities is of the order of $1 - \bar{n}/2$. In the numerical example given above ($\bar{n} = 0.1$) this ratio is equal to 0.95, in other words 95% of the pulses recorded at the anode will originate from single electrons.

The shape of the decay curve of the light can now be determined by repeating the pulse many times and sampling the intensity at various instants. The accuracy is then equal to the spread of the transit time in the photomultiplier (i.e. 0.13 ns in the case of the XP 1210). This can be seen as follows. Although every photoelectron received can have been emitted at any arbitrary instant during the duration of the scintillation, the probability that a photoelectron will be emitted between t and $t + \Delta t$ is simply:

$$\frac{f(t)\Delta t}{\int f(t)dt},$$

if $f(t)$ is the curve representing the variation of the luminous intensity with time. In other words, if we determine the arrival times of N single photoelectrons, the number of electrons observed between the moments t and $t + \Delta t$ is equal to:

$$\Delta N = \frac{N f(t)\Delta t}{\int f(t)dt}. \quad \ldots\ldots \quad (10)$$

Thus, except for a constant factor, $f(t)$ is given by $\Delta N/\Delta t$. The speed of the effects that can be investigated is no longer limited by the pulse response of the photomultiplier but by the spread of the transit times in the photomultiplier, since this limits the accuracy with which the instant of emission of a photoelectron can

be measured. It is obvious that a very large number of measurements will be required if an exact curve is to be obtained, particularly since very few of the measurements will correspond to an observable pulse. Incidentally, this procedure can also be usefully applied to the light pulses in measuring the speed (the pulse response) of the photomultiplier: their duration is measured, and if this is not negligibly short compared with the response of the photomultiplier (1.1 ns), the information obtained enables corrections to be applied.

4) Frequency response (modulated light).

The detection of modulated light is an increasingly frequent application of photomultipliers [17]. The function $G(\omega)$ that gives the gain of a photomultiplier as a function of modulation frequency can in principle be calculated from the pulse response $R(t)$, using the well-known relation:

$$G(\omega) = \int_{-\infty}^{\infty} R(t)\, e^{-j\omega t}\, dt. \qquad (11)$$

According to this relation the curve representing the function $G(\omega)$ assumes a more extended form the narrower the function $R(t)$. Thus, the bandwidth of a photomultiplier used as a detector of modulated light benefits directly from the improvements of the pulse response. On page 273, the bandwidth B, i.e. the upper limit of the passband, has already been defined as the value of ω for which $G(\omega)/G(0) = 1/\sqrt{2}$. The following table gives the calculated bandwidth B for the three fast photomultipliers mentioned earlier:

Tube	Passband
56 AVP	65 MHz
XP 1020	100 MHz
XP 1210	300 MHz

We should mention that a *heterodyne method* can be used to detect light modulated at frequencies very much higher than the limit of the multiplier bandwidth. The stage K-S_1 can be made smaller in size, thus giving a bandwidth considerably higher than 1000 MHz. If a suitable control electrode is now introduced into this stage, the electron beam can be modulated (by means of a local oscillator) at a frequency f_1 close to the frequency f_2 at which the light is modulated. This gives a frequency conversion in which the difference-frequency signal is brought within the bandwidth of the photomultiplier. This has been done in our laboratories [18] with an experimental tube PM 994 (*fig. 17*) at a frequency close to 1000 MHz; the transmitted signal was that of a television picture.

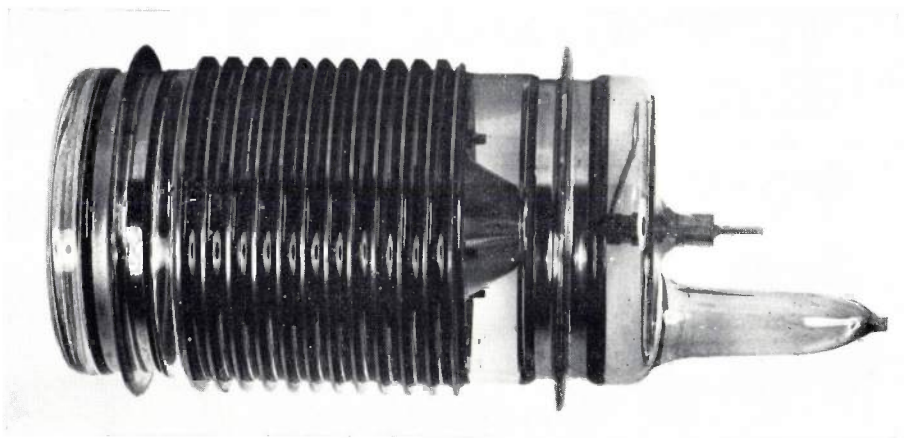


Fig. 17. Experimental photomultiplier type PM 994, used for the heterodyne detection of modulated light. Total length 14 cm.

5) Measurements of speed of response.

Since the phenomena for which we are trying to develop photomultipliers of suitable time resolution are already at the limits of the measurable, it is reasonable to ask how characteristics such as the extremely fast speed of response can be checked. In such procedures it is necessary to use extremely short light pulses which are well-defined in time. The shortest pulses that can be produced are those due to the passage of relativistic charged particles in transparent materials, due to the Cerenkov effect. This procedure, however, is seldom feasible, and the light pulses usually employed are derived from a high-voltage spark discharge between two closely adjacent electrodes [19]. This method has the advantage that absolutely synchronous electric pulses are available which can be used to give a time reference (for triggering an

[17] C. Cernigoi, I. Gabrielli and G. Iernetti, Nucl. Instr. Meth. **6**, 193-200, 1960.

[18] J. Nussli, J. Physique, Suppl. Phys. appl., **26**, 113A-114A, 1965. G. Marie and J. Nussli, C.R. Acad. Sci. Paris **258**, 5179-5182, 1964.

[19] G. Piétri, J. Nussli and M. Brault, Perfectionnements récents apportés aux photomultiplicateurs rapides destinés à la physique nucléaire, Electronique Nucléaire, Compte rendu Colloque int. S.F.E.R., Paris 1963, pp. 793-805. F. Kirsten, UCRL 8706, pp. 1-7, University of California, Febr. 1959. J. H. Malmberg, Rev. sci. Instr. **28**, 1027-1029, 1957.

oscilloscope). However, the duration of these sparks is always of the order of 10^{-9} s, which is short enough for tubes like the 56 AVP but about equal to the response time of the XP 1210. A proper measurement can therefore only be made if one can determine in some other way the exact shape of the light pulse, say by means of an ultra-fast photoemissive cell, like the TVHC 20 or the TVHC 40 (*fig. 18*; these cells have a rise time of 0.2 ns). This presupposes that the light from the spark is sufficient to produce a signal that does not have to be amplified before application to the oscilloscope. Knowing the shape of the light pulse one can then calculate, in certain cases at least, the pulse response of the photomultiplier from the shape of its output pulse.

If, on the other hand, the delay in the triggering of the oscilloscope is known, one can calculate the mean transit time in the photomultiplier. If the light produced by the flash is focused on a point of the cathode, and if this point is moved across the cathode surface, the centre-to-edge spread can be estimated from the change in the transit time, that is to say from the horizontal displacement of the oscilloscope trace. Finally, if the light is attenuated to such an extent that there is only very little probability of two photoelectrons being emitted simultaneously, the jitter observed on the trace enables an estimate to be made of the transit time fluctuations. (Fortunately there are other methods for this purpose that are more exact, e.g. photographic recording of a large number of traces, or time-amplitude converters [20].)

The frequency response may in principle also be determined directly with the aid of Pockels-effect modulators, or by the detection of beats between the different modes of a laser.

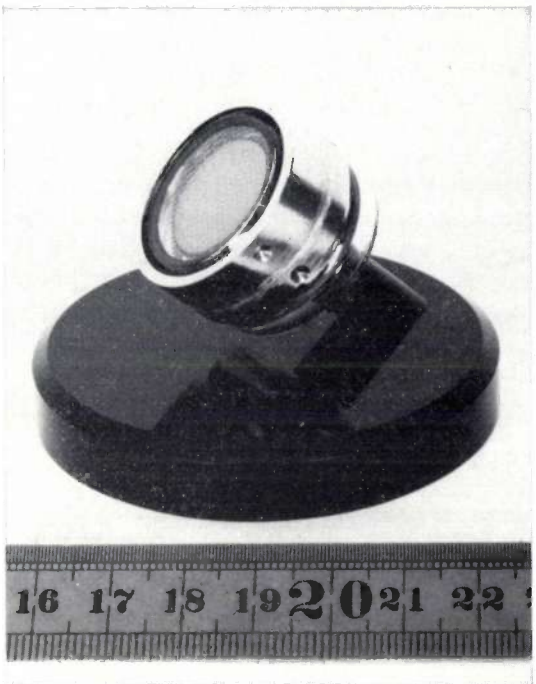


Fig. 18. Ultra-fast photoemissive cell TVHC. It contains a photocathode deposited on a metal disc and a grid-type anode. The distance between these two planar electrodes is about 2 mm, which reduces the transit time of the photoelectrons and hence the duration of the current they induce. This cell is produced by the transfer technique and may be integral with one end of a coaxial cable.

Extension of linearity

The non-linearity which arises because the gain and sensitivity depend on the current introduces so many problems that we cannot deal with them all here. However, we should at least look into what happens when the anode current becomes a significant fraction α of the d.c. current I_p which flows in the resistances connected in parallel with each stage and ensures that the desired distribution of voltage over the stages is maintained.

Let V_0 be the interdynode potential. Between S_n and the anode there is therefore a voltage drop αV_0, but since this stage is only a collection stage, a fall in voltage does not cause a fall in current. However, as the high voltage is stabilized, the excess potential αV_0 is applied to the other stages, which amounts to multiplying the voltage applied to the photomultiplier by a factor of $1 + \alpha/n$. Since the gain varies approximately as V_0^n, the gain of the photomultiplier will be multiplied approximately by

$$(1 + \alpha/n)^n \approx 1 + \alpha, \quad \ldots \quad (12)$$

provided that α is less than, say, 0.5. The photomultiplier is thus seen to be "superlinear", i.e. its response increases faster than the signal.

With light pulses, the situation is different. In this case, capacitors can be arranged at the ends of the dynodes to stabilize the potentials. As a result, during short intervals very large currents can be obtained which, however, are limited by the space charge in the final stages, so that the response will be sublinear. The only way to reduce the space-charge effect is to increase the electric field, particularly on the last dynode. That is why the anode is nearly always in the form of a grid situated at a short distance (1 mm) from the last dynode. By applying to the last dynodes progressively higher voltages (doubling at each stage) up to 500 V between S_n and the anode, the linearity range of the 56 AVP can be extended up to 300 mA. In the XP 1210 the grid-type anode could not be used, as we noted earlier. The plate adopted for the anode in this tube does not permit such an extended linearity region, the limit being reached at 80 mA, or 4 V across 50 ohms. This is nowadays sufficient, owing to the use of semiconductor electronic systems.

[20] A. E. Bjerke, Q. A. Kerns and T. A. Nunamaker, UCRL 9976, University of California, Febr. 1962. G. Present, A. Schwarzschild, I. Spirn and N. Wotherspoon, Fast delayed coincidence technique: the XP 1020 photomultiplier and limits of resolving times due to scintillator characteristics, Nucl. Instr. Meth. **31**, 71-76, 1964. G. Bertolini, V. Mandl, A. Rota and M. Cocchi, Time resolution measurements with XP 1020 photomultipliers, Nucl. Instr. Meth. **42**, 109-117, 1966.

[21] G. Breuze, J.-P. Fertin and R. Petit, Photomultiplicateur rapide à fort courant de sortie et grand domaine de linéarité, Electronique Nucléaire, Compte rendu Colloque int. S.F.E.R., Paris 1963, pp. 113-125.

In certain very special cases even higher limits of linearity are required. Types XP 1140 and 1141 are fast tubes designed for voltages of 2000 V between the last dynodes. The limit of linearity here may be as high as 3 amperes [21].

Study of particular problems

Tubes of unusually large or small dimensions

In order to increase the probability of detecting particles whose paths are not known beforehand, it is necessary to increase the volume and the surface area of the scintillators and hence the surface area of the cathode used in the photomultiplier. On the other hand, in some equipment the space available is very limited.

The best compromise between the various characteristics of a photomultiplier has been found for diameters of 40 to 50 mm. Because of the way in which secondary emission varies with voltage, the interdynode potentials have to be between 100 and 200 V (200 to 500 V between cathode and S_1); any significant change of scale is therefore reflected in a corresponding variation of the electric fields. Scaling up the dimensions linearly moreover gives a linear increase in the magnitude (and spread) of the transit times. Scaling down the dimensions, on the other hand, gives rise to problems of insulation and dark current (field emission). A typical example is that of the 58 AVP (*fig. 19*), which had to be fitted with a cathode of 110 mm — 2.5 times as large as the cathode of the 56 AVP — but still gives the same speed [14]. Linear scaling-up would have meant a proportionate increase in response time: it was therefore necessary to keep the same multiplier system. Now, in the input optics, if the speed of response and the dimensions of the arrival area on the first dynode were to remain the same, the potentials would have to have been increased by a factor of $(2.5)^2$ and the electric fields by a factor of 2.5. This was unacceptable for various reasons, in particular because of field emission. A compromise solution was found by fitting an extra electrode (like the one in the XP 1020) which is set to the potential of one of the later dynodes; the negative side of the compromise is that some deterioration in the centre-to-edge spread and the collection efficiency has to be accepted.

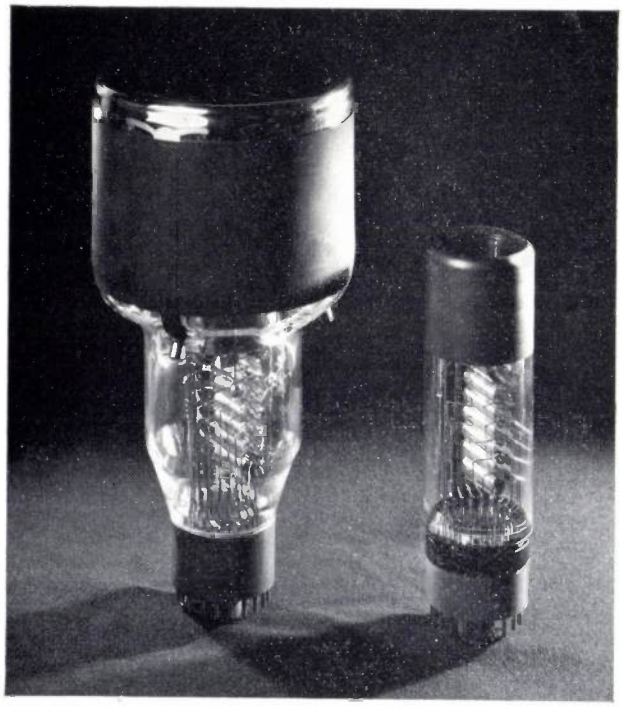


Fig. 19. Fast photomultipliers 56 AVP (on the right) and 58 AVP. Length of tubes 18 and 27 cm, respectively.

Another example is the 152 AVP [2], which has a cathode of only 19 mm diameter. Linear scaling-down would have required impossibly high electric fields and impossibly small dimensions for the dynodes. Reducing the length of the dynodes would have upset the operation of the multiplier as a result of charge effects on the surface of the spacers. A multiplier with special cup-shaped dynodes was therefore developed for this tube (*fig. 20*).

Improvement of mechanical strength

Most photomultipliers are not particularly rugged, because of the nature of their construction. Insulation requirements, for example, may often make it necessary to use insulators which are more fragile than mica, such as glass or ceramics. The XP 1210, however, has been specially constructed to withstand vibrations with accelerations greater than 25 *g*. The 152 AVP can withstand the same conditions, because of its small size (and some slight modifications). A specially ruggedized version of this type (*fig. 21*) has been developed at the works at Brive, in which the tube is shaped like a long rectangular block; all the electrodes are connected to pins at two opposite faces. With this arrangement the tube can easily be mounted between two printed-wiring boards and the whole assembly encapsulated in a potting resin.

Future developments

It is unlikely that there will be any major advances for some time in the speed characteristics of photomultipliers intended for nuclear research, since the performance of present-day tubes is often right at the limits required for present-day experiments. Similarly, we should not expect any further substantial increase in the quantum efficiency of the photocathode, which is already very high in modern photocathodes (in blue light), although we can look forward to a wider application of these cathodes and, in the longer term, to the

development of cathodes that are more sensitive in the 0.7-1 μm region. This would meet more effectively the requirements of photometry in the near infra-red

There are, however, two photomultiplier characteristics which are still capable of considerable improvement: these are the dark current and the statistical fluctuations of the gain.

We have seen that the dark current is chiefly related to the presence of caesium on surfaces where it is not wanted. At the present time a new manufacturing technique is being studied, the "transfer technique", in which the photocathode can be activated *in situ* without contaminating the rest of the tube. The two parts of the tube are mounted separately inside the bell-jar of a vacuum equipment and the cathode is activated from a small evaporator placed right against it. Once the cathode has been formed, the two parts are brought together — still under vacuum — and a seal is effected by an indium pressure-weld. As a result, the spontaneous emission from the dynodes is eliminated and emission from the cathode reduced to its thermionic component alone. These advantages are gained at the expense of a somewhat lower gain (so that the voltage has to be increased).

The statistical fluctuations of the gain are known to be related to the low values of the mean secondary-emission coefficients, which lie between 3 and 8. These fluctuations make it impossible, for example, to identify from their amplitude alone the pulses due to bursts of one, two or three photoelectrons — it would be very useful to distinguish between such bursts in investigations of very weak scintillations like those produced by soft beta rays. The solution is to be found in other processes of electron multiplication. We may mention the following:

a) Channel multiplication [22]. Already used (in the form of single channels) in "windowless photomultipliers" for the detection of low-energy charged particles or soft X-rays or for far UV detection, this technique is now being extended to image intensification, making use of multiple channels.

b) Multiplication by a field effect in thin insulating layers [23].

c) Multiplication by the creation of electron-hole pairs in semiconductor junctions (like those already in use for detecting nuclear radiations). In this case, photoelectrons are accelerated by a voltage of the order of 40 kV and are detected as beta particles in the semiconductor junction. Encouraging results have already been obtained in our laboratories [24]. The relatively low gain (10^4), however, necessitates the use of special low-noise narrow-band amplifiers.

The PM 994 photomultiplier described earlier, which is used for the heterodyne detection of modulated light, could well be the forerunner of a new type of tube in which the information is processed at the photocath-

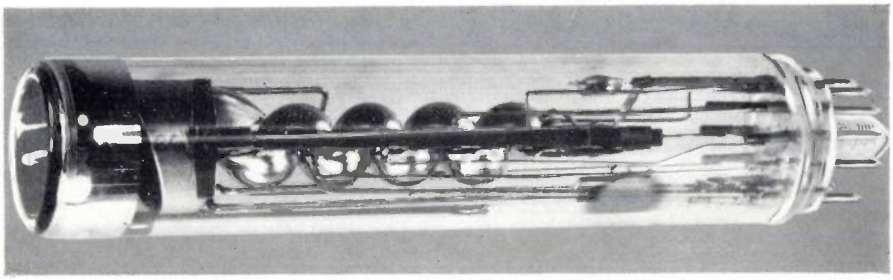


Fig. 20. Small-diameter photomultiplier 152 AVP. Length 10 cm.

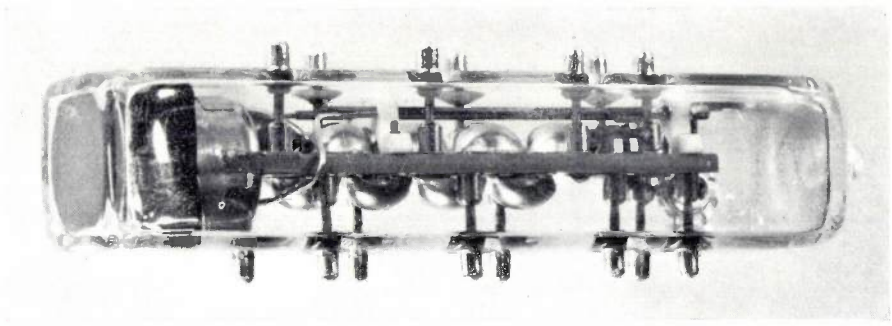


Fig. 21. Ruggedized version of the 152 AVP, type PM 400. Length 9 cm.

ode. This technique, applied to pulses, amounts in fact to a sampling technique. Another possible technique is to use an optical system between cathode and multiplier to form an image of the cathode, and to scan the image by means of a magnetic deflection system: very accurate information about the position of a point of light on the surface of the cathode can be obtained in this way. This method is at present being used for star-tracking equipment in space satellites [25].

The material presented above has demonstrated how the development of photomultipliers is moving towards a greater specialization of types and towards a better adaptability to a whole range of different requirements. This evolution will of course be affected by the development of *solid-state nuclear detectors*, whose performance in the special field of high-resolution spectrometry is and will remain incomparably superior to that of photomultipliers, at any rate for medium-energy radiation [26]. For low-energy radiation the penetration of the radiation into a junction is too small, and for very high energies too little energy is transferred in a junction.

We may conclude that for a long time to come photomultipliers will retain their supremacy as the ideal instruments wherever information has to be derived from effects in which light is produced. In the detection of nuclear radiation, the photomultiplier will continue in demand as an indispensable instrument (with potential for further perfection) in the investigation of extremely rapid phenomena or of very high or low energy radiations.

[22] J. Adams and B. W. Manley, IEEE Trans. **NS-13**, No. 3, 88-99, 1966. J. Adams and B. W. Manley, The channel electron multiplier, a new radiation detector, Philips tech. Rev. **28**, 156-161, 1967 (No. 5/6/7).
[23] H. M. Smith, J. E. Ruedy and G. A. Morton, Performance of a photomultiplier with a porous transmission dynode, IEEE Trans. **NS-13**, No. 3, 77-80, 1966.
[24] P. Chevalier and J. Nussli, Photomultiplicateur à haute résolution utilisant un multiplicateur semi-conducteur, C.R. Acad. Sci. Paris **264B**, 462-465, 1967 (No. 6).
[25] W. D. Atwill, Electronics **33**, No. 40, 88-91, 1960.
[26] W. K. Hofker, Semiconductor detectors for ionizing radiation, Philips tech. Rev. **27**, 323-336, 1966.

Summary. The manufacture of photomultipliers now encompasses an astonishing variety of types of tube, in which attention is paid to the requirements of widely different applications. After a description of the general principles underlying the design of photomultipliers and their characteristics, the article examines the statistical effects that limit the performance of photomultipliers, in relation to the introduction of noise and speed of response. It is shown how the study of these effects gives a better understanding of the relations existing between the various parameters and leads to an overall optimization of performance. Thus, the appropriate use of high electric fields, which give a high speed of response but tend to increase the dark current, has resulted in a pulse response of nearly 1 ns in the fastest photomultipliers, and a transit-time fluctuation of nearly 0.1 ns. The article also gives a description of several of the newest developments such as heterodyne detection of modulated light, and the concluding section indicates the probable trends of future development.

Recent scientific publications

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

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

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

G. A. Allen: The activation energies of chromium, iron and nickel in gallium arsenide.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 593-602, 1968 (No. 5). *M*

G. Baker & D. E. Charlton (Mullard Ltd., Southampton): Low frequency noise in copper doped germanium infrared detectors caused by thermal impedance fluctuations.
Infrared Phys. **8**, 15-24, 1968 (No. 1).

J. R. A. Beale & J. A. G. Slatter: The equivalent circuit of a transistor with a lightly doped collector operating in saturation.
Solid-State Electronics **11**, 241-252, 1968 (No. 2). *M*

P. Beekenkamp & J. M. Stevels: A suggestion for a new nomenclature to describe the structure of vitreous systems and their related crystalline compounds.
Phys. Chem. Glasses **9**, 64-68, 1968 (No. 2). *E*

F. Berz: Large signal a.c. field effect in depletion layers for wide gap semiconductors.
Surface Sci. **10**, 58-75, 1968 (No. 1). *M*

G. Blasse & A. Bril: Investigations on Bi^{3+}-activated phosphors.
J. chem. Phys. **48**, 217-222, 1968 (No. 1). *E*

G. Blasse & A. Bril: The influence of crystal structure on the fluorescence of oxidic niobates and related compounds.
Z. phys. Chemie Neue Folge **57**, 187-202, 1968 (No. 3-6). *E*

G. Blasse & A. Bril: Fluorescence of Eu^{2+}-activated alkaline-earth aluminates.
Philips Res. Repts. **23**, 201-206, 1968 (No. 2). *E*

G. Blasse, W. L. Wanmaker (Philips Lighting Division, Eindhoven), **J. W. ter Vrugt** (*idem*) **& A. Bril:** Fluorescence of Eu^{2+}-activated silicates.
Philips Res. Repts. **23**, 189-200, 1968 (No. 2). *E*

J. Bonnefous: Propriétés directionnelles de quelques substances ferromagnétiques polycristallines. (Thèse, Orsay 1967.)
Acta electronica **11**, 7-112, 123-179, 1968 (Nos. 1, 2). *L*

J. van den Boomgaard: On the system Cr-SiC.
Philips Res. Repts. **23**, 270-280, 1968 (No. 3). *E*

G. A. Bootsma: Gas adsorption on cadmium sulphide.
Surface Sci. **9**, 396-406, 1968 (No. 3). *E*

G. Bosch (Philips Radio, Gramophone and Television Division, Eindhoven): Photoconduction of *n*-type SiC.
Philips Res. Repts. **23**, 139-141, 1968 (No. 2).

A. J. Bosman & C. Crevecoeur: Dipole relaxation losses in CoO doped with Li or Na.
J. Phys. Chem. Solids **29**, 109-113, 1968 (No. 1). *E*

H. Bouma & J. J. Andriessen (Institute for Perception Research, Eindhoven): Perceived orientation of isolated line segments.
Vision Res. **8**, 493-507, 1968.

G.-A. Boutry & M. Jatteau: Sur la mesure des températures de surface par télévision dans l'infrarouge.
C.R. Acad. Sci. Paris **266B**, 214-217, 1968 (No. 4). *L*

S. E. Bradshaw (Associated Semiconductor Manufacturers Ltd., Wembley, England): The effects of gas pressure and velocity on epitaxial silicon deposition by the hydrogen reduction of chlorosilanes.
Int. J. Electronics **23**, 381-391, 1967 (No. 4).

S. D. Brotherton (Associated Semiconductor Manufacturers Ltd., Wembley, England): Dependence of MOS transistor threshold voltage on substrate resistivity.
Solid-State Electronics **10**, 611-616, 1967 (No. 6).

S. D. Brotherton & J. B. Phillips (Associated Semiconductor Manufacturers Ltd., Wembley, England): The control of threshold voltage of complementary pair m.o.s. transistors.
IEE Conf. Publn. 30: Integrated circuits, p. 126-131, 1967.

K. H. J. Buschow & A. S. van der Goot: Intermetallic compounds in the system samarium-cobalt.
J. less-common Met. **14**, 323-328, 1968 (No. 3). *E*

H. J. Butterweck: A theorem about lossless reciprocal three-ports.
IEEE Trans. **CT-15**, 74-76, 1968 (No. 1). *E*

P. J. Buysman & W. Luiten: The relationship between the structure and the strength of ceramic shell moulds, &: Permeability of ceramic shell moulds.
Proc. 12th Conf. European Investment Casters' Federation, Eindhoven 1967, paper No. XI, 12 & 14 pp. *E*

J. R. Chamberlain, A. C. Everitt & J. W. Orton: Optical absorption intensities and quantum counter action of Er^{3+} in yttrium gallium garnet.
J. Physics C (Proc. Phys. Soc.), ser. 2, **1**, 157-164, 1968 (No. 1). *M*

F. J. du Chatenier: Space-charge-limited photocurrent in vapour-deposited layers of red lead monoxide.
Philips Res. Repts. **23**, 142-150, 1968 (No. 2). *E*

R. W. Cooper, W. A. Crossley, J. L. Page & R. F. Pearson: Faraday rotation in YIG and TbIG.
J. appl. Phys. **39**, 565-567, 1968 (No. 2, Part I). *M*

H. J. van Daal: Polar optical-mode scattering of electrons in SnO_2.
Solid State Comm. **6**, 5-9, 1968 (No. 1). *E*

M. B. Das (Associated Semiconductor Manufacturers Ltd., Wembley, England): Switching characteristics of m.o.s. and junction-gate field-effect transistors.
Proc. IEE **114**, 1223-1230, 1967 (No. 9).

M. B. Das (Associated Semiconductor Manufacturers Ltd., Wembley, England): Dependence of the characteristics of MOS-transistors on the substrate resistivity.
Solid-State Electronics **11**, 305-322, 1968 (No. 3).

J. B. Davies & B. J. Goldsmith: An analysis of general mode propagation and the pulse-shortening phenomenon in electron linear accelerators.
Philips Res. Repts. **23**, 207-232, 1968 (No. 2). *M*

A. M. van Diepen (Natuurkundig Laboratorium der Universiteit van Amsterdam), **H. W. de Wijn** (*idem*) **& K. H. J. Buschow:** Nuclear magnetic resonance in $DyAl_3$, $HoAl_3$, and α-$ErAl_3$.
Physics Letters **26A**, 340-341, 1968 (No. 8). *E*

C. Z. van Doorn: Nature of donor-acceptor pairs in cadmium sulphide.
J. Phys. Chem. Solids **29**, 599-608, 1968 (No. 4). *E*

J. J. Engelsman, J. Knaape & J. Visser: Volumetric determination of the U/O ratio in uranium oxides.
Talanta **15**, 171-176, 1968 (No. 2). *E*

U. Enz & H. van der Heide: Domain-wall mobility in Si-doped YIG.
J. appl. Phys. **39**, 435-437, 1968 (No. 2, Part I). *E*

I. Flinn: Surface properties of *n*-type gallium arsenide.
Surface Sci. **10**, 32-57, 1968 (No. 1). *M*

B. Frank, E. Roeder & S. Scholz: Verdichtung und Formgebung von Glaspulvern.
Ber. Dtsch. Keram. Ges. **45**, 231-233, 1968 (No. 5). *A*

J. G. C. de Gast: Berekening en constructie van hydrostatische lagers.
Polytechn. T. Werktuigbouw **23**, 141-150, 187-196, 1968 (Nos. 4, 5). *E*

J. G. C. de Gast: Een nieuwe variabele voorrestrictie voor hydrostatische dubbelfilm-lagers en zijn toepassing in precisiegereedschapswerktuigen.
Ingenieur **80**, O 51-62, 1968 (No. 19). *E*

C. A. A. J. Greebe & W. F. Druyvesteyn: Some considerations on Alfvén-wave resonances in non-viscous liquid metals.
Philips Res. Repts. **23**, 121-130, 1968 (No. 2). *E*

C. A. A. J. Greebe, W. F. Druyvesteyn & A. J. Smets: Alfvén wave resonances in liquid Na.
Physics Letters **26A**, 337-338, 1968 (No. 8). *E*

J. A. Greefkes & F. de Jager: Continuous delta modulation.
Philips Res. Repts. **23**, 233-246, 1968 (No. 2). *E*

G. J. van Gurp: Flux-transport noise in type-II superconductors.
Phys. Rev. **166**, 436-446, 1968 (No. 2). *E*

E. F. de Haan & K. R. U. Weimer: The beam-indexing colour television display tube.
Roy. Telev. Soc. J. **11**, 278-282, 1967/68 (No. 12). *E*

C. Haas: Spin-disorder scattering and magnetoresistance of magnetic semiconductors.
Phys. Rev. **168**, 531-538, 1968 (No. 2). *E*

S. H. Hagen: Surface-barrier diodes on silicon carbide.
J. appl. Phys. **39**, 1458-1461, 1968 (No. 3). *E*

E. E. Havinga: Band structure and superconductivity of non-transition metals.
Physics Letters **26A**, 244-246, 1968 (No. 6). *E*

H. F. van Heek: Current saturation and negative resistance in evaporated CdSe layers.
Physics Letters **26A**, 175-176, 1968 (No. 5). *E*

H. F. van Heek: Hall mobility of electrons in the space-charge layer of thin film CdSe transistors.
Solid-State Electronics **11**, 459-467, 1968 (No. 4). *E*

J. C. M. Henning: Magnetic, optical and e.p.r. properties of Cs_3ZnCl_5 with dilute Co^{2+} ions.
Z. angew. Physik **24**, 281-287, 1968 (No. 5). *E*

J. C. M. Henning: Semi-empirical calculation of ^{14}N hyperfine coupling factors in heterocyclic anions.
Chem. Phys. Letters **1**, 678-680, 1968 (No. 13). *E*

A. R. Hill: Uses of fine focused ion beams with high current density.
Nature **218**, 202-203, 1968 (No. 5137). *M*

A. H. Hoekstra (Philips Lighting Division, Eindhoven): The chemistry and luminescence of antimony-containing calcium chlorapatite.
Thesis, Eindhoven 1967.

D. van Houwelingen & J. Volger: The superconducting-dynamo properties and applications.
Philips Res. Repts. **23**, 249-269, 1968 (No. 3). *E*

G. H. Jonker: The application of combined conductivity and Seebeck-effect plots for the analysis of semiconductor properties.
Philips Res. Repts. **23**, 131-138, 1968 (No. 2). *E*

J. A. Kerr (Associated Semiconductor Manufacturers Ltd., Wembley, England) **& L. N. Large** (SERL, Baldock, England): Semiconductor devices made by ion implantation.
Proc. int. Conf. on applications of ion beams to semiconductor technology, Grenoble 1967, p. 601-618.

F. M. Klaassen: Thermal noise in space-charge-limited solid-state diodes.
Solid-State Electronics **11**, 377-378, 1968 (No. 3). *E*

H. Klotz & B. Schmidl: Verfahrenstechnik zur Rückgewinnung von Indiumverbindungen aus Abgasen in der Lampenindustrie.
Verfahrenstechnik **2**, 25-27, 1968 (No. 1). *A*

H. G. Kock, D. de Nobel, M. T. Vlaardingerbroek & P. J. de Waard: Continuous-wave planar avalanche diode with restricted depletion layer.
Proc. IEEE **56**, 105, 1968 (No. 1). *E*

J. van Laar, H. J. Endeman (Laboratorium voor Kristalchemie der Rijks-Universiteit, Utrecht) **& J. M. Bijvoet** (*idem*): Remarks on the relation between microscopic and macroscopic crystal optics.
Acta cryst. **A 24**, 52-56, 1968 (No. 1). *E*

J. G. M. de Lau: High-frequency and microwave properties of hot-pressed fine-grained ferrites.
Proc. Brit. Ceramic Soc. **10**, 275-284, 1968. *E*

D. B. Lee (Associated Semiconductor Manufacturers Ltd., Wembley, England): Diffusion into silicon from an arsenic-doped oxide.
Solid-State Electronics **10**, 623-624, 1967 (No. 6).

W. Lems & Th. Holtwijk: Magnetization reversal of electrodeposited FeNi films.
J. Physique **29**, Colloque C 2, 140-143, 1968. *E*

P. R. Locher: Nuclear magnetic resonance of copper and cobalt in metallic conducting spinels.
Z. angew. Physik **24**, 277-280, 1968 (No. 5). *E*

J. Loeckx: Syntactische analyse en programmeertalen.
Informatie **9**, 277-282, 1967 (No. 12). *B*

F. K. Lotgering: Mixed crystals between binary sulphides or selenides with spinel structure.
J. Phys. Chem. Solids **29**, 699-709, 1968 (No. 4). *E*

F. K. Lotgering & R. P. van Stapele: Magnetic properties and electrical conduction of copper-containing sulfo- and selenospinels.
J. appl. Phys. **39**, 417-423, 1968 (No. 2, Part I). *E*

P. Massini & G. Voorn: Optical and photochemical properties of chlorophyll *a* solubilized in aqueous solutions of surfactants.
Biochim. biophys. Acta **153**, 589-601, 1968 (No. 3). *E*

G. Meijer: Rapid growth inhibition of gherkin hypocotyls in blue light.
Acta bot. neerl. **17**, 9-14, 1968 (No. 1). *E*

R. Memming & G. Neumann: On the relationship betwee surface states and -radicals at the germanium-electrolyte interface.
Surface Sci. **10**, 1-20, 1968 (No. 1). *H*

R. F. Mitchell: Some new materials for ultrasonic transducers.
Ultrasonics **6**, 112-116, 1968 (No. 2). *M*

R. F. Mitchell & M. Redwood (Dept. of Electrical & Electronic Engng., Queen Mary College, University of London): Frequency response of a distributed piezoelectric source of sound.
Electronics Letters **4**, 107-109, 1968 (No. 6). *M*

F. D. Morten & R. E. J. King (Mullard Ltd., Southampton): Multi-element infrared detectors for high information rate systems.
Infrared Phys. **8**, 9-14, 1968 (No. 1).

B. J. Mulder: Symmetry of the fluorescence and absorption spectra of anthracene crystals.
J. Phys. Chem. Solids **29**, 182-184, 1968 (No. 1). *E*

A. G. Nassibian (Associated Semiconductor Manufacturers Ltd., Wembley, England): Effect of diffused oxygen and gold on surface properties of oxidized silicon.
Solid-State Electronics **10**, 879-890, 1967 (No. 9).

A. G. Nassibian (Associated Semiconductor Manufacturers Ltd., Wembley, England): Effect of gold on surface properties and leakage current of MOS transistors.
Solid-State Electronics **10**, 891-896, 1967 (No. 9).

M. Noé & G. Vandewiele (Université Libre de Bruxelles): The calculation of distributions of Kolmogorov-Smirnov type statistics including a table of significance points for a particular case.
Ann. math. Statistics **39**, 233-241, 1968 (No. 1). *B*

J. M. Noothoven van Goor: Densities of charge carriers in bismuth.
Physics Letters **26A**, 490-491, 1968 (No. 10). *E*

J. M. Noothoven van Goor & H. M. G. J. Trum: Distribution coefficient and donor valency of tellurium in bismuth.
J. Phys. Chem. Solids **29**, 341-345, 1968 (No. 2). *E*

H. Ocken & J. H. N. van Vucht: Phase equilibria and superconductivity in the molybdenum-platinum system.
J. less-common Met. **15**, 193-199, 1968 (No. 2). *E*

C. van Opdorp: Evaluation of doping profiles from capacitance measurements.
Solid-State Electronics **11**, 397-406, 1968 (No. 4). *E*

A. E. Pannenborg: Decision making in research policy at the industrial level.
Ciba Foundation and Science of Science Foundation Symp. on decision making in national science policy, 1968, Churchill, London, p. 99-103. *E*

R. F. Pearson & A. D. Annis: Anisotropy of Fe^{3+} ions in yttrium iron garnet.
J. appl. Phys. **39**, 1338-1339, 1968 (No. 2, Part II). *M*

J. B. H. Peek: The measurement of correlation functions in correlators using "shift-invariant independent" functions.
Thesis, Eindhoven 1967. *E*

L. G. Pittaway, J. Smith & B. W. Nicholls: Laser induced electron emission using a pulsed argon laser.
Physics Letters **26A**, 300-301, 1968 (No. 7). *M*

T. Poorter (Philips Electronic Components and Materials Division, Eindhoven): The relation between reverse-secondary-breakdown behaviour and the properties of the epitaxial layer of planar epitaxial power transistors.
Philips Res. Repts. **23**, 281-308, 1968 (No. 3).

A. Rabenau, H. Rau & G. Rosenstein: Über die Bleisulfidhalogenide $Pb_5S_2J_6$ und $Pb_7S_2Br_{10}$.
Naturwiss. **55**, 82, 1968 (No. 2). *A*

J. E. Ralph & M. G. Townsend: Near-infrared fluorescence and absorption spectra of Co^{2+} and Ni^{2+} in MgO.
J. chem. Phys. **48**, 149-154, 1968 (No. 1). *E, M*

P. Reijnen: Phase equilibria in the system MgO-FeO-Fe_2O_3.
Philips Res. Repts. **23**, 151-188, 1968 (No. 2). *E*

E. Roeder & J. Hornstra: Grain growth during hot-pressing of tantalum carbide.
J. Amer. Ceramic Soc. **51**, 224-225, 1968 (No. 4). *A, E*

D. A. Schreuder (Philips Lighting Division, Eindhoven): Road tunnels in the Netherlands.
Light and Lighting **60**, 370-377, 1967 (No. 12).

L. A. Æ. Sluyterman: The rate-limiting reaction in papain action as derived from the reaction of the enzyme with chloroacetic acid.
Biochim. biophys. Acta **151**, 178-187, 1968 (No. 1). *E*

M. J. Sparnaay: Physisorption on heterogeneous surfaces.
Surface Sci. **9**, 100-118, 1968 (No. 1). *E*

I. Teramoto & A. M. J. G. van Run: The existence region and the magnetic and electrical properties of MnSb.
J. Phys. Chem. Solids **29**, 347-355, 1968 (No. 2). *E*

D. R. Tilley: Influence of planar defects on superconducting surface nucleation fields.
J. Physics C (Proc. Phys. Soc.), ser. 2, **1**, 293-298, 1968 (No. 2). *M*

M. G. Townsend: Visible charge transfer band in blue sapphire.
Solid State Comm. **6**, 81-83, 1968 (No. 2). *M*

M. G. Townsend & W. A. Crossley: Magnetic susceptibility of rare-earth compounds with the pyrochlore structure.
J. Phys. Chem. Solids **29**, 593-598, 1968 (No. 4). *E, M*

B. Tuck: Conduction band density of states in impure GaAs.
J. Phys. Chem. Solids **29**, 615-622, 1968 (No. 4). *M*

A. A. Turnbull & G. B. Evans: Photoemission from GaAs-Cs-O.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 155-160, 1968 (No. 2). *M*

W. A. J. J. Velge & K. H. J. Buschow: Magnetic and crystallographic properties of some rare earth cobalt compounds with $CaZn_5$ structure.
J. appl. Phys. **39**, 1717-1720, 1968 (No. 3). *E*

M. L. Verheijke: Efficiencies of a coincidence spectrometer for positron annihilation radiation. Addendum.
Nucl. Instr. Meth. **58**, 347-348, 1968 (No. 2). *E*

Q. H. F. Vrehen: Interband magneto-optical absorption in gallium arsenide.
J. Phys. Chem. Solids **29**, 129-141, 1968 (No. 1). *E*

H. W. Werner & H. A. M. de Grefte: Ein Massenspektrometer zur Untersuchung Dünner Schichten.
Vakuum-Technik **17**, 37-41, 1968 (No. 2). *E*

J. S. C. Wessels: Isolation and properties of two digitonin-soluble pigment-protein complexes from spinach.
Biochim. biophys. Acta **153**, 497-500, 1968 (No. 2). *E*

J. S. van Wieringen: Note of Mössbauer fraction in powders of small particles.
Physics Letters **26A**, 370-371, 1968 (No. 8). *E*

W. J. Witteman & H. W. Werner: The effect of water vapour and hydrogen on the gas composition of a sealed-off CO_2 laser.
Physics Letters **26A**, 454-455, 1968 (No. 10). *E*

P. Wodon: Quelques aspects du calcul non numérique sur ordinateurs.
Rev. MBLE **10**, 59-67, 1967 (No. 2). *B*

G. Zanmarchi & C. Haas: Magnon drag at optical frequencies and the infrared spectrum of MnTe.
J. appl. Phys. **39**, 596-597, 1968 (No. 2, Part I). *E*

Contents of Philips Telecommunication Review **27**, No. 4, 1968:

J. Larcher & J. Noordanus: Frequency synthesizers for radio equipment (p. 149-166).

Contents of Electronic Applications **28**, No. 1, 1968:

B. J. M. Overgoor: Impedance matching network for capacitor microphones (p. 1-5).
J. Tuil: Intermodulation in aerial amplifiers (p. 6-21).
A. Cense: Recent developments in circuits and transistors for television receivers, VI. Clamping circuit for colour difference signals (p. 22-28).
H. Schmidt: Thermal feedback in integrated circuits (p. 29-39).

Contents of Mullard Technical Communications **10**, No. 92, 1968:

J. Lavallee: Voltage regulation of lower-powered alternators (p. 34-39).
J. Merrett: Voltage transient and $d\dot{V}/dt$ suppression in thyristor bridges (p. 40-51).
B. J. M. Overgoor: Junction field effect transistors: their structure and operation (p. 52-56).
D. J. G. Janssen: Circuit logic with silicon controlled switches (p. 57-64).
D. Humphries & J. Lavallee: Thyristor control of fan cooling and heating systems (p. 65-72).

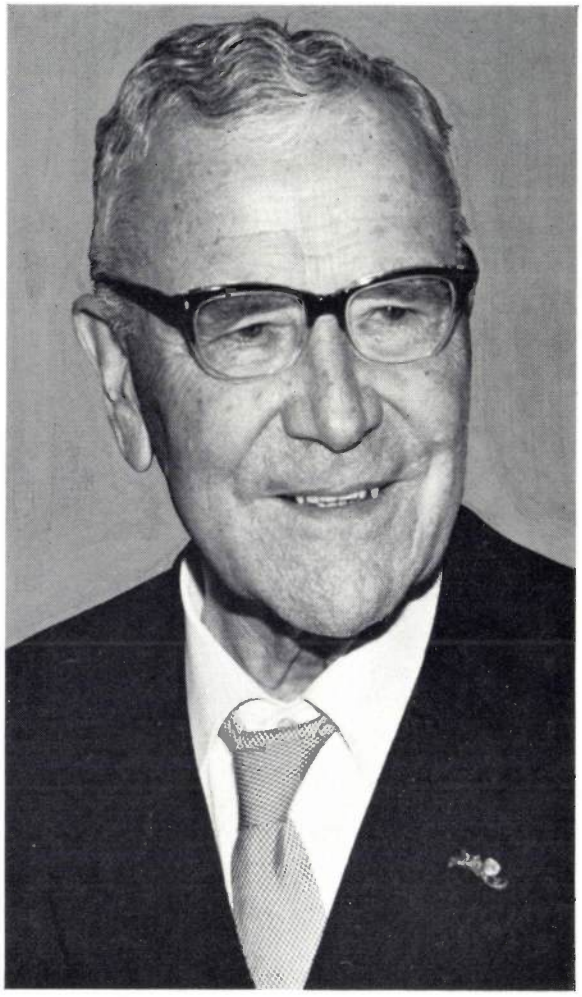
Contents of Valvo Berichte **14**, No. 1, 1968:

U. J. Pittack: Hochleistungs-Impulsklystrons in Linearbeschleunigern (p. 1-25).
E. Ginsberg: Entwurf monolithischer integrierter Schaltungen. Schaltungselemente der integrierten Technik, Grundregeln für die Ausführung integrierter Schaltungen (p. 26-36).

PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 10

In Memoriam Prof. G. Holst



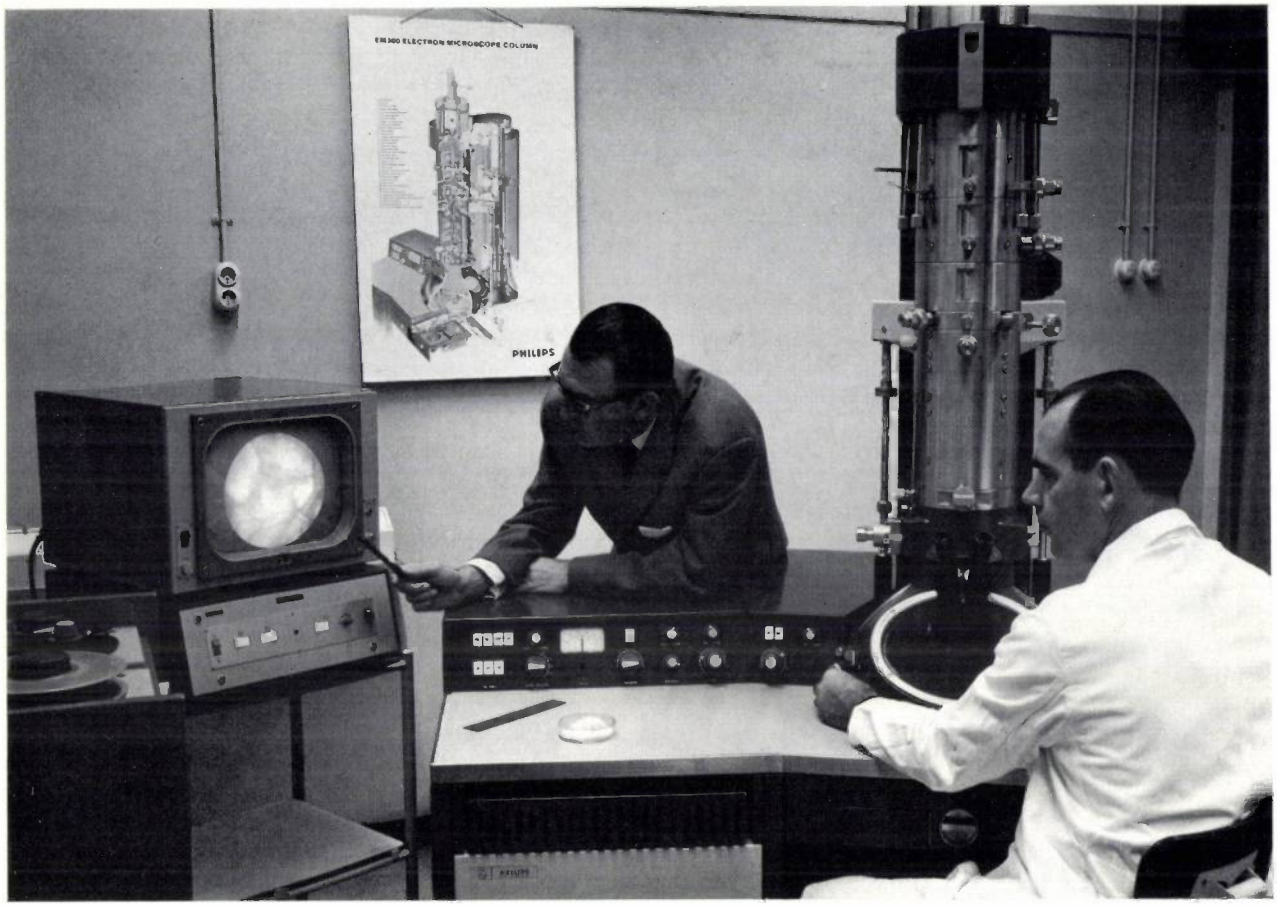
G. Holst, founder of the Philips Research Laboratories, died on 11th October at the age of 82.

Gilles Holst was born on 20th March 1886 at Haarlem. From 1904 to 1908 he studied at the Technical University of Zürich, where he received his doctorate in 1914. From 1910 until the end of 1913 he was an assistant of Kamerlingh Onnes at Leyden. On 1st January 1914 Holst joined our Company as a physicist, and became the founder and director of our Research Laboratories, to which he gave inspiring leadership until he retired from the position of director in 1946. After his retirement he remained with the company for a further ten years as a member of the Supervisory Board. As Chairman of the Board of Governors of Delft Technical University and as chairman of a number of official committees he did a great deal in these later years for higher technical education in the Netherlands.

In this journal, which he himself founded, Holst and the history of the "Natuurkundig Laboratorium" have several times been the subject of attention in the course of the years. Now, with his going, we become only too aware that our pages have given an inadequate idea of the great and beneficial influence which he came to exercise on the development of science and technology, on our Company, on our Laboratories and on each of us personally.

We who knew Holst, and who have endeavoured to continue his work, will always remember him with gratitude and respect. May a younger generation realize that they too are pursuing a tradition which Holst established.

H. B. G. Casimir



Electron microscopy with television display system

The electron image in an electron microscope contains much more information than the eye can pick out from the image which appears on the fluorescent screen. This is particularly so at very high magnifications, since the ability of the eye to perceive detail or contrast is poor at the low level of brightness (luminance) which the image then has. If the image is not viewed on the screen but recorded directly on a photographic emulsion, the result is much better because of the integrating character of the photographic process and the better resolving power of the emulsion, whose grain is much finer than that of the fluorescent layer of the screen. For accurate focusing, however, it is still necessary to view the image on the screen. Because of the limitations in doing this, the microscopist has usually no choice but to take a whole series of photomicrographs with systematically varied focus, in the hope that the picture he wants will be included.

The viewing can be fundamentally improved by converting the electron image into a television signal: this can then be displayed on a monitor with a much greater brightness than the original screen image, and with variable contrast.

This method was first tried out by Haine and Einstein in 1960 [1]. They replaced the fluorescent screen of the microscope by an amorphous selenium film whose conductivity could be locally varied by electron bombardment. The television signal was obtained by scanning the selenium film with a second, very narrow electron beam, as in the familiar types of TV camera tube. This procedure was beset by many practical difficulties, one of which was that the selenium film soon became contaminated.

The idea of converting the electron image into a television signal has now been successfully put into practice by the use of a special "Plumbicon" [*] camera tube. The face plate of this tube has been replaced by a fibre-optics disc. A fluorescent layer (*Fl*) mounted directly on top of this disc converts the electron image of the microscope into an optical image, whose light is trans-

mitted by the fibre optics to the photoconducting layer of the "Plumbicon" tube. This arrangement has considerable advantages over the symmetrical two-lens optical system often used in such cases. It collects about 3 times as much light, the arrangement is much more compact and no aberrations are introduced, while the reproduction of contrast is very little inferior to that of lenses. The diameter of the fibres is only 7 μm, i.e. four times smaller in diameter than the scanning spot of the "Plumbicon" tube, which means that the fibre-optics window has very little effect on the resolving power.

Some practical details are shown in *fig. 1*. The fluorescent layer *Fl* is covered with a very thin aluminium film *A*, which is transparent to the electrons of the microscope. Apart from providing electrical conduction for charge removal, this film has two other functions: it intercepts light from outside (such as the room lighting) and it approximately doubles the efficiency of the fluorescent layer by reflecting to the fibre optics the light emitted in the wrong direction. (To simplify manufacture the layer *Fl* and the film *A* are in fact applied to a separate fibre-optics disc which is later brought into optical contact with the fibre-optics disc of the "Plumbicon" tube.) The face of the camera tube is inserted inside the vacuum system through a vacuum seal arranged in such a way that the electron beam strikes the camera tube when the ordinary fluorescent screen has been swung aside [2].

The favourable noise performance of the "Plumbicon" tube plays an essential part in this system: it is possible to work with a very low illumination of the photoconducting layer, without the bright picture on the monitor being unduly affected by noise from the television system [3].

The very bright monitor picture makes it a simple matter to select the required part of the electron image and to focus it sharply. Measurements of image details are also easily carried out. Since the image can be viewed in a normally lighted room by several persons at the same time, without previous dark-adaptation, the method is also very suitable for instructional and teaching purposes (see title photograph).

The increase in sensitivity can also be used in just the opposite way: for reducing the intensity of the electron beam incident on the specimen. For taking a

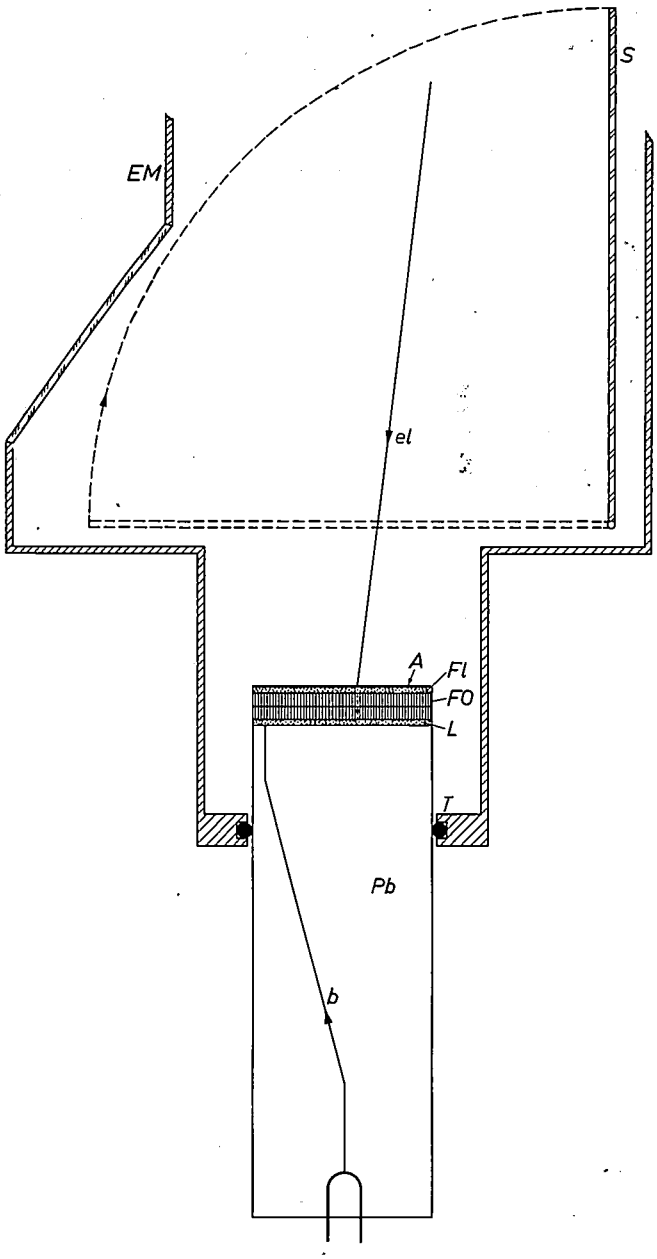


Fig. 1. The fluorescent layer *Fl* converts the electron image into an optical image, which the fibre optics *FO* transmit to the light-sensitive electrode *L* in the "Plumbicon" tube *Pb*. *A* aluminium layer. *el* imaging electron beam in the electron microscope. *b* scanning beam of the "Plumbicon" tube. *T* vacuum-tight feed through the wall of the electron microscope *EM*. *S* is the ordinary fluorescent screen, here swung out of the beam.

[*] Registered Trade Mark for television camera tubes.

[1] M. E. Haine and P. A. Einstein, Proc. Eur. Reg. Conf. on electron microscopy, Delft 1960, Vol. I, p. 97.

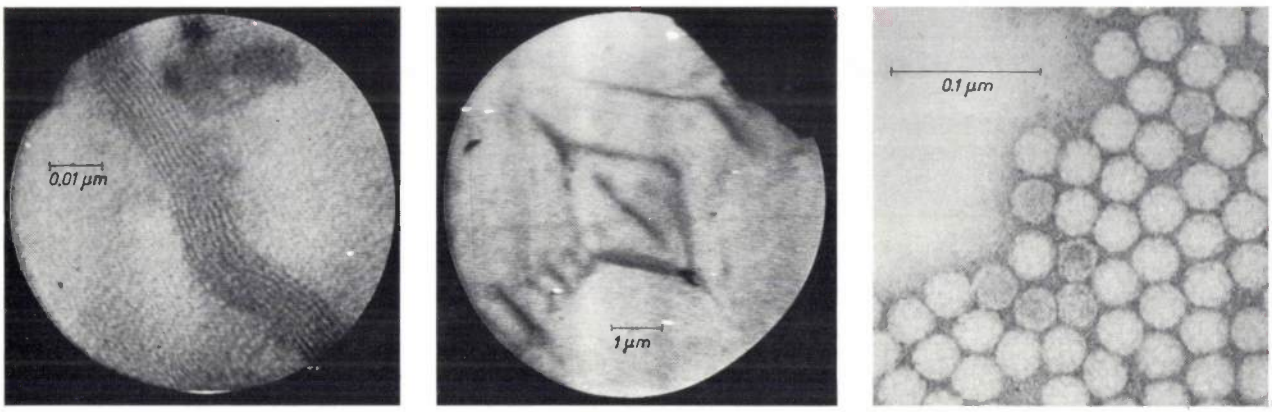
[2] This is one of the design features of the Philips EM 300 electron microscope. *(An article on this instrument will shortly be published in this journal. Editor.)* A recent version includes a further step: connected between the screen *Fl* and the "Plumbicon" tube there is an image intensifier, again coupled to both by means of fibre optics. This gives a 10 to 20 times higher sensitivity, although at the expense of some resolving power. For more information see: H. F. Premsela, A versatile t.v. image display system for electron microscopy using an image intensifier and a "Plumbicon" pick-up tube with fibre-optics window, and W. Kühl, Information transfer with electron microscope t.v. systems, both articles in Bulletin No. EM30, published by the Scientific and Analytical Equipment Department of the Philips Industrial Equipment Division (PIT).

[3] A. G. van Doorn, Philips tech. Rev. **27**, 1, 1966. The dark current of the "Plumbicon" tube is very low, even compared with the signal current at low brightness. This means that the statistical variations of the dark current are of no significance as a source of noise.

photograph of the monitor screen, for example, using the same exposure time, the beam intensity may be several hundred times lower than for a normal electron micrograph. This is important when studying specimens that are sensitive to electron bombardment, for instance those easily affected by heat or damaged in some other way [2].

Electronic image conversion like that used in a television display system also offers a number of quite new ways of extracting information from the electron microscope image. One that immediately comes to mind is the combination with a video tape recorder; such an arrangement is indispensable for studying what is happening in a process, and if a goniometer specimen-holder is used, the three-dimensional structure of the object can be made visible by slowly rotating the specimen while the image is recorded. Various forms of image processing also become possible, such as image addition and subtraction for the suppression of unwanted structures, and contour enhancement. The read-out of individual lines of the monitor image for displaying an intensity profile, and digital recording of the two-dimensional intensity distribution of the image for processing in a computer are yet further possibilities.

Because of the more comfortable working conditions and the other advantages we have mentioned, the use of a television system in electron microscopy is becoming increasingly popular. We should point out, however, that the method is not only convenient, it is also a fundamental advance in observation at the highest magnifications. These bring such small object details within our range that it is even possible to make out the effect of the potential field due to individual atoms of the specimen on the electron beam [4].



a *b* *c*

Fig. 2. Electron micrographs with Philips electron microscope EM 200 equipped with the television system described.
a) Monitor-screen micrograph of a platina phthalocyanine crystal. Because of slight defocusing the $(20\bar{1})$ lattice planes of crystals which are favourably situated on the substrate are visible as a result of interference between the transmitted beam and beams which have undergone Bragg reflections at these planes. The distance between the planes is 12 Å. The phase structure of the substrate (amorphous carbon) is also unavoidably visible here. Voltage 80 kV.
b) Monitor-screen micrograph of polyethylene crystals; focused image. Here the increase in sensitivity is used for reducing the beam current in order to avoid overheating the specimen, and for this reason a relatively small magnification is used. Current density on the fluorescent screen 8×10^{-14} A/cm². (For a direct photomicrograph with an exposure of 1 second a current density of 5×10^{-11} A/cm² would have been necessary.) Voltage 60 kV.
c) Turnip-yellow mosaic virus (TYMV), negative staining. Direct photomicrograph taken after adjusting the stigmator and exact focusing with the aid of the monitor image. The "golf-ball" structure of the virus is clearly visible because of the absence of substrate phase structure (see *a*), which would appear after even the slightest defocusing. Voltage 80 kV. (The specimen was provided by Drs. J. E. Mellema, Laboratorium voor Structuurchemie, University of Groningen.)

The interference patterns (phase structures) thus caused have periodicities of the order of 0.5 nm (5 Å) and become visible when the microscope is not focused exactly on the object plane. This may be useful for displaying certain details that cannot be made visible by any other means. On the other hand, the substrate of amorphous carbon on which the specimens are usually mounted may also exhibit phase structure in these circumstances, sometimes making it difficult to

interpret the image. To achieve the desired effect, or avoid the undesired one, it is necessary to work with an accurately determined degree of defocusing. However, this is possible only if the stigmator of the microscope has been extremely accurately adjusted (giving compensation of the slight astigmatism caused by the small but inevitable departure from circular symmetry of the magnetic field of the objective lens) and if the setting for exact focus has been found to start with. But finding this "parafocal" region with the very dim images obtained on the ordinary fluorescent screen at the highest magnifications is very time-consuming, and in practice virtually impossible. The brighter image obtained with the aid of our arrangement using the "Plumbicon" tube has completely altered this situation. The method of adjustment by observing the phase structures now presents no particular difficulties for the experienced microscopist, and with the EM 300 electron microscope it will enable a resolution of 3.5 to 4 Å to be obtained in a single micrograph (*fig. 2*), with no need for making a systematic series of micrographs. On the monitor screen the resolution is still a good 5 to 10 Å.

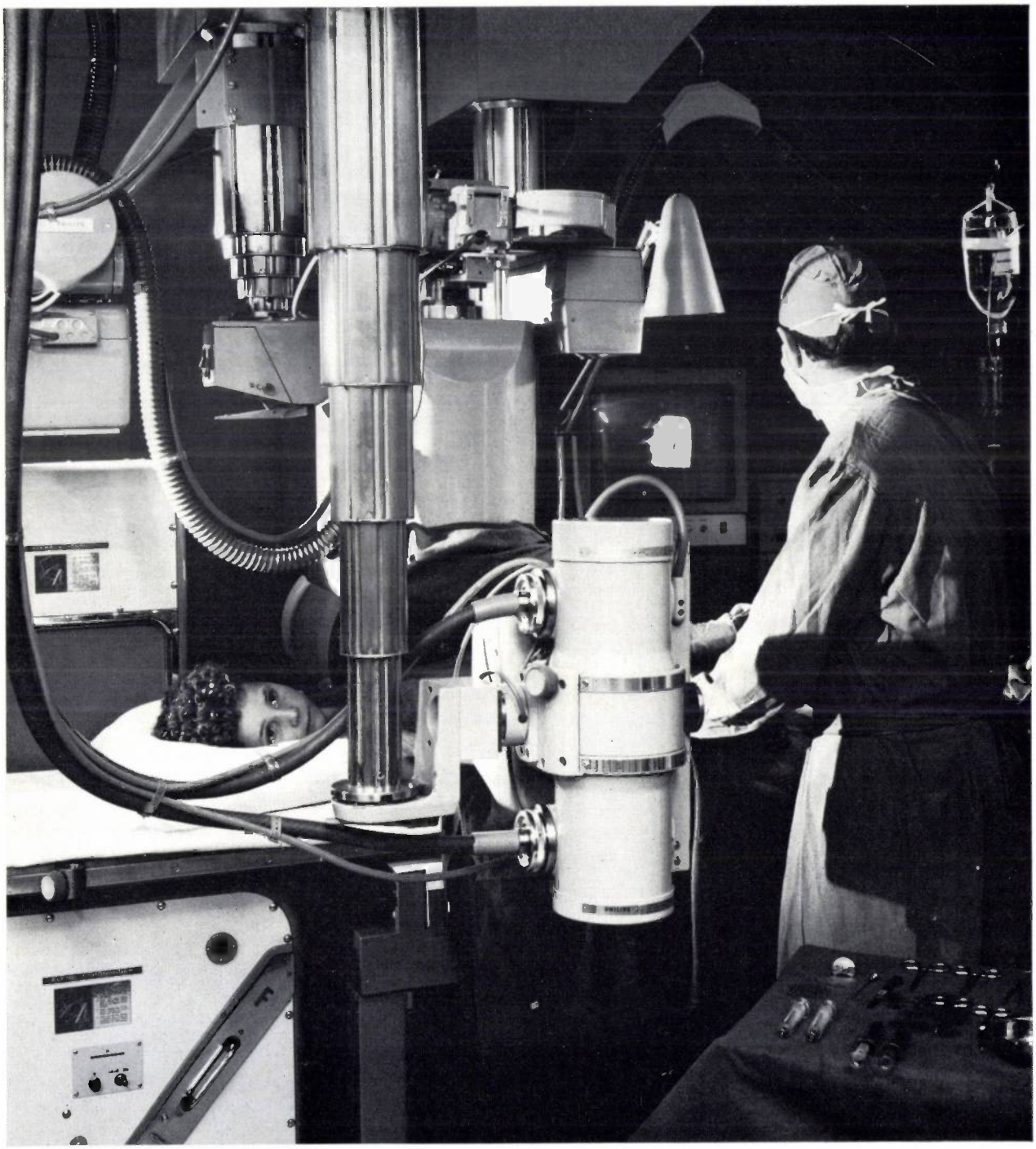
In the further development of electron microscopes the phase structures which can now be viewed so easily will be of great significance, since the monitor image enables all manner of imperfections in the performance of the electron microscope to be traced and diagnosed.

P. H. Broerse
A. C. van Dorsten
H. F. Premsela

[4] A. C. van Dorsten and H. F. Premsela, Sixth int. Congress for electron microscopy, Kyoto 1966, Vol. 1, p. 21.

Ir. A. C. van Dorsten (scientific adviser), Drs. P. H. Broerse and H. F. Premsela are with Philips Research Laboratories, Eindhoven.

Installation for cardiovascular examination



The photograph shows a heart catheterization unit. A thin flexible tube, the catheter, is inserted into a peripheral blood vessel (in the upper arm, for example) and pushed in until the end enters the heart. Lateral and frontal fluoroscopy — each with image intensifiers, which can be seen here behind and above the patient, and which are connected to a closed-circuit television system — enable the surgeon to follow the progress of the catheter. For good visibility the catheter is made of a material that strongly absorbs X-rays. The catheter is used to give measurements of blood pressure, blood oxygen, etc., in the heart ventricles. The blood circulation rate can also be determined by injecting a dye into the heart through the catheter; similarly a contrast medium can be injected to show up anomalies such as abnormal connections between blood vessels. In addition, radiographs can be taken (with the unit on the left), and the X-ray images on the image intensifiers can be photographed, filmed (up to 200 frames per second) or recorded on tape with the aid of the television circuit.

The design of road lighting for given luminance and uniformity

J. Vermeulen and B. Knudsen

When a road-lighting installation is planned, a compromise has to be found between the quality of the installation and the expenditure required. This compromise can be arrived at today from the recommendations which various national and international bodies have published, specifying values for minimum permissible average luminance and maximum permissible non-uniformity. The design procedure described here makes use of a direct and practical way of expressing the relation between these quantities and the multiplicity of parameters for a road-lighting system — which include the complicated reflection characteristics of the road surface. The outcome is a new procedure giving a rapid determination of the maximum permissible spacing of the lamps and of the most suitable type of lantern fitting.

The first essential for good observation on a lighted road is an adequate level of luminance, i.e. the luminous flux which the road reflects per unit surface and per unit solid angle in the direction of the observer. Another important factor is that there should be no great difference in luminance; it should be as uniform as possible. When producing a design for a road lighting installation it is therefore necessary to determine the luminance distribution of the design. The luminance of a given part of the road surface depends on its illumination level and reflection characteristics. Generally speaking, a road surface is neither a perfect (specular) reflector nor a perfect diffuser, but something in between [1]. Owing to the complicated directional dependence of the reflection characteristics, the luminance distribution of a lighting installation cannot be expressed in a simple mathematical expression; it can only be derived from a large number of measurements carried out point by point. Until recently this meant that it was not possible to produce several alternative designs at short notice — which is why international road-lighting recommendations have only quoted illumination levels.

Nowadays, however, computers can speed up the processing of the data obtained from point-by-point measurements, and measuring instruments have been developed which can determine in a single measurement the average luminance of a given road surface in given lighting conditions. The large quantity of data made available in this way has made it possible to classify road surfaces in existing lighting installations, and this has made it simpler to produce new designs. Another result is that international recommendations can now specify average luminance levels and values of uniformity, expressed for example as the ratio of the average and minimum luminance [2].

With a given lantern at a given height above the road surface, the *spacing of the light sources* is the variable that determines the uniformity of the luminance and its average level. The usual practice is for comparative designs to be made for a number of lanterns and to choose the type which permits the greatest spacing between light sources. This is because the cost of a lighting installation is largely determined by the number of lamp posts required. Progressive standardization of the posts sets a limit to the choice

Ir. J. Vermeulen is with, and B. Knudsen was formerly with, the Philips Lighting Laboratory, Eindhoven.

[1] Here it is very important whether the road surface is dry, damp or wet; see J. Bergmans, Lichtreflectie door wegdekken, Thesis, Delft 1938; J. B. de Boer and A. Oostrijck, Reflection properties of dry and wet road surfaces and a simple method for their measurement, Philips Res. Repts. 9, 209-224, 1954; W. Kebschull, Die Reflexion trockner und feuchter Strassenbeläge, Thesis, Berlin 1968.

Since the reflection characteristics vary with the wetness of the road surface in a way which is difficult to predict, the design procedures described in this article relate only to *dry* surfaces.

[2] International recommendations for the lighting of public thoroughfares, Publication C.I.E. (Commission Internationale de l'Eclairage) No. 12 (E 3.3.1), Paris 1965.

of the *height* of the light source, and in some cases the mounting height is determined by local conditions. As a rule the procedure in drawing up a design is to determine first of all the greatest spacing between light sources which is still compatible with the uniformity required. A calculation is then made of the light-source spacing that gives the required average luminance level; the smaller of the two distances is the maximum spacing that may be used in the design.

These two stages of the design procedure will now be dealt with, in the same order. First we must describe the geometrical configuration of the light source, the road surface and the observer.

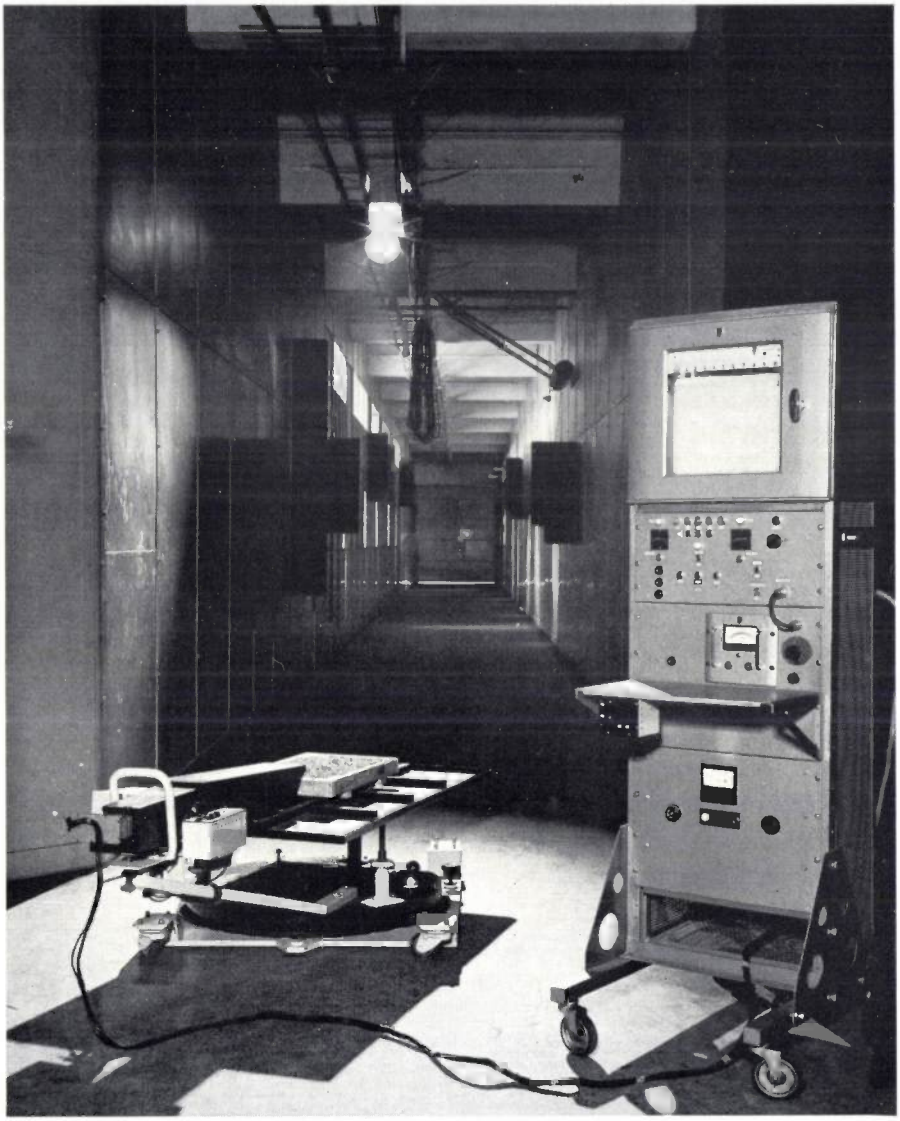


Fig. 2. Arrangement for measuring luminance factors on a sample of road surface in the laboratory. The sample is placed on a turntable to which a luminance meter is attached. The rotation of the table corresponds to the variation of the angle δ (see fig. 1). The light source, which gives a narrow beam of light, travels along a rail 35 metres long at a height of 2.9 metres. The beam of light remains automatically directed upon the sample of road surface. In this way the angle γ is varied, while the luminous intensity I in this arrangement is independent of γ and δ. The measuring device records the product $q \cos^3 \gamma$.

Configuration of light source, road surface and observer

Fig. 1 illustrates a road lighting installation. A light source S (which may be treated as a point source) is at a height h above the road surface. The point P of the road surface illuminated by S is seen by an

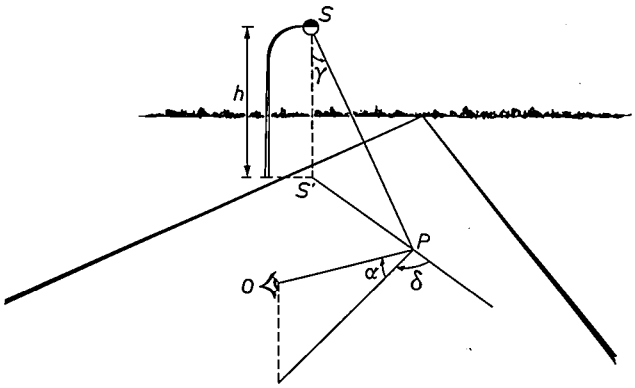


Fig. 1. Configuration of a road lighting installation. S light source, P a point of the road surface. O observer.

observer O. For a given position of the observer the configuration is completely described by the angles α, γ and δ indicated in the figure. The observed luminance L_p at P is proportional to the illumination E of the horizontal plane at P:

$$L_p = qE. \quad \ldots\ldots\ldots \quad (1)$$

The proportionality constant q is called the *luminance factor*. In general the luminous intensity I of the lantern is different in different directions: $I = I(\gamma,\delta)$. The illumination E at P is given by:

$$E = \frac{I(\gamma,\delta)\cos\gamma}{\overline{SP}^2} = \frac{I(\gamma,\delta)}{h^2}\cos^3\gamma.$$

The luminance observed at P is then:

$$L_p = q\,\frac{I(\gamma,\delta)}{h^2}\cos^3\gamma. \quad \ldots\ldots \quad (2)$$

In this expression q is a function of α, γ and δ which depends on the nature of the road surface. In practice the dependence on α is found to be very small as long as α is a small angle, as it is for a road user at a considerable distance from point P. If, with future designs in mind, a measurement is required of the distribution of q over the road surface for an existing installation, it is sufficient to take a constant value of 1° for α; this is the angle with the road surface at which a car driver observes point P from a distance of about 90 metres. To determine q as a function of γ and δ it is then necessary to take measurements for a large number of points on the road surface, and for this purpose the luminous intensity $I(\gamma,\delta)$ of the light source must be known. A laboratory arrangement for carrying out and recording these measurements on road surface samples indoors is shown in *fig. 2*. This arrangement yields the measured values of $q\cos^3\gamma$ directly.

Isoluminance diagram

If the values of $q\cos^3\gamma$ found by measurements from a given observer's position for one lantern are plotted on a plan of the road, and lines are drawn through the points with the same values of $q\cos^3\gamma$, the result is the *luminance-factor diagram* for that position of the observer (*fig. 3a*). This diagram is typical of the reflection characteristics of the road surface (see fig. 3*b*, p. 302). It is usual to take the observer's position at the point 13.5 h from the light source and on a line through the source and parallel to the direction of the road.

If we now put the light source into a lantern and multiply the $q\cos^3\gamma$ value of each point by the value of $I(\gamma,\delta)$ of the lantern for that point, we find the luminance at each point (eq. 2) for that lantern. The lines joining up the points of equal luminance form

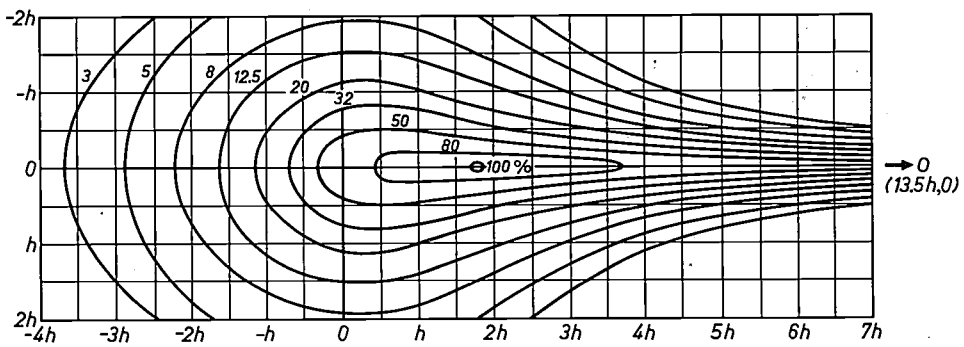


Fig. 3*a*. Example of a luminance-factor diagram. The light source is at a height h above point (0,0), and the observer is at the point (13.5h,0).

the *isoluminance diagram* (*fig. 4a, b*). This diagram is applicable to the combination of a given type of road surface with a given type of lantern, and to the specified position of the observer.

The height of the light source h (the mounting height) appears in these diagrams as the unit of length. The luminance distribution changes uniformly with changes in the mounting height h.

which will give errors never greater than 17% for turning angles likely to be encountered on practical roads. In such a situation the part of the diagram behind the lantern is not rotated, and the part between the lantern and the observer is rotated towards the observer. In fact, the error of 17% is only found for bends with a small radius of curvature, as found in smaller roads.

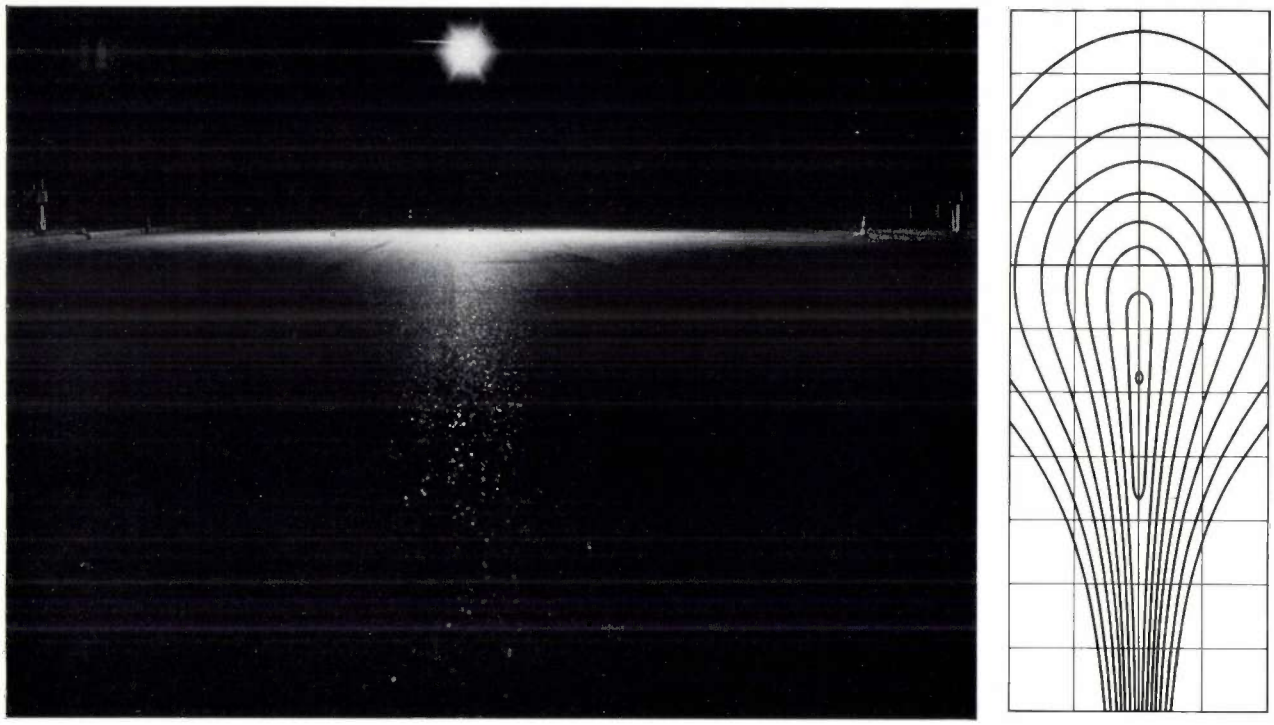


Fig. 3*b*. The road surface to which the luminance-factor diagram of fig. 3*a* applies, lighted by a lamp without a lantern (I = constant). The luminance distribution thus obtained is proportional to the luminance factor distribution (shown again on the right); the photograph gives a kind of perspective picture of the distribution.

Both the luminance-factor diagram of a road surface and the light distribution of a lantern, given by $I(\gamma,\delta)$, can be measured in the laboratory. The point-by-point multiplication of the two distributions has been simplified by the introduction of computers; the isoluminance diagram resulting from this multiplication can therefore be produced entirely in the laboratory, making it possible to dispense with time-consuming measurements out of doors, with their dependence on weather conditions.

If the observer is not situated on the line through the light source and parallel to the direction of the road, or if there is a bend in the road, the luminance-factor diagram is then rotated with respect to the light distribution on the road surface. The product of the two will then yield a different isoluminance diagram. However, it has been shown in our laboratory [3][4] that there is a simple way of using the isoluminance diagram for an observer under the row of lanterns

The design of road lighting of sufficient uniformity

On a road lighted by a row of lanterns there is a patch of minimum luminance between each pair of lanterns. The ratio of this minimum luminance L_{min} to the average luminance $\bar{L}$ of the lighted road surface is a measure of the uniformity of the luminance. The present internationally recommended value for the ratio $L_{min}/\bar{L}$ is at least 0.4 [2]. Since $\bar{L}$ is not so easy to determine, investigations have been made to find out whether the ratio between minimum and maximum luminance L_{min}/L_{max} might also be a useful measure of the uniformity. Both ratios were calculated for 2160 lighting installations; the results are presented in *fig. 5* [4]. The figure shows that there is a high degree of correlation between the two ratios: the two dashed lines enclosing 95% of the cases lie close

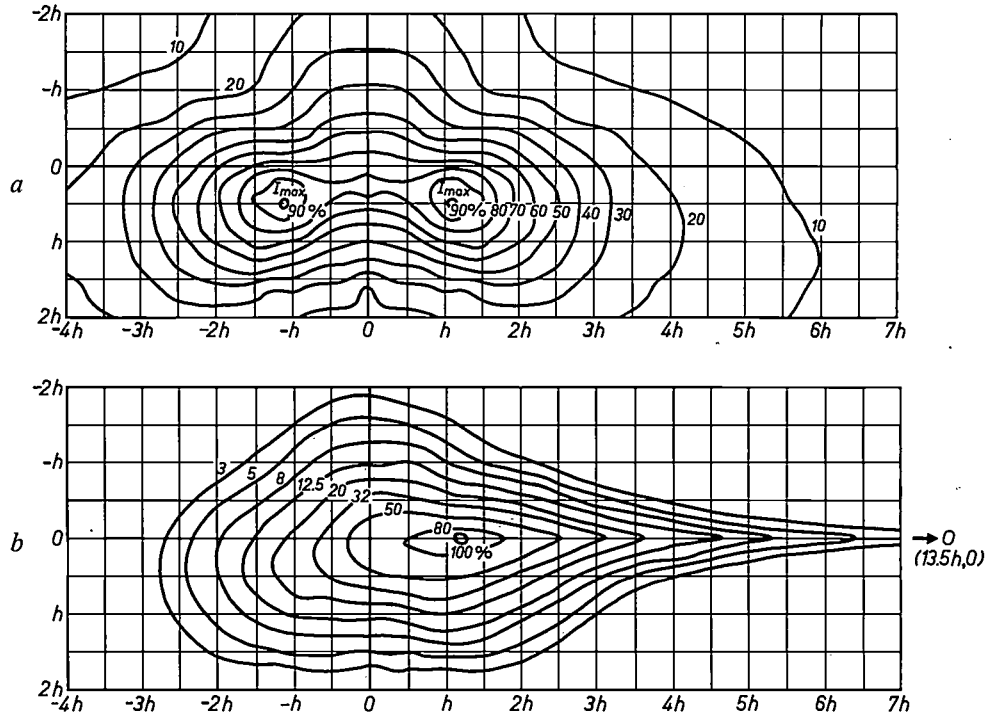


Fig. 4. *a*) Light distribution of a lantern, represented by lines of equal luminous intensity (isocandela diagram) in a horizontal plane. The light source is at a height h above point (0,0). *b*) Isoluminance diagram, obtained as a product of the distribution (*a*) with the luminance-factor diagram in fig. 3*a*.

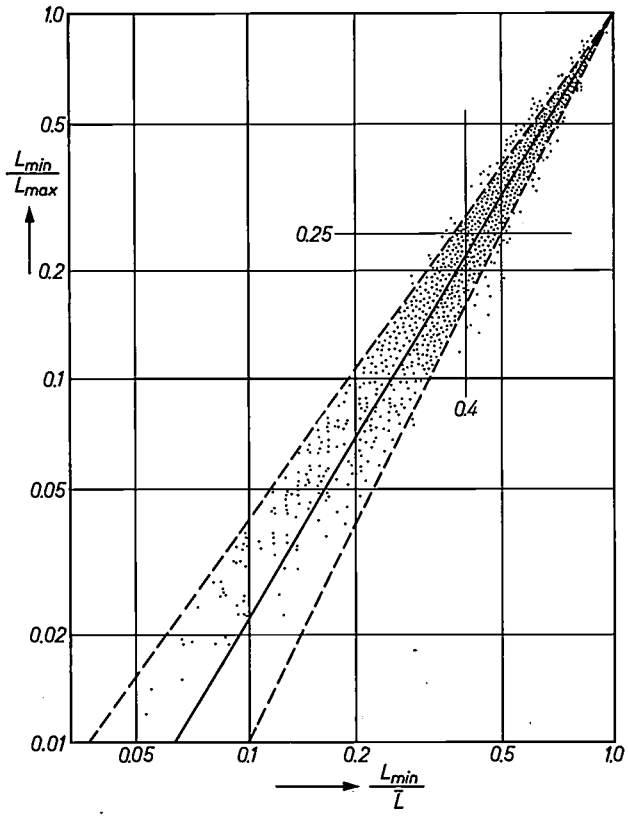


Fig. 5. The values of $L_{min}/\bar{L}$ and L_{min}/L_{max} for 2160 lighting installations. The condition $L_{min}/\bar{L} \geqq 0.4$ is met by nearly all installations for which $L_{min}/L_{max} \geqq 0.25$.

together. Nearly all cases where $L_{min}/L_{max} \geqq 0.25$ meet the condition $L_{min}/\bar{L} \geqq 0.4$. In the procedure described here the design for uniformity is therefore based on the condition $L_{min}/L_{max} \geqq 0.25$. This requires a determination of the luminance at only two points.

The point where the luminance is at a minimum lies between two lanterns and its luminance is determined mainly by these two lanterns. The small contributions from more distant lanterns are neglected in the design procedure. The position for maximum luminance is usually close to a lantern and, provided the distance between the posts is not too small, more than 90% of its luminance is due to this lantern. Here too the contributions of more distant lanterns are neglected. Designing for the best uniformity is thus confined to the luminance distribution under two successive lanterns.

The initial assumption in this design is that each of the two lanterns contributes equally to the luminance at the point of minimum luminance. If the minimum luminance is $0.25L_{max}$, the contribution

[3] B. Knudsen, Public lighting design based upon uniformity of luminance, to be published in Illumination Engineering.

[4] The lighting installations are specified in J. B. de Boer (editor), Public lighting, Philips Technical Library, Eindhoven 1967. See also reference [3].

from each lantern is $0.125L_{max}$ and the minimum luminance occurs at the point of intersection of the isoluminance curves for $0.125\ L_{max}$ of the two lanterns. The design procedure therefore consists in moving the isoluminance diagrams of two successive lanterns over each other on the plan of the road to be lighted until a situation is reached where no point of the road surface between two lanterns falls *outside* the $0.125L_{max}$ isoluminance curves of *both* lanterns (*fig. 6a*). To achieve this, the $0.125\ L_{max}$ curves are moved apart until their point of intersection coincides with the edge of the road. This is where the point of minimum luminance lies. In the immediate vicinity of this point the luminance may perhaps fall to below the permissible minimum of $0.25\ L_{max}$ if the contribution from one lantern decreases faster than that from the other increases. In a final design this should be investigated by calculating the luminance at a number of points around the point of intersection of the curves.

The spacing of the light sources, again expressed with h as the unit of length, is now found from the location of the isoluminance diagrams on the plan.

If the observer is not under the row of lanterns — a case which applies to the drivers of oncoming vehicles on the offside of O in fig. 6*a* — then all the isoluminance diagrams are rotated (fig. 6*b*), and this is also done if there is a bend in the road (fig. 6*c*). If the cases in fig. 6*a* and *b* indicate different light-source spacings, the smaller value must be chosen.

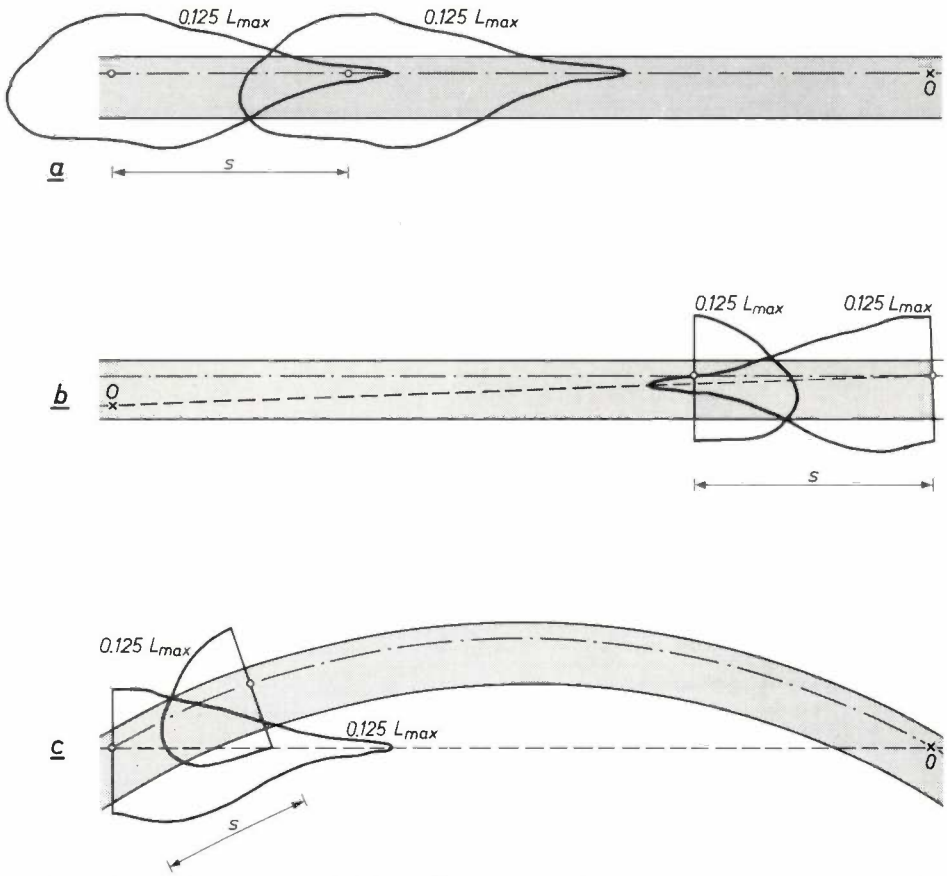


Fig. 6. Isoluminance curves $0.125\ L_{max}$ for two successive lanterns whose spacing s is such that no point of the road falls *outside both* curves.
a) Straight road, observer O (driver of a car) directly under the row of lanterns.
b) Straight road, observer O not directly under the row of lanterns (driver of oncoming vehicle with respect to *a*).
c) A bend in the road.

Average luminance

Once the maximum light-source spacing compatible with the required uniformity has been established, the next problem is to find out whether the average luminance is high enough at this spacing. Since the luminance distribution over the road surface is now known, the average luminance can be calculated as the average of a large number of points. This is time consuming, however, and a faster method has there-

fore been developed. To describe this method we must introduce two new parameters; we shall also describe an instrument that can be used for measuring the magnitude of these parameters for an existing road surface. These parameters are the average luminance factor q_0 — which is small for a dark road surface and large for a light road surface — and the quantity κ, which is a measure of the spread of the luminance factor q over the road surface [5]. The quantity κ is defined as follows:

$$\kappa = \log_{10} \frac{q_0}{q_{\min}},$$

where $q_{\min}$ is the smallest value of q encountered. If the road surface is a perfect diffuser, then q is everywhere equal to q_0; in this hypothetical case, therefore, $q_{\min} = q_0$ and $\kappa = 0$. The more the road surface differs from a perfect diffuser, that is to say the more specular reflection it gives, the higher becomes the value of κ. Since the ratio $q_0/q_{\min}$ can reach very high values for wet road surfaces (e.g. 10^4, compared with values between 1 and 4 for dry road surfaces), it is better to use the logarithm of this ratio as the characteristic quantity.

The quantity κ characterizes the reflection characteristics of the road surface so completely that the luminance-factor diagrams of all road surfaces with the same κ may be assumed to be similar (to the accuracy required in calculating the average luminance). The consequence is that a difference in the average luminance factor q_0 of different road surfaces with the same κ just means that the road surface is lighter or darker. This amounts to saying that q_0 characterizes the amount of the reflected light and κ its distribution, and this approach is what is used in the calculation of the average luminance.

In practice it is necessary to use the quantity

$$\kappa_p = \log_{10} \frac{q_0}{q_p},$$

where q_p is the luminance factor directly under the lantern. This is because q_p is a value that can be measured straight away, whereas the position for minimum q — which is not as a rule directly under the lantern although it is close to it [6] — would first have to be established. A statistical investigation on 73 road surfaces has shown that, with a slight spread, $\kappa = 1.14\ \kappa_p$.

The use of the quantities q_0 and κ_p is attractive because in principle each of them can be determined by a single measurement on a given road surface. Both measurements can be made with the same instrument, developed at the Philips Lighting Laboratory. This instrument, called a *road-surface reflectometer*, will now be described.

The road-surface reflectometer

The quantity q_0 is the luminance factor averaged over all the angles of incidence encountered in practice (the angles γ and δ in fig. 1). Instead of determining q_0 as the average over all the points illuminated from a single point, we can determine the luminance factor of a single point on the road surface which is illuminated from all directions simultaneously. This method can be put into practice by mounting above the road surface a light-radiating ceiling which is large enough to cover all the angles γ and δ likely to be encountered in practical road lighting. The point P to be measured (see *fig. 7*) lies under this ceiling. If the luminance measured at this point is to be proportional to the average luminance factor q_0 of the road surface, the factor $I(\gamma,\delta)$ in equation (2) must be constant. For this condition to be met each surface element of the ceiling should have the same luminous intensity I in the direction of P. Given a diffusely radiating ceiling this implies that the luminance of the ceiling must be inversely proportional to $\cos\gamma$. To achieve this a uniformly illuminated opal-glass ceiling is used, which is covered with a mask of opaque circles as illustrated in *fig. 8*. The "effective" luminance of the ceiling is then propor-

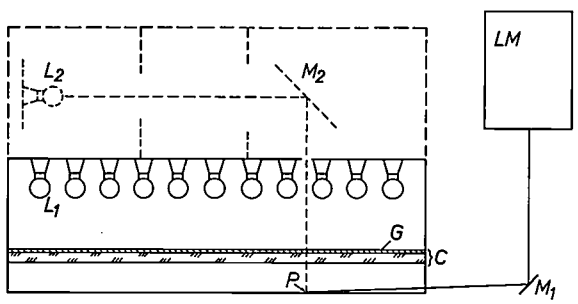


Fig. 7. Cross-section of the road surface reflectometer. The luminance at point P of the road surface is measured with a luminance meter LM by means of a mirror M_1. Point P is illuminated by an opal-glass ceiling C, which in turn is uniformly illuminated by incandescent lamps L_1 and has a mask G of opaque circles applied to it; the fraction of the surface which transmits light is made proportional to $1/\cos\gamma$, thus ensuring that every part of the surface has the same luminous intensity in the direction of point P. In this case the measured luminance is proportional to the average luminance factor q_0.
For the measurement of q_p the ceiling C is removed from the instrument and the point P is illuminated perpendicularly from above by lamp L_2 via mirror M_2. The value of κ_p is found from q_0 and q_p.

[5] J. B. de Boer and J. Vermeulen, Simple luminance calculation based on road surface classification, paper read at the C.I.E. Conference at Washington in 1967.

[6] Although the luminance factor q close to the lantern is at a minimum, the luminance L around it is at a maximum. This is because $\cos^3\gamma$ at this position has a pronounced maximum (see eq. 2).

tional to the fraction of the surface that transmits light.

The ceiling is situated in the instrument at a height of 30 mm above the road surface under investigation; the point P to be measured is perpendicularly below the centre-point of the concentric circles. A small area around P is observed by a luminance meter from an angle of 1° with the road surface [7]. The size of this area, 10×100 mm, is a compromise: it must be small compared with the luminous ceiling to approximate sufficiently to a point, and it must be large compared with the structural details of the road surface to yield a representative measurement. For road surfaces of coarser structure the second requirement is not satisfied, and this means that more than one measurement is necessary on such surfaces. The instrument is calibrated by means of a standard surface of known average luminance factor.

To determine κ_p it is necessary to measure the luminance factor q_p for perpendicularly incident light. For this measurement the instrument contains a special lamp which illuminates the point P directly from above when the ceiling has been removed.

The reflectometer, shown standing on a road surface in *fig. 9*, is supplied from a storage battery and is easily transportable.

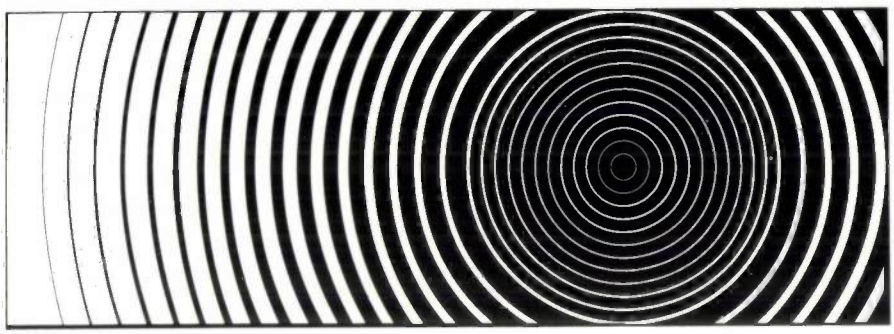


Fig. 8. Mask of opaque circles on the ceiling of the road surface reflectometer. The luminance is measured at the point perpendicularly below the centre of the circles.

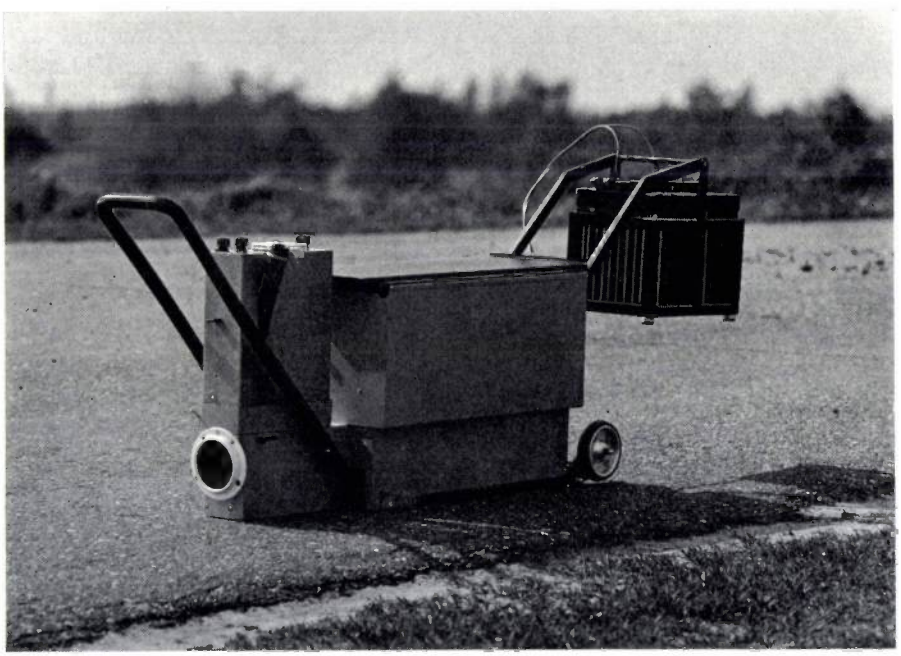


Fig. 9. The reflectometer in position on a road surface.

Road-surface classification and design procedure

The average luminance $\bar{L}$ of the installation now has to be determined with the aid of the values obtained for q_0 and κ_p. The quantity $\bar{L}$ is first of all proportional to the average luminance factor q_0,

to the output of the lamp in the lantern, expressed by its luminous flux Φ, and inversely proportional to the surface area to be illuminated per lantern. If the road is w metres wide and the spacing between the lamp posts is s metres, this area is ws m². There is also a more complicated dependence of the average luminance on the light distribution of the lantern, the κ_p value of the road surface, the configuration of the installation (single-sided, staggered, central, etc.) and the position of the observer. All these effects are included in a single proportionality factor τ, and we can therefore write:

$$\bar{L} = q_0 \tau \Phi/ws.$$

The value of τ now has to be calculated for all possible cases from the data relating to the road surface and lanterns.

To summarize the many situations which can occur, a road surface classification by κ_p value has been introduced. The following division into five classes is in our view sufficiently accurate, and has been proposed to the International Commission on Illumination.

Class I: $\kappa_p < 0.22$
II: $0.22 \leqq \kappa_p < 0.33$
III: $0.33 \leqq \kappa_p < 0.44$
IV: $0.44 \leqq \kappa_p < 0.55$
V: $0.55 \leqq \kappa_p$.

With increasing class number the reflection of the road surface becomes less diffuse. Nearly all (dry) road surfaces encountered come under classes I to IV. A road surface may come into class V if it has been worn very smooth.

Since road surfaces with identical κ (or κ_p) have similar luminance-factor diagrams, the luminance distribution of a road surface with a given lantern can be found from the light distribution of the lantern and the κ_p value of the road surface. The average luminance, and hence the value of τ for this combination, can also be obtained by calculating the average over a large number of points. This has been done with the aid of a computer for a large number of combinations, assuming a lighting installation of *standard configuration*. Any practical installation can be built up from its elements, and the τ value may then be determined as follows.

The standard installation (see *fig. 10*) consists of a row of lanterns at a mounting height of h metres above the axis of a road $4h$ wide. The road is taken as being divided in the longitudinal direction into strips

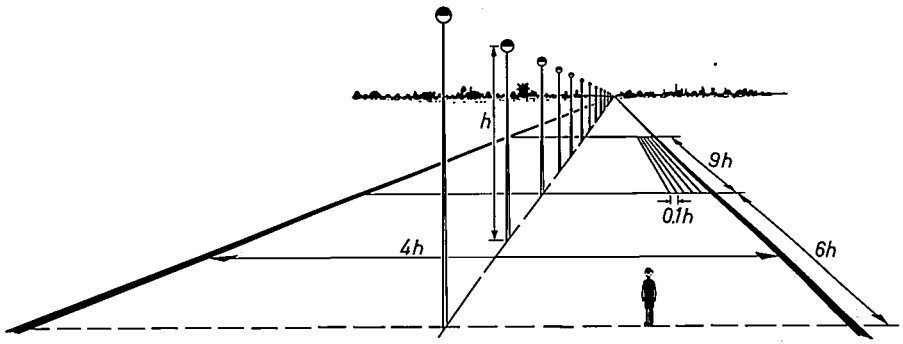


Fig. 10. Standard configuration of a road with lanterns. Any conceivable configuration for which the average luminance is to be calculated can be built up from the strips into which the road is divided. If there are more rows of lanterns the calculation is done in steps and the results added.

$0.1h$ wide, extending from $6h$ to $15h$ in front of the observer. Every practical installation will consist of a number of such strips (where there are lanterns on both sides of the road, two installations are superimposed). The value of τ is calculated separately for each strip and the average luminance of the jth strip (where j indicates the number of the strip and runs from —20 to +20) is given by:

$$\bar{L}_j = q_0 \tau_j \frac{\Phi}{0.1\, hs}.$$

If, for example, a design for an installation consists of the strips —1 to 10, the average luminance $\bar{L}$ follows from:

$$\bar{L} = \frac{1}{11} \sum_{j=-1}^{10} \bar{L}_j = q_0 \frac{\Phi}{1.1\, hs} \sum_{j=-1}^{10} \tau_j. \quad . \quad (3)$$

[7] The luminance meter we use is described in W. Morass and F. Rendl, Portable measuring instrument for road-surface luminance, Light and Lighting **60**, 157-160, 1967.

It simplifies the calculation if τ_j is not quoted separately for each strip but straight away as the sum

$$\sum_{j=1}^{n} \tau_j ,$$

as a function of n. This is done by giving $\Sigma\tau_j$ curves. An example is shown in *fig. 11*, for a lantern with the light distribution of fig. 4*a* and a road surface belonging to class II. The figure shows three curves,

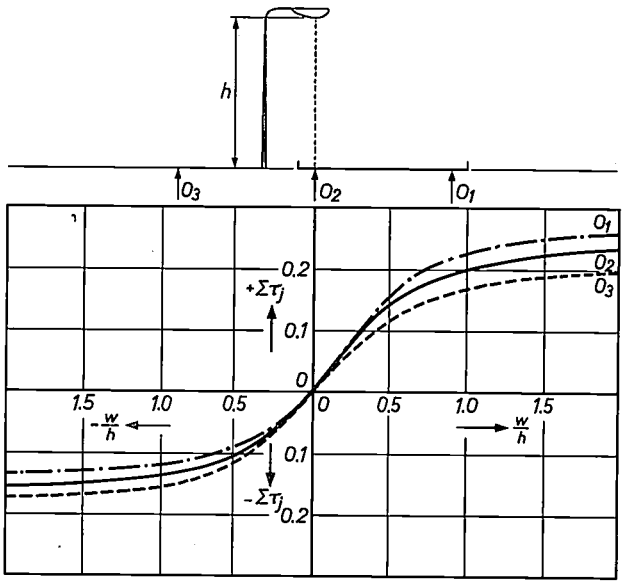


Fig. 11. Graph of $\Sigma\tau_j$ plotted on the transverse cross-section of the standard installation (fig. 10). These curves apply to the lantern whose light distribution is given in fig. 4 and a road surface of class II. Above the graph there is a diagram of the design used for the calculated example in the text, in the correct situation with respect to the graph and to the correct scale. There are three curves for the three observation positions O_1, O_2 and O_3.

corresponding to the three different observation positions O_1, O_2 and O_3 shown in the little drawing above the graph in fig. 11. The $\Sigma\tau_j$ value is read from the graph by drawing the transverse cross-section of the design on the horizontal axis, to scale. The $\Sigma\tau_j$ value for the right-hand kerb, minus that for the left-hand kerb, gives the $\Sigma\tau_j$ value for the width of the road. The road illustrated in the little drawing above the graph extends in width from $-0.1h$ to $+1.0h$. The $\Sigma\tau_j$ value for the observation position O_2 in this case is $\Sigma\tau_j = 0.20 - (-0.03) = 0.23$.

Once the value of $\Sigma\tau_j$ has been determined in this way, the required spacing s between the lamp posts at the desired average luminance $\bar{L}$ can be found from equation (3).

In some lanterns lamps of different light output can be used without causing much change in the light distribution. If the spacing s required for the specified average luminance is appreciably less than that required for adequate uniformity, it may be cheaper to use a more powerful lamp.

If the manufacturer of a road-lighting lantern supplies the isoluminance diagram and the set of $\Sigma\tau_j$ curves for the lantern with different positions of the observer, for each of the five κ_p classes, the procedures we have described enable different lanterns to be quickly compared when drawing up a new design. The procedures are accurate to within $\pm 15\%$. Once a particular lantern has been decided upon, a point-by-point calculation for this lantern can be carried out if greater accuracy is required, and, if necessary, corrections can be applied to the calculated spacings between the posts.

Summary. The light sources of a road-lighting installation produce on the road surface, as observed by a road user, a luminance distribution of complicated form. Every road-lighting installation has to meet certain requirements for the uniformity of the luminance distribution and the average level of the luminance. Both characteristics are usually improved by reducing the distance between the light sources, but this makes the installation more expensive. A rapid procedure for determining the maximum permissible spacing of light sources has been developed at the Philips Lighting Laboratory. With the aid of this procedure different lanterns can be compared to see which is the most suitable. The method makes use of a classification of road surfaces according to their light-reflecting characteristics. An instrument is described which establishes the average luminance factor of a road surface in one measurement, and in a second measurement a quantity which, together with the average luminance factor, determines where the road surface stands in this classification.

Superconducting magnets

A. L. Luiten

When superconductivity was discovered, more than half a century ago, there was every indication that the effect could soon be put to use in the generation of intense magnetic fields by means of superconducting coils. In fact, however, this has only become a reality in recent years. The article below gives a general introduction to the subject, followed by a description of a number of superconducting magnets constructed at Philips Research Laboratories in Eindhoven.

Very soon after the discovery of superconductivity, by Kamerlingh Onnes and Holst in 1911, it was realized that the use of superconducting material might enable large electromagnets to be constructed which would give a very high field strength [1]. When windings of normally conducting material are used, a considerable amount of electrical power has to be supplied, and this power is converted into heat, which is not easily conducted away. Superconducting magnet windings, on the other hand, do not need any electrical power to maintain a magnetic field, regardless of the field strength and dimensions of the coil. It is of course necessary to keep the windings at a temperature below the transition point T_c of the coil material, i.e. the temperature at which the material changes from the normal to the superconducting state. Since T_c is no more than a few degrees Kelvin (e.g. for lead 7.2 °K and for tin 3.7 °K), this necessitates the use of liquid helium.

A magnet with a cooled lead winding, made by Kamerlingh Onnes, would not however give a field greater than a few hundred oersteds; when the current was increased there was a particular value at which the winding recovered its normal electrical resistance. This led to the discovery of the critical field strength H_c, which is the maximum value of an external field at which a given material can be in the superconducting state. In all known superconductors at that time, H_c was found to be smaller than 1000 Oe (8×10^4 A/m).

In 1931 De Haas and Voogd [2] found a bismuth alloy which had a much higher H_c (it was 15 000 Oe), but wires of this material changed to the normal state at a current for which the magnetic field generated by the current itself was very much weaker than H_c [3]. In establishing this fact they had for the first time come up against the third important phenomenon. In wire of superconducting alloys and compounds the maximum possible current is much lower than that at which the field generated at the surface of the wire by the current itself reaches the value H_c. A characteristic of such materials is the critical current density j_c, which depends on the temperature and on the magnitude of any external magnetic field that may be present.

The first magnet capable of generating a reasonably high field (7000 Oe) was not built until more than 20 years later, by Yntema [4] in 1955. The coil material used for this magnet was cold-worked niobium wire.

The vigorous development now in progress can be said to have really begun, however, in 1961. In that year Kunzler *et al.* [5] reported on a magnet wound from molybdenum-rhenium wire, which could give 15 000 Oe. Shortly afterwards some materials were discovered which combined a high transition temperature with a high critical field strength, and could be made to give a high critical current density by a suitable metallurgical treatment and cold working. Coils of niobium-zirconium wire were soon being made which gave field strengths of 60 to 70 kOe, while niobium-tin wire seemed to offer even greater promise. Getting through the "100 kOe barrier" gave much more difficulty, however, than had been expected, and this was in fact only achieved in 1963. Progress in the 100 kOe region is much slower. A magnet to give about 140 kOe is under construction in the U.S.A. With the use of niobium-tin wire a field strength of over 200 kOe is

Ir. A. L. Luiten, now with the Philips X-Ray and Medical Equipment Division, was formerly with Philips Research Laboratories, Eindhoven.

[1] A discussion of the most important features of superconductivity including the state of current theoretical knowledge is to be found in: J. Volger, Philips tech. Rev. **29**, 1, 1968 (No. 1).
[2] W. J. de Haas and J. Voogd, Comm. Phys. Lab. Univ. Leiden **19**, No. 214c, 1931.
[3] W. H. Keesom, Comm. Kamerlingh Onnes Lab. Univ. Leiden **21**, No. 234f, 1935.
[4] G. B. Yntema, Phys. Rev. **98**, 1197, 1955.
[5] J. E. Kunzler, E. Buehler, F. S. L. Hsu, B. T. Matthias and C. Wahl, J. appl. Phys. **32**, 325, 1961.

theoretically attainable. The extremely high permissible current density of this material — at 100 kOe, for example, a current density of about 200 kA/cm² is possible — permits field strengths as high as this to be achieved with coils of acceptable dimensions. On the other hand, now that superconducting coils can be made, fairly high fields can also be produced which extend over a region of several cubic metres — something that was previously unthought of.

A unique feature of superconducting magnets is that they can be used to give a perfectly constant magnetic field by short-circuiting the coil with a superconducting switch after the energizing current has been switched on (*fig. 1*).

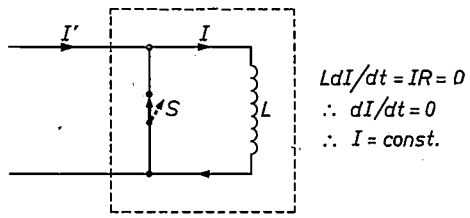


Fig. 1. When the coil in which a current I is flowing is short-circuited by means of a superconducting switch S, thus forming a closed superconducting circuit — i.e. one whose resistance R is zero — the current I then no longer reacts to variations in the current I' in the external circuit. Variation of I' only causes a change in the current flowing in S.

The problems encountered in the design of a superconducting magnet are very different from those encountered in the design of an ordinary electromagnet. The problem of removal of the heat due to Joule heating is one that does not arise here, and this is of course a great advantage, but the designer is faced with a number of other problems whose nature and seriousness depend on the dimensions and field strength of the magnet, and which thus differ from case to case. These problems include those due to the properties of the coil material and those that arise because in large magnets — or to be more exact, magnets of high field energy — an abrupt transition of the coil wire to the normal state can have catastrophic results if appropriate precautions are not taken. Since the resistance of the winding wire is often very high in the normal state, a considerable amount of Joule heat is developed. Because of this, the extent of the normal region rapidly increases and so too does the resistance of the coil. This causes the current to fall at an increasing rate, giving rise to a fairly large induced voltage which can cause breakdown. Also, the field energy released in the form of Joule heat in the normally conducting region can cause local heating intense enough to damage the insulation or even melt the wire.

The discussion of these and related problems will take up a large part of this article. First of all, however, we shall briefly consider various superconducting materials and their suitability for application in magnets. The article concludes with a section on the design, construction and use of experimental magnets, in which a number of the magnets made in our Laboratories are described (*fig. 2*).

Superconducting materials

Superconductors of three kinds

A distinction is made between superconductors of the first, second and third kinds [1]. Those of the first kind, also known as "soft superconductors", are the familiar ones like tin, lead, mercury and indium. The comments made above about the critical field strength H_c are applicable to these substances: apart from local effects due to demagnetization, the whole of a sample subjected to an external field of this value changes from the superconducting to the normal state. Moreover, below H_c the flux density B is zero in the whole sample apart from a thin surface layer. In this surface layer there is a current, the "screening" or "Meissner" current, which produces a flux density inside the sample which exactly cancels the flux density due to the external field (Meissner effect). The magnetic behaviour of these materials is represented in *fig. 3* by curve *1*. The field strength H_c is at the most a few hundred oersteds.

In superconductors of the second kind the field *does* gradually penetrate. When the external field is increased, this penetration begins at a relatively low value H_{c1} (a few tens of oersteds) but is only completed at a value H_{c2} which is usually much higher than the critical field strength H_c of the soft superconductors. Between H_{c1} and H_{c2} material is said to be in the "mixed" state. The flux penetration takes place in the form of millions of flux "threads" each of which carries at its core a flux Φ_0, the "flux quantum". Each flux thread in turn has a screening current flowing around it, and for this reason the threads are sometimes called vortices. At $H = H_{c2}$ the normal cores of the flux threads occupy the entire volume exactly.

In superconductors of the third kind, the "hard superconductors", the situation is basically the same, except that here the flux threads cannot penetrate directly into the whole volume. First of all, they occupy a number of small domains at the surface, and, as the field increases, deeper domains are gradually "conquered" by the magnetic field. We shall return to this presently.

The three types of superconductor behave very differently when a transport current flows. This aspect of the behaviour of superconductors, which was mentioned in the introduction and is perhaps the most impor-

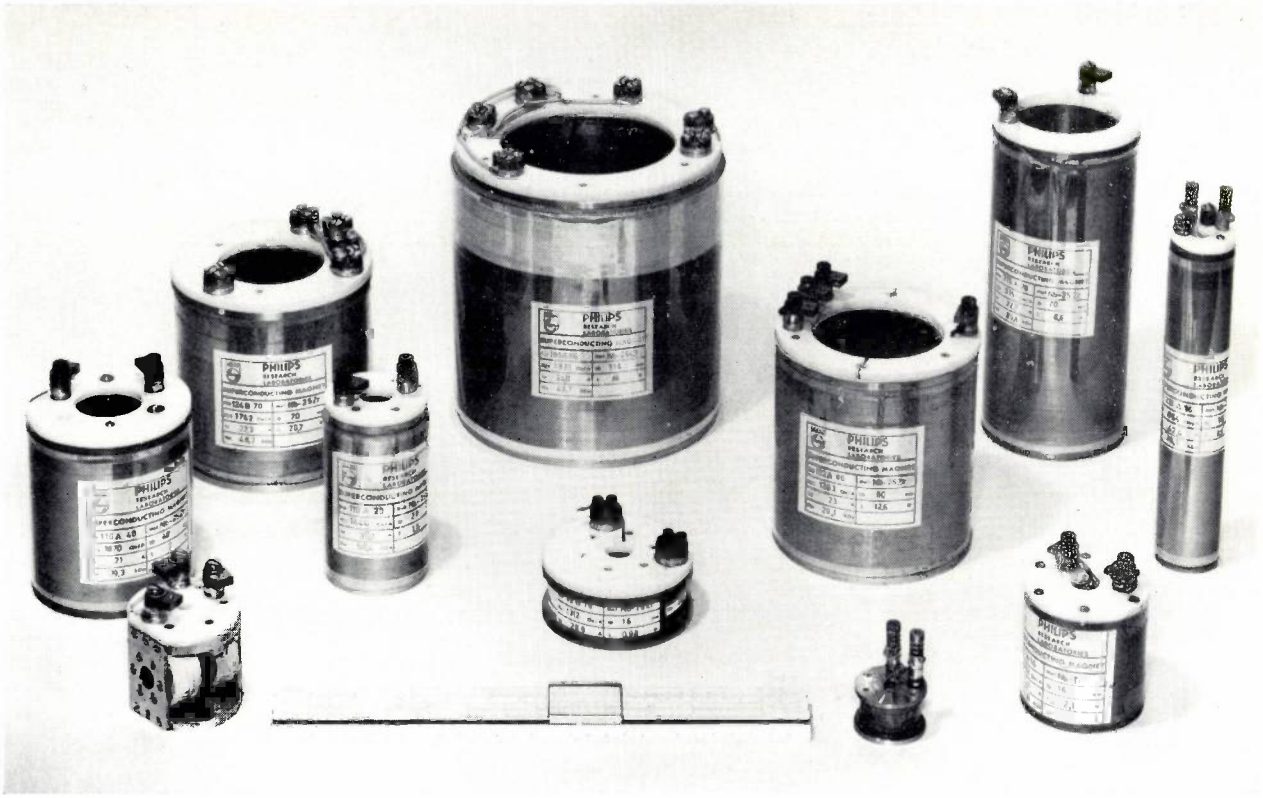


Fig. 2. Some experimental superconducting magnets made at the Philips Research Laboratories in Eindhoven. The small magnet at the front and on the left has an air gap and has been specially made for optical research. The magnet coil at the extreme right gives a highly uniform and readily calculable field, and one of its uses is for calibrating Hall elements for field-strength meters.

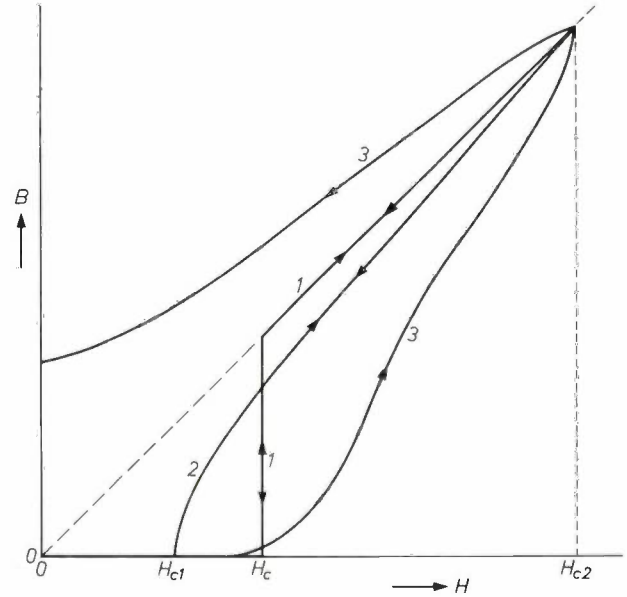


Fig. 3. Variation of flux density B with field strength H for the three kinds of superconductor. In superconductors of the first kind (curve *1*) the flux density in the region $0 < H < H_c$ is zero (Meissner effect). At $H = H_c$ the material changes to the normal state and the field can penetrate freely. In the region $H > H_c$, $B = H$. In superconductors of the second kind (curve *2*) the penetration of the field when H is increased begins at the relatively low field strength H_{c1}, but is not completed until the relatively high field strength H_{c2} is reached. Curves *1* and *2* apply both for increasing and for decreasing field strength. Much the same kind of thing happens in superconductors of the third kind as in those of the second kind, but in addition to the reversible component there is a second component which shows hysteresis, so that in this case B is not uniquely determined by H.

tant to the designer, will now be examined at greater length. The further details of the magnetic behaviour, which will be given later, will then be easier to follow.

In a superconductor of the first kind, i.e. one that exhibits the Meissner effect ($B = 0$), the transport current flows in a thin surface layer, just like the Meissner current. The highest value I_c that transport currents can have in superconductors of this type is determined by the magnetic field strength which this current itself produces at the surface of the superconductor (Silsbee's rule). When an external transverse field H is applied, a transport current can again flow, provided the maximum field strength occurring at the surface of the superconductor does not exceed the value H_c.

A superconductor of the second kind behaves identically as long as H is less than H_{c1}, but in the region $H_{c1} < H < H_{c2}$ the situation is completely different. Theoretically, no supercurrent can flow in an ideal superconductor of the second kind because such a current, however small, would cause the superconducting state to change immediately into the normal state. Superconductors of the third kind, on the other hand, have a very high current-carrying capacity. The physical background to these two situations will now be described.

The flux threads in a superconductor of the second kind in the mixed state are uniformly distributed over

the cross-section. When a transport current is passed through a wire in the mixed state, the flux threads move in a direction perpendicular to the direction of the current, so that $dB/dt \neq 0$ locally. As a result small currents are induced which generate heat, for one reason because they also flow through the normal core of the flux threads. The superconducting material thus exhibits, surprisingly, a certain resistance; a persistent current cannot flow in such a superconductor. In a superconductor of the third kind the flux threads can move only over a small distance; they very soon come up against an obstacle. This could typically be a lattice defect of some kind or other, a grain boundary, and so on. This explains to some extent why the current-carrying capacity of a given material can be increased by cold working, since the deformation is bound to create a large number of lattice defects.

This pinning of the flux threads also affects the magnetic behaviour. When an increasing external field exceeds the value H_{c1}, then, as we have said, only a part of the sample is invaded by flux threads, and broadly speaking this is the part that lies outside the outermost barriers. In the absence of a transport current the domains where the field has penetrated are uniformly filled with flux threads.

Inside these domains the wall currents of the flux threads therefore cancel one another out, and only at their surface is there a net current. Macroscopically, the situation now is that the screening current no longer flows everywhere at the surface but partly along the deeper-lying walls between the domains. As the external field increases, the mutual repulsion of the flux threads becomes greater and the barrier bounding the domain can be overcome. Owing to the heat generation associated with the flux displacement the barrier loses its superconduction and consequently collapses completely, and as a result the next domain "fills up" with flux threads, and so on.

The penetration of the magnetic field into a sample thus proceeds in jumps, and each flux jump is accompanied by the generation of a certain amount of heat. When the domains are relatively large, the heat generated can become quite considerable. The same applies when a preceding barrier is much higher than a number of successive ones, so that when it collapses several domains fill up at the same time. Flux jumps also occur when a transport current changes.

The change in B cannot be greater than the difference in ordinate of curve *3* (fig. 3) and the straight line $B = H$; this maximum value decreases as H approaches H_{c2}.

Since the screening currents flowing in the domain wall oppose *every* change of field in the material, a decrease in the external magnetic field will also tend to prevent the *escape* of the flux that has penetrated into a superconductor of the third kind. This explains the considerable magnetic hysteresis shown by these superconductors. The hysteresis increases with the current-carrying capacity of the material and the size of the sample.

Superconducting material for magnet coils

It will be evident from what has been said above that only superconductors of the third kind can be considered for use in the winding of magnet coils. The transition temperature of the material should be well above the operating temperature, which is usually 4.2 °K, the boiling point of helium. The reason for this is that the critical current I_c in the wire and the critical field strength H_{c2} both increase with the difference between T_c and the operating temperature. The critical field strength is important because H_{c2} is the absolute upper limit of the field strength. Finally, if the dimensions of the coil are to be kept within reasonable bounds, the critical current density j_c at the operating temperature must not be too small either.

Pure metals are no longer used for making superconducting coils; most superconducting chemical elements are superconductors of the first kind, the others being unsuitable because of their relatively low H_{c2} (cf. Yntema's niobium coil). The materials nowadays used are therefore either alloys or chemical compounds of two metals. These two groups will now be briefly discussed.

The principal *alloys* now used for winding coils are Nb-Zr and Nb-Ti; Mo-Re has also been used in the past. The critical field strength H_{c2} of these alloys depends on the proportions of the constituents of the alloy, which can usually be varied between wide limits. For example, niobium with 25% zirconium has a critical field strength of about 68 kOe, but if the zirconium content is increased, the value of H_{c2} can be raised to about 90 kOe. The highest attainable current density, on the other hand, is obtained with a zirconium content of 25%. The critical current density can be considerably increased by metallurgical treatments which promote the formation of dislocations, inhomogeneities and precipitates. *Fig. 4* shows the critical current I_c for a particular Nb-Zr alloy as a function of the magnitude H of a transverse external magnetic field, for various temperatures. It can be seen that $I_c(H)$ and H_{c2} decrease with increasing temperature.

A higher critical field strength than that of Nb-Zr alloys can be achieved with 60Nb40Ti, i.e. 120 kOe. At field strengths below 40 kOe, however, materials of this composition until recently gave greater flux jumps than Nb-Zr, and it was therefore only used for the innermost windings of the coil, where the field strength

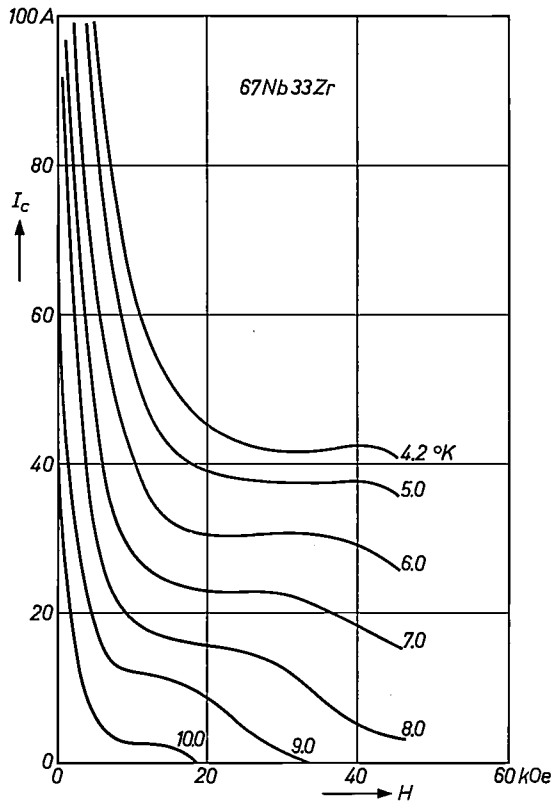


Fig. 4. The critical current I_c at various temperatures as a function of the magnitude H of the magnetic field for short samples of a given type of wire (67Nb33Zr, ¼ mm dia.). The transition temperature of this material is between 10 and 11 °K. The field strength at which I_c becomes zero is the value of H_{c2} for the temperature in question.

is greatest. Nb-Ti alloys with 40-80% Ti are now available which are perfectly suitable for use at relatively low fields too, and indeed are even better than Nb-Zr (*fig. 5*).

In spite of the extensive cold working the wire made from alloys is sufficiently ductile to be used for winding coils. This is unfortunately not the case with all the materials from the other group, the *intermetallic compounds*. The most suitable of these are very brittle. In other respects the majority of these materials are more useful than the alloys. For example, Nb_3Sn has a very high transition temperature (18.1 °K) and at 4.2 °K a critical field strength of no less than about 220 kOe. Below 180 kOe it also has the highest critical current density of all the superconductors known at present (fig. 5).

The compound Nb_3Sn is used in three different forms: plain wire, cabled wire, and ribbon. In recent years the ribbon has almost entirely supplanted the other two. The great difficulty with wire and cable is that because of its brittleness the Nb_3Sn can only be formed from the elements after winding. This necessitates a lengthy annealing process, and therefore severely restricts the choice of the other materials to be used in the coil.

The Nb_3Sn ribbon is usually made by applying a thin layer of Nb_3Sn (about 10 μm thick) to both sides of a thin strip of stainless steel (thickness of about 40 μm), which acts as a base. The Nb_3Sn is formed by the deposition of niobium and tin from vaporized niobium chloride and tin chloride at about 1000 °C. In another method the tin is diffused at high temperature into the niobium ribbon where it combines to form a thin surface layer of Nb_3Sn. In both methods the superconducting Nb_3Sn is coated with a layer of copper or silver, which considerably improves the electrical properties of the ribbon, as we shall see below.

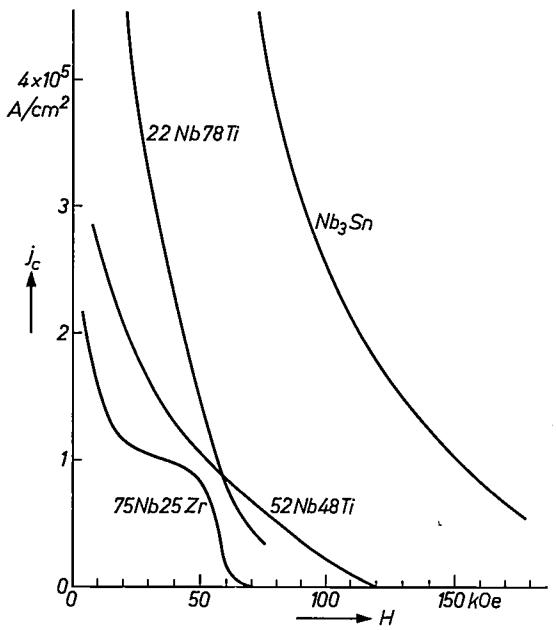


Fig. 5. The critical current density j_c as a function of H, the magnitude of a transverse magnetic field at 4.2 °K, measured for short, straight samples of various coil materials. The two curves for Nb-Ti refer to the best materials of this type of alloy at present available.

Electrical and magnetic characteristics of superconducting magnets

The design of ordinary electromagnets is much the same whether they are large ones or small ones. This is far from true for superconducting magnets; some problems encountered with large magnets are of no significance with small magnets, like providing safeguards against the effects of a sudden transition to the normal state. There are many other respects in which superconducting magnets differ from ordinary types. Before discussing designs, therefore, we shall briefly consider the electrical and magnetic characteristics of superconducting magnets.

Magnetic energy: inductance

In energizing a superconducting magnet, energy is only required for making the current *increase*. This energy Q is converted reversibly into magnetic field energy. If the inductance L of the coil is constant, we can write $Q = \frac{1}{2} LI^2$, where I is the final value of the current. In practice L is not entirely constant, on account of the magnetic processes in the coil wire which we have discussed above. As the current increases, the material of the superconducting coil tends to oppose the penetration of the field, giving rise to a negative magnetization which reduces the inductance of the coil. With increasing field strength this magnetization gradually decreases, and the inductance approaches the value L_1 for a copper solenoid (*fig. 6*). If the current is now allowed to decrease, the screening current first reverses direction before the magnetic field in the wires changes. This gives a small reversible region in the L-I curve for which the value of the inductance is only L_0, the inductance of a hypothetical coil wound from perfectly diamagnetic material.

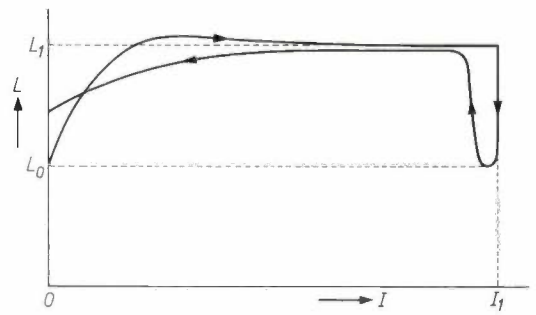


Fig. 6. When the current I through a superconducting coil is increased from zero to I_1 and then returned to zero again, the inductance L varies in the manner shown by the curve. The value L_1 is the inductance of a coil of normally conducting copper wire. The much lower value at low field strength and immediately after reversal of the current (minimum value L_0) is due to the magnetization of the coil wire.

Field linearity, residual field and uniformity

The field strength in a normally conducting solenoid with no iron core is proportional at all points to the magnitude of the current in the coil and can be calculated exactly. In a superconducting magnet the field produced by this current has superimposed upon it a magnetic field due to the magnetization of the coil, and the form and magnitude of this field depend on the magnitude of the current in the winding. This contribution is by no means negligible, particularly in the parts of the bore where the field strength is lower than 10 kOe. At the ends of a coil the "magnetization field" is the greater and may be as high as a few hundred oersteds. It is much weaker at the middle of the coil (a few tens of oersteds) and its direction depends on the shape of the coil (*fig. 7*). Because of the magnetization the relation between field strength and current in a superconducting magnet is not completely linear and the field strength is not uniquely determined by the current (i.e. there is hysteresis).

After the current has been switched off, a part of the magnetic flux that has penetrated remains "frozen in" as a kind of remanence in the superconductor. The magnetization thus causes a residual field, which is fairly high at the ends of the coil. The residual field can only be removed by heating the coil until its temperature is higher than the transition temperature, when the magnetization disappears.

A factor of great practical importance is the effect of the magnetization field on the uniformity of the magnet field. The field is more uniform when the current is increasing than when it is decreasing since, in the first case, the magnetization field gives a positive contribution which increases towards the ends of the coil, thus partly compensating the decrease in the "current field".

This magnetization field, which is not uniquely de-

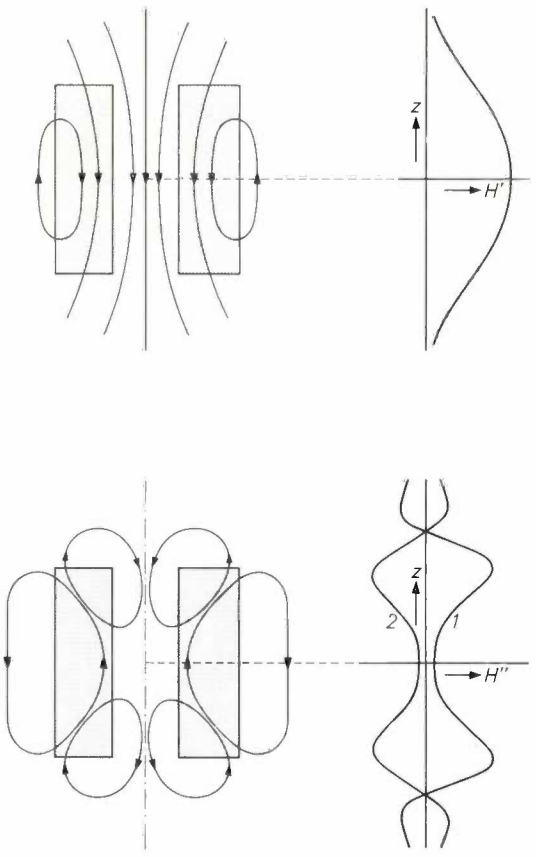


Fig. 7. Upper left, the lines of force of the magnetic field generated by the current flowing in a superconducting coil of the cross-section shown. Lower left, the lines of force of the magnetic field due to the magnetization of the windings. The direction of the arrows refers to the situation in which the current is increasing. Upper right, variation of the field strength H' of the "current field" at the centre-line of the coil as a function of the co-ordinate z. Lower right, the magnitude of the "magnetization field" H for increasing current (curve *1*) and decreasing current (curve *2*).

termined by the magnitude of the current, makes it very difficult to construct coils that will give a uniform field. The difficulties are least with coils for which the thickness of the winding is small compared with the diameter and which only have to supply a relatively low field.

Current-degradation and training effect

An effect found with the first magnets of Nb-Zr was that when the current was increased the winding changed locally to the normal state at a current that was no more than 30 to 50% of the value of I_c found from the measurements on short lengths of wire (fig. 4). This "degradation effect", as it is called, was found to be greater in large coils than in smaller ones. With large magnets it was not at first possible to produce very high fields; in fact, increasing the number of turns led to a lower maximum field strength. A peculiarity of the maximum current was that with different magnets wound from the same type of wire, it was found, over a considerable range to be almost independent of the field generated at that particular current. A second peculiarity that was noted was that the maximum current was not usually reached until the coil had first gone normal a few times at a lower current; this is known as the training effect. The nature of the degradation effect has meanwhile been established, and it has now become possible to construct large coils which can generate very high fields. The physical background of the training effect however has not been fully clarified as yet.

The cause of the degradation effect is to be found in the generation of heat in the flux jumps that occur when the current is increased. When the heat associated with a flux jump is sufficient to cause the wire to change locally to the normal state, a return to the superconducting state will only take place at a relatively low current. If the current is higher than a particular threshold value I_{ms}, the amount of heat generated in the region that has gone normal is greater than the amount of heat dissipated, and this region expands. This danger increases with the length of the piece of wire that has been returned to normal by the flux jump; with a short piece the longitudinal dissipation of heat is relatively greater.

Fig. 8 gives a three-dimensional representation of the combinations of I, T and H at which a given material changes from one state to the other. The points inside the surface which cuts the co-ordinate planes at the curves shown on the figure correspond to situations in which the material is superconducting. From a figure of this type we can derive the way in which the temperature increment ΔT_n which just brings about the change from the superconducting to the normal state depends on current, operating temperature and magnet field. It can be seen that ΔT_n decreases as H approaches H_{c2} and as the current increases. On the other hand, as we have seen, the magnitude of the flux jump decreases with increasing field strength. This decrease is so considerable that the chance that ΔT_n will be exceeded is greatest at low field strengths and zero at field strengths close to H_{c2}. This is illustrated schematically in *fig. 9*. The region of I-H combinations that can lead to a transition to the normal state is

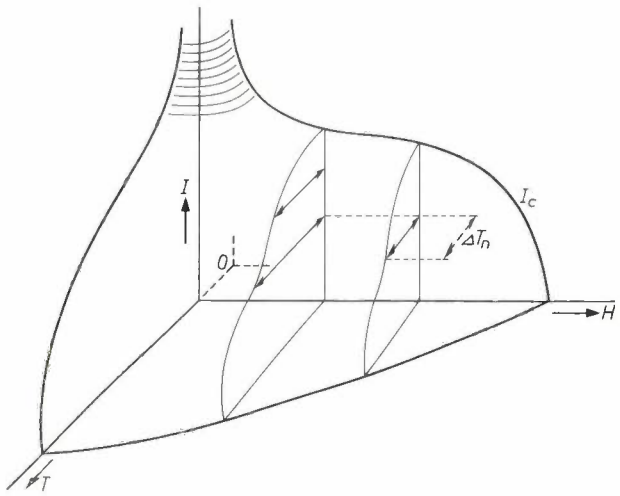


Fig. 8. Schematic representation of the three-dimensional surface in the I-H-T space inside which the points correspond to the combinations of I, H and T which give the superconducting state. At a given temperature T the temperature increment ΔT_n which just causes the sample to go normal increases with H and I. The I-H plane (cf. fig. 9) is drawn at the T-value corresponding to the operating temperature.

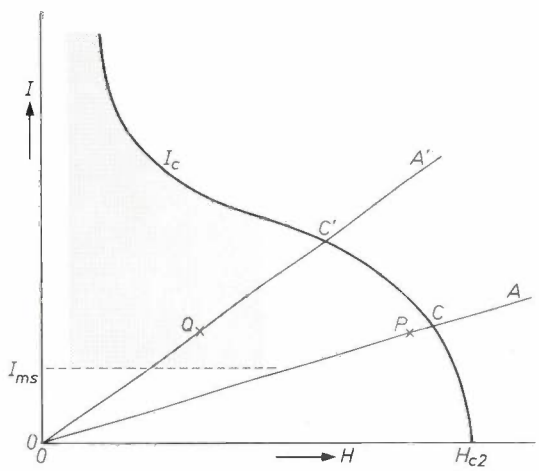


Fig. 9. I-H diagram containing an I_c-H curve like the one in fig. 4, and two operating characteristics for one of the innermost windings (OA) and an outer winding (OA'). The characteristic for the outer winding is steeper because the field in the winding decreases towards the outside. With the combinations of I and H in the shaded region there is a certain chance of the coil going normal because of the generation of heat associated with flux jumps. This effect makes it difficult to reach the point C' on the line OA'. (In practice this is the current-degradation effect.) One peculiar feature is that the parts of the coil in which the field strength is highest are the ones least subject to this particular difficulty. (At currents lower than I_{ms} an element that has gone normal does not continue to grow but ultimately returns to the superconducting state.)

schematically indicated in this figure by the shaded area. The figure shows that it depends on the slope of the operating characteristic — i.e. the relation between the current I and the field H which it generates — whether or not this region will be traversed when the current is increased. If this happens, then there is a certain chance that the maximum current I_m will not be equal to the ordinate value at the point of intersection of the operating characteristic and the curve for I_c: in such a case I_m lies somewhere between I_c and I_{ms}. The paths of the operating characteristics are obviously not the same for the different parts of a coil. For instance, the innermost winding, in which the highest field is found at a particular current, could have the characteristic OA, and a winding nearer the outside of the coil the characteristic OA'. At the current corresponding to the point P for the innermost winding — a permissible situation for this winding — the operating point Q of the other winding is in the dangerous region. Paradoxical though it may seem, a coil often breaks down because a part where the field is relatively *low* goes normal. Because of the current-degradation effect, it is impossible to predict with certainty what maximum current may be expected in a given magnet coil.

In large coils the flux jumps usually give rise to greater pulses of heat than they do in small ones, and I_m is correspondingly lower; in very large coils I_m is sometimes very little greater than I_{ms}, perhaps 10% of I_c. The reason for this is that a local flux jump is capable of inducing several flux jumps of different magnitudes elsewhere in the coil. The probability that one of these will cause the coil to go normal increases of course with the number of flux jumps induced, i.e. it increases with the size of the coil. The occurrence of induced flux jumps might also explain why, if repeated attempts to increase the current are made, the wire of the coil does not always go normal at the same place.

Finally, a few words about the training effect. The principal cause is undoubtedly the actual magnetization of the windings themselves, once the current has been switched on at low temperature. A subsidiary cause may be heat generated mechanically, for example by the movement of the windings.

The effect of the magnetization may be described as follows. When a coil is cooled to below the transition temperature and the current is increased until some part of the coil changes to the normal state, then after switching off the current a part of the coil will still remain superconducting. This part is larger the more rapidly the current is switched off. When the current is switched on again and its magnitude increased, the superconducting part is thus already "premagnetized", and therefore no more flux jumps can occur in that part. In the second attempt to raise the current, the probability that too great a flux jump will occur somewhere in the coil is therefore less than in the first attempt, and generally speaking the current reached will be higher, and so on.

If the temperature of a coil is increased to room temperature after it has been used the magnetization disappears, and the same thing happens when the coil is used again. After completing the "training sequence" the same maximum current is then reached. Even the current strength at which the coil goes normal for the first time is often nearly the same.

The need for a training sequence for coils of medium field energy (5 to 10 kJ) and above is a nuisance for the user. During the training sequence the field energy is dissipated several times, which causes considerable evaporation of the helium and may perhaps lead to the risk of damage in some types. The need for special safeguards is therefore even more imperative than might be assumed from the first section.

Experience has shown that the probability of flux jumps causing the coil to go normal increases with the rate at which the value of H is increased. In practice this should not be increased at more than 100 to 300 oersteds per second. This means that a training series can take up a lot of time, particularly if the desired field strength is more than 50 kOe.

Stability of superconducting magnets

What has just been said about the training effect and current degradation makes it clear that, in the absence of special provisions, an energized superconducting coil is in a metastable state. If the coil suddenly goes normal, say as a result of too great a flux jump, the affected region spreads until the current has decreased to a value below I_{ms}. Measures to reduce this effect can be based on various principles. Which principle is adopted, and the extent to which the stability is increased, will depend on the conditions and on the requirements to be met. We can begin by saying that it is possible to make a coil perfectly stable, i.e. not only insensitive to the largest flux jumps likely to be encountered, but even to an overshoot of I_c.

The simplest method is to premagnetize the coil, before current flows in it, by means of an external magnetic field of 10 to 20 kOe, produced by an auxiliary coil. When the field strength is further increased by switching on the current in the main coil and increasing it, there can then be no large flux jumps in the main coil, but only small ones. The degradation effect can be considerably reduced in this way, so that I_c is very nearly reached (*fig. 10*). In practice the auxiliary field is usually generated by means of the outer winding of the magnet itself, which is separately energized for the

purpose. Current degradation does of course occur in this winding (or in the auxiliary coil).

Now let us look at the methods of improving stability which are related exclusively to the *coil materials*. All of these methods are based on coating the superconducting core with a layer of non-superconducting metal. This has two useful consequences. In the first place the heat capacity per unit length is increased and reduces the temperature fluctuation occurring at a given flux jump. In the second place a conductor is introduced, in parallel with the superconductor, into which

Fig. 10. Two superconducting wire-wound coils, which can generate a magnetic field of 90 kOe when the small coil is placed in the bore of the larger one.

most of the current from the core can transfer without excessive heat generation, when the core changes locally to the normal state. Which of the two effects contributes most to the stability improvement depends on the nature of the material used for coating the superconducting core.

To obtain a relatively low resistance when the core is in the normal state, the coating is made of silver, copper or aluminium. For improving the heat capacity a particularly suitable material is indium, which at 4.2 °K has a relatively high specific heat and fairly good electrical conductivity. Other useful materials are lead, bismuth and copper. A copper coating is useful since both its heat capacity and its electrical conductivity are fairly high. Indium is more generally used for impregnating cabled wires than for coating a single wire.

What has been said above applies in particular to wire made from one of the superconducting *alloys* quoted earlier: Nb-Zr, Nb-Ti, etc. The heat capacity of wire and ribbon based on the compound Nb_3Sn is already sufficiently high because of the way in which it is made: Nb_3Sn wire has a niobium coating, which in turn is usually surrounded with a further coating of monel; Nb_3Sn ribbon has a base of stainless steel or niobium. Since this base is less effective than a sleeve and has a fairly high resistance, ribbon is also coated with a high-conductivity outer layer of copper or silver.

As well as improving the wire thermally and electrically in this way, steps are usually taken to aid the conduction of the heat away from the winding turns. Wire-wound coils, for example, have been impregnated with a synthetic resin containing a relatively large amount of aluminium oxide powder, which is a good heat conductor. This has the effect of reducing the current degradation. In our coils of Nb_3Sn ribbon we have improved the conduction of heat from the coil by introducing insulated copper foil between the winding, and we have done the same with wire-wound coils. Large magnet coils are sometimes provided with cooling channels between the layers of the winding.

Practical aspects and considerations

In the stabilizing method in which the heat capacity is increased, known as enthalpy stabilization, the current degradation can be substantially reduced, making it possible to approximate closely to the critical current measured for a sample.

The other method of stabilization gives even better results than enthalpy stabilization. With a sufficiently thick coating the effect of a local flow of current in excess of I_c may be reduced to negligible proportions. Most of the current is transfered locally to the coating and, provided the excess current remains within reasonable bounds, the heat generated can be removed by the liquid helium, so that the normal region does not spread. If the current drops to a value below I_c, the entire conductor returns to the superconducting state. This is described as *full stabilization*.

If the coil is fully stabilized, then in theory it is quite permissible for the superconducting core to have a break in it at some point. This clearly demonstrates that we are dealing with a case on the border-line: a fully stabilized coil, cooled down to 4.2 °K, may also be regarded as a copper or aluminium coil whose resistance has been very considerably lowered because most of the winding has become superconducting at the working temperature. The leading design parameter of such a coil is therefore not so much the critical current density of the superconducting material as the maximum heat transfer to boiling helium. This can be no more than about 0.8 W/cm^2: if the temperature difference between wall and helium becomes greater than that corresponding to this heat transfer, the helium starts to boil in another mode (a transition from nucleate boiling to film boiling takes place), which in fact greatly reduces the heat transfer. The limit in practice is 0.3 W/cm^2.

The method of stabilization and the degree of stability to be chosen in any given case depend not only on the magnitude of the field energy and of the flux jumps, but also on a number of practical considerations and on the particular requirements. In general, full stabilization is required only:

1) Where the magnetic energy is so great (e.g. more than 10^6 joules) that sudden complete dissipation of this energy could be detrimental to the magnet coil or would cause excessive helium evaporating.

2) Where interruptions in operation are not permissible, e.g. because an abrupt change in field strength would be detrimental to the equipment set up in the bore.

3) Where simplicity of construction, minimum safety precautions and minimum operating risks are demanded.

The advantage of full stabilization is gained at the expense of the packing factor, and hence of the effective current density. The maximum field strength is lower than the value that can be achieved using less stabilization and the dimensions of the windings are much larger. This is a particular disadvantage for coils whose inside diameter is small compared with the outside diameter. Where the highest possible field is required, or the dimensions of the windings have to be kept small, the best method is enthalpy stabilization; for large magnets with a field strength of 100 to 150 kOe it is in fact the only solution. The method of stabilization and various other details of five of the world's largest magnets now under construction or completed are given in *Table I*.

In cases where full stabilization is not possible, but where protection of the coil is still important, the use of a copper coating is to be recommended. As we noted earlier this combines the advantages of both stabilizing methods, which is a particularly attractive safety feature.

Some experimental superconducting magnets

All our superconducting magnets are transportable experimental magnets for laboratory use. The inside diameter of most of them is between 16 mm and 116 mm; one coil, which also differs very considerably from the others in design, is much larger and has an internal bore with a diameter of 185 mm. The maximum field strength of the small coils is between 10 and about 90 kOe, and that of the large coil is about 40 kOe.

We shall now discuss some special features in the design and use of the small magnets, and then after looking at some of the safety measures we shall discuss the large magnet.

Most of our experimental superconducting coils were wound from Nb-Zr or from Nb-Ti wire, some from cable and a few from Nb_3Sn ribbon. The Nb-Zr wire used was provided with a thin copper coating. The wire and cable coils were made using the conventional winding and insulation methods used for non-superconducting coils. Special care was taken, however, to ensure that the coils were wound tightly, with no chance of movement between the turns. Care was taken to ensure good insulation between the winding and the coil former.

To reduce current degradation, insulated copper foil has been wound between the layers of some coils, as described above (*fig. 11*). The thermal stability of some coils has been improved even more by connecting low-resistance shunts between the turns in the same layer, for example by means of copper strips arranged parallel to the axis of the coil. This was in fact a necessity in coils wound from narrow Nb_3Sn ribbon. The use of

Table I. Some details of five of the world's largest superconducting magnets now under construction or completed.

Place	Application	H_{max} (kOe)	Length (cm)	Inside diameter (cm)	Field energy (MJ)	Super-conducting material	Stabilization	Stage (Oct. '68)
AVCO Research Lab. (Everett, U.S.A.)	Model magn.-hydr. generator	40	120	30	4.6	75Nb25Zr	full	completed
Argonne National Laboratory (U.S.A.)	Bubble chamber spectrometer	20	304	478	80	Nb-Ti ribbon	full	nearly completed
Brookhaven National Laboratory (U.S.A.)	"	30	183	490	about 725	under investigation	full	under construction
CERN (Geneva, Switzerland)	"	35	420	470	750	Nb-Ti	full	design
NASA (Cleveland, U.S.A.)	Magnetic bottle	75 to 137	330	15-25	—	Nb_3Sn ribbon Nb-Ti	enthalpy full	under construction

shunts does of course increase the time constant of the coil, which may be a disadvantage in some cases, and it also prevents the field strength from being changed rapidly since any change induces a current in the shunts, thus generating heat.

Our coils were usually connected to the supply leads by etching the wire clean and clamping it between small copper terminal blocks. With the copper-coated Nb-Zr wire we also soldered copper ribbon to the coating of the wire. The heat generated at the connection was kept within reasonable limits by making sure that the contact resistance was always less than 10 $\mu\Omega$.

Since the heat conduction of these coils is relatively small, joints between the lengths of wire forming the coil could not be located inside the coil. Although fully superconducting connections can be made by means of certain spot-welding techniques, these too must be located outside the coil, because even a relatively weak magnetic field is sufficient to make them go normal.

In our wire-wound coils the critical current is between 20 and 40 A. A current of this magnitude can be supplied from a storage battery and may be fed to the coil without too much helium being vaporized by the heat generated in the supply leads or by the penetration of heat due to leakage along the leads. The output voltage of the battery can be low, since it does not have to exceed the sum of voltage drop along the supply leads and the voltage induced across the coil when the current increases. As storage batteries have a low internal resistance, a variable series resistance had to be included in the circuit to provide a means of limiting and adjusting the current. In fact, however, we have usually used a stabilized current source. This provides a more constant current, and also does away with the need for a series resistance. To limit the rate of build-up of the field strength, a damping resistance of 0.3 to 1 Ω was shunted across the coil.

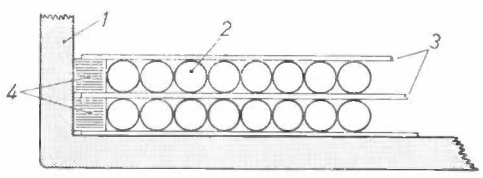


Fig. 11. Method of winding a coil of Nb-Zr wire for an experimental superconducting magnet. *1* coil former. *2* Nb-Zr wire with thin copper coating; outside diameter 0.25 mm. *3* "Mylar" [*] layer. *4* "Mylar" ribbon, insulating the winding from the coil former.

As we noted in the first section of the article the most constant field is obtained by applying a superconducting short-circuit between the ends of the coil. In practice, the terminals of such a short-circuit sometimes give a slight contact resistance which causes a gradual decrease of the current. This is less troublesome for coils of larger inductance. In our smaller coils a time constant of 10^6 to 10^8 seconds is obtained, which is quite sufficient for many experiments. Spot-welded connections give a completely superconducting connection and hence an absolutely constant field.

Coils wound from cable or ribbon require a very high current (100 to 2000 A in our coils) and the current source therefore has to be very much larger than one of the small experimental magnets, and very thick supply cables have to be used. The thick cables are a particular nuisance, since they conduct a good deal of heat into the cryostat and this causes an appreciable helium loss. The problem can be neatly solved by energizing the coil by means of a superconducting dynamo, as described by Volger [6]. This dynamo is mounted

[*] Registered Trade Mark, Du Pont de Nemours.

[6] J. Volger and P. S. Admiraal, Physics Letters **2**, 257, 1962. One of the earlier versions of the dynamo is also described in Philips tech. Rev. **25**, 16, 1963/64. A survey of various possible versions is given in: J. van Suchtelen, J. Volger and D. van Houwelingen, Cryogenics **5**, 256, 1965. See also reference [1].

near the coil in the cryostat, so that coil, dynamo and supply cables from a single superconducting unit.

Fig. 12 shows one of the experimental combined arrangements of this type which we have used [7]. The coil is at the bottom, the dynamo at the top. The maximum current through the coil is about 1400 A, producing a field of about 40 kOe in the bore of the coil (70 mm dia.). The optimum situation is achieved when the cable of the coil winding is just able to carry the maximum current from the dynamo. The required field can then be generated with the minimum number of windings, which means that the coil has the smallest possible inductance and the current therefore reaches its final value in the shortest possible time.

Fig. 12. A superconducting magnet coil (at the bottom) combined with a Volger superconducting dynamo (at the top). The dynamo shaft can be seen at the top of the picture. Between the dynamo and the coil there are two short-circuited windings in which a persistent current starts to flow when the coil produces a field, thus screening the dynamo. The maximum current which the dynamo can supply is 2400 A. The maximum current of the coil is 1400 A (field strength 35 kOe). This current is produced by the dynamo in about 20 minutes. This time can be reduced if the coil and dynamo are optimally matched.

Since the Volger superconducting dynamo is not entirely free from losses, it is not so suitable when a very high power is required. (To energize a coil in one hour to a field energy of 10^8 joules — see Table I — requires a power of 30 kW.) The helium evaporation, with the dynamo in operation, is then greater than when the current is supplied from outside through cables which are cooled in the cryostat by rising helium vapour. It is however useful to include a small superconducting dynamo in the short-circuiting lead of such an arrangement for fine adjustment of the current and for keeping the current constant if there should be any contact resistances in the circuit.

Safety measures

As well as the measures described above for protection of the coils against overheating if part of the coil should go normal, we have in addition adopted three other safety measures of a different nature. These three methods each ensure that not all of the heat is generated inside a limited part of the coil volume. In the first method (uniform distribution) the heat is dissipated over a considerable extent of the coil volume, in the second (internal displacement) the heat is mainly released *outside the coil* but inside the cryostat, and in the third it is mainly liberated *outside the cryostat*.

The heat generation can be distributed uniformly by introducing short-circuited rings or layers of copper foil between the layers or sections of the winding. As soon as the field strength decreases, a high current is induced in each of these short-circuited windings, so that a considerable part of the field energy is dissipated in them. Another very effective method is that of subdividing the coil into a large number of sections, each of which is short-circuited by means of a small resistance (*fig. 13a*). When a resistance occurs in a particular section, the current in that section will decrease and the current in the neighbouring winding sections will increase because of the coupling of flux. The difference between this increased current and the externally supplied current then flows through the parallel resistance. The increase of current in the other sections can cause these in their turn to go normal, so that the energy dissipation takes place simultaneously at a number of places in the coil and also to some extent in the shunt resistances.

The internal displacement of the dissipation is a variant on the method of uniform distribution by means of short-circuited rings. The ring or rings in this case are applied *outside* the coil. In some instances we surrounded the coil with a jacket of pure copper or aluminium, and in others the coil former itself acted as the ring. If the time constant of the rings is long compared with the "natural" quench time of the coil, the field energy is almost completely — and uniformly — dissipated in the ring and transferred by this directly to the helium.

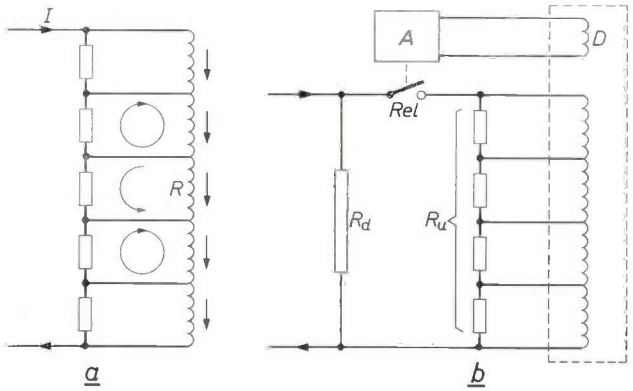


Fig. 13. *a*) Schematic circuit of coil divided into five sections each of which is short-circuited to distribute evenly the heat generated when a part of the coil goes normal. The arrows indicate where and how current starts to flow in the circuit when the part of the coil that has gone normal (*R*) is in the central section.
b) Schematic representation of superconducting magnet whose field energy is mainly dissipated outside the cryostat into a resistance R_u when a part of the winding goes normal. As soon as the field strength begins to decrease, a voltage induced in the detection coil *D* causes the circuit *A* to open the relay *Rel*, thus isolating the coil from the current source. R_n damping resistance.

The displacement of the dissipation to a point outside the cryostat was achieved in a different way. The coils of the magnets are "short-circuited" by a resistance R_u (fig. 13*b*). They are also connected to a relay which isolates the coil from the current source as soon as the field strength in the coil begins to decrease. If R_u is large compared with the resistance of the part of the coil that has gone normal, and provided the relay operates in a time short compared with the natural quench time of the magnet, the greater part of the energy is dissipated in R_u. The switching time of the relay had to be 10 to 100 ms for our cable-wound coils, and about 1 ms for the wire-wound coils.

Set against the very considerable advantages of external dissipation there is the disadvantage that the coil has to be well insulated, because of the self-induced voltage which is initially equal to $I_c R_u$. The value chosen for R_u must therefore be no higher than is necessary to dissipate a good part of the energy of the coil externally. This limit to R_u is fairly sharply defined.

As well as by increasing the value of R_u, the externally dissipated fraction of the coil energy can also be further increased by dividing the coil into sections, each of which has its own "short-circuiting" resistance. When one section goes normal, the energy is dissipated entirely outside the cryostat via the other sections, which are still completely superconducting. An incidental advantage is that these sections have a longer time constant than the section of coil that has gone normal.

With external dissipation of the energy, evaporation of helium from the cryostat is kept low. This enables the experimenter to check the critical current before each experiment and put the coil through a training series. The maximum current then reached is often higher than is obtained with other methods of protection; since only a little heat is generated in the coil on switching off, most of the coil retains its magnetization and no further large flux jumps can take place inside it.

The large magnet

We conclude with some details of the large magnet we have made (*fig. 14*). The coil is wound from seven-

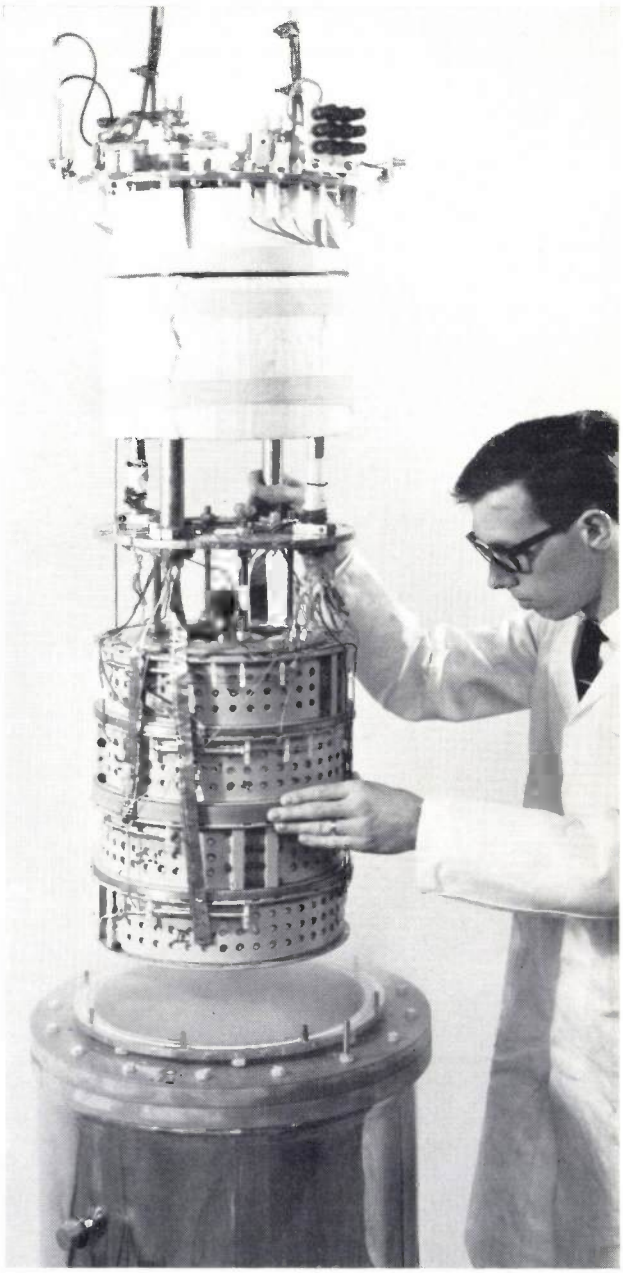


Fig. 14. The coil of the largest superconducting magnet made at Philips Research Laboratories. The inside diameter of the bore is 186 mm. The coil is wound in eight sections arranged in pairs on four coil formers. The maximum current is about 180 A, and the maximum field strength about 40 kOe.

[7] A more detailed description is given in: D. van Houwelingen and J. Volger, Philips Res. Repts. **23**, 249, 1968 (No. 3).

core Nb-Zr cable, in eight sections each containing 750 to 800 metres of cable. The sections are arranged in pairs on four coil formers. At a current of about 180 A this magnet produces a field of about 40 kOe in a cylindrical bore with a diameter of 18.5 cm. The winding is not stabilized and gives a certain amount of current degradation as well as some training effect. The winding is sufficiently well insulated to permit the use of the method mentioned above for the external dissipation of the field energy (about 120 kJ); at 4.2 °K the breakdown voltage is everywhere higher than 2 kV. External dissipation was highly desirable here since 120 kJ of energy set free in the cryostat would vaporize 45 litres of liquid helium!

Temperature fluctuations due to flux jumps are limited by provision of very good thermal contact between the cable and the helium. This is achieved by sheathing the cable with open-spun nylon webbing, by making the insulation between layers of spaced, axially arranged strips of insulating paper, and by providing a large number of holes in the coil former to facilitate the supply of liquid helium and the removal of vapour bubbles.

Great attention has also been paid to the mechanical stability. For example, special clamping devices were designed to make sure that the windings stayed in position. With these three features — mechanical stability, good heat removal and external energy dissipation — it was possible to raise the critical current to the maximum value obtainable with a non-stabilized coil. It is true that we could have reached a higher critical current if we had adopted stabilization with copper, but this would have given a much poorer packing factor, and hence a much larger outside coil diameter. This would have made the magnet less suitable for certain potential applications. The design which we adopted gives an effective current density of no less than 12 000 A/cm^2 at a current of 180 A.

Summary. Superconducting magnet coils for generating very high fields can only be wound from material in which flux can penetrate without destroying the superconductivity, and in which this situation can exist up to very high field strengths and very large currents. Suitable materials include Nb-Zr alloys and Nb_3Sn, in the form of either single or cabled wire, or ribbon. The operating temperature is generally 4.2 °K (the boiling point of helium). Heat generated by sudden local penetrations of flux may cause parts of a coil to go normal before the critical current or field is reached. This can be reduced by increasing the heat capacity per unit length and by ensuring good heat transfer to the helium. It can in fact be eliminated completely by surrounding the wire with a high-conductivity sleeve or coating into which the current transfers locally. Experimental magnets have been made (inside diameter of coil 16 to 116 mm; field strengths up to 90 kOe) which have very good thermal characteristics. The safety measures used include ensuring a more uniform distribution or a displacement of the heat generation by means of short-circuited windings, or dissipation of the field energy outside the cryostat.

Recent scientific publications

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

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

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

H. Bacchi & M. Blanchet (Services Techniques des Armées): Générateurs d'impulsions haute tension de 1 nanoseconde à mi-hauteur.
Onde électr. **48**, 430-437, 1968 (No. 494). *L*

K. H. Beckmann: Methods and results of optical investigations on semiconductor interfaces.
Angew. Chemie: Int. Edit. in English **7**, 161-171, 1968 (No. 3); German Edit. **80**, 213-224, 1968 (No. 6). *H*

P. Beekenkamp: On the structure of glasses in the system K_2O-B_2O_3-Al_2O_3.
Phys. Chem. Glasses **9**, 14-20, 1968 (No. 1). *E*

G. D. Bergman (Associated Semiconductor Manufacturers Ltd., Wembley, England): Semiconductor materials and devices.
Semiconductor devices in power engineering, editor J. Seymour, publ. Pitman, London 1968, chap. 1 (p. 1-36).

P. F. Bongers & G. Zanmarchi: Infrared absorption spectrum and Faraday rotation of ferromagnetic $CdCr_2Se_4$.
Solid State Comm. **6**, 291-294, 1968 (No. 5). *E*

A. H. Boonstra: Some investigations on germanium and silicon surfaces.
Thesis, Eindhoven 1967. *E*

C. R. Borley & L. H. Guildford: A 100 line thermal viewer.
Infrared Phys. **8**, 131-134, 1968 (No. 1). *M*

A. Bril, G. Blasse & J. A. A. Bertens: Measurement of quantum efficiencies of Eu^{3+}-activated phosphors using excitation to selected Eu^{3+}-levels.
J. Electrochem. Soc. **115**, 395-397, 1968 (No. 4). *E*

H. B. G. Casimir: Maatschappij, wetenschap en wetenschapsbeleid.
Ingenieursblad **37**, 383-388, 1968 (No. 10). *E*

H. B. G. Casimir: Betrachtungen über integrierte Schaltungen und Mikrotechnik.
Tagungsbroschüre VDE-Fachtagung Elektronik 1968, 6 pp. *E*

C. Crevecoeur & H. J. de Wit: Thermoelectric force versus electrical resistivity of MnO at 1100 °C. An example of "Jonker's pear".
Solid State Comm. **6**, 295-296, 1968 (No. 5). *E*

K. Decker (Nervenklinik der Universität München), **H. O. Lincke** (Nervenklinik U. M.) **& W. J. Oosterkamp:** Die Beurteilung von Röntgenbildern und ihre Erleichterung durch Farbsubtraktion.
Deutscher Röntgenkongress 1967, Baden-Baden, Teil A, p. 78-79; Thieme, Stuttgart 1968. *E*

J. Drenth (Rijksuniversiteit Groningen), **W. G. J. Hol** (R.U. Gr.), **J. W. E. Visser** (R.U. Gr.) **& L. A. Æ. Sluyterman:** Papain in water-rich and in methanol-rich media; crystallization and conformation.
J. mol. Biol. **34**, 369-371, 1968 (No. 2). *E*

W. F. Druyvesteyn: Size-effects in the magnetoresistance of Al and In films.
Phil. Mag. **18**, 11-26, 1968 (No. 151). *E*

G. Engelsma: The influence of light of different spectral regions on the synthesis of phenolic compounds in gherkin seedlings, in relation to photomorphogenesis, IV. Mechanism of far-red action.
Acta bot. neerl. **17**, 85-89, 1968 (No. 2). *E*

U. Enz & H. van der Heide: Two new manifestations of the photomagnetic effect.
Solid State Comm. **6**, 347-349, 1968 (No. 6). *E*

G. Eschard & R. Polaert: Tube obturateur pour photographie ultra-rapide.
Onde électr. **48**, 426-429, 1968 (No. 494). *L*

C. Fosseprez: La structure et le fonctionnement des calculatrices digitales. Introduction à leurs applications. 1re partie: Les organes essentiels.
Rev. MBLE **10**, 111-122, 1967 (No. 3). *B*

B. Frank & J. Liebertz: Glasbildung in den Systemen $CaO\text{-}Al_2O_3$, $SrO\text{-}Al_2O_3$ und $BaO\text{-}Al_2O_3$ durch Abschrecken schmelzflüssiger Tröpfchen.
Glastechn. Ber. **41**, 253-254, 1968 (No. 6). *A*

J.-M. Goethals & P. Delsarte: On a class of majority-logic decodable cyclic codes.
IEEE Trans. **IT-14**, 182-188, 1968 (No. 2). *B*

J. Haisma & K. Roozendaal (Philips Industrial Equipment Division, Eindhoven): Investigation of the behaviour of an expansion ejector in the low temperature region beyond the λ-transition of helium.
Proc. XIIth Int. Congress of Refrigeration, Madrid 1967, paper No. 1.37, 9 pp.; 1968. *E*

J. E. Knowles: The simulation of domain wall motion in square loop ferrites, and other aspects of wall behaviour.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 821-830, 1968 (No. 7). *M*

E. Laviron (Commissariat à l'Energie Atomique), **C. Delmare** (Comm. En. At.) **& H. Bacchi:** Caméras électroniques de 1 ns de durée d'ouverture.
Onde électr. **48**, 421-425, 1968 (No. 494). *L*

J. C. Mackellar (Mullard Radio Valve Co.): Semiconductor switching techniques and their application to power control.
Semiconductor devices in power engineering, editor J. Seymour, publ. Pitman, London 1968, chap. 3 (p. 63-85).

K. Mouthaan & W. Fulop (Brunel University, London, Acton, England): On diffusion of electrons across a magnetic field.
J. appl. Phys. **39**, 2021-2023, 1968 (No. 4). *E*

B. J. Mulder: Diffusion and surface reactions of singlet excitons in anthracene.
Thesis, Eindhoven 1967. *E*

J. Neirynck: La collaboration entre l'industrie et l'Université en matière de recherche.
Rev. MBLE **10**, 105-110, 1967 (No. 3). *B*

C. B. Oliver (Associated Semiconductor Manufacturers Ltd., Wembley, England): Degradation of evaporated aluminium contacts on silicon planar transistors.
Proc. 1967 Reliability Physics Symp., Los Angeles, p. 209-215; 1968.

L. G. Pittaway, J. Smith, E. D. Fletcher & B. W. Nicholls: Laser-induced electron emission from tungsten, tantalum and lanthanum hexaboride using ruby and argon lasers.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 711-719, 1968 (No. 6). *M*

J. Polman: Resonanties in een inhomogene plasmakolom.
Ned. T. Natuurk. **34**, 102-111, 1968 (No. 4). *E*

C. J. M. Rooymans: Structural investigations on some oxides and other chalcogenides at normal and very high pressures.
Thesis, Amsterdam 1967. *E*

E. Tscholl: The photochemical interpretation of slow phenomena in cadmium sulphide.
Thesis, Eindhoven 1967. *E*

Contents of Electronic Applications **28**, No. 2, 1968:

Contents of Valvo Berichte **14**, No. 2, 1968:

Volume 29, 1968, No. 10 pages 293-324 Published 31st December 1968

PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 11

An experimental light-weight colour television camera

P. H. Broerse, J. H. T. van Roosmalen and S. L. Tan

The article below describes the portable colour television camera which was illustrated in this journal a few months ago. The construction of such a small camera has been made possible by the development of an (experimental) "Plumbicon" [*] camera tube which is smaller than the "Plumbicon" tubes used in the studio camera. The pictures taken with this miniature camera compare well in quality with those taken with the studio camera at present in use.

The days when television cameras were only used in studio or outside broadcast work now seem very distant. Quite early on, it became apparent that television would be of service in industry, in education, and for many other purposes. A striking example in medical education is the televising of surgical operations so that they can be seen by large numbers of students [1]. A more recent application is the use of a television camera in conjunction with an endoscope, enabling doctors to look at a patient's internal organs on a television screen. The use of closed-circuit television is also becoming more and more widespread in other areas of education, in research work, and in industry. One interesting example is the use of a television camera in conjunction with a microscope, an arrangement which can be very useful in checking the manufacture of integrated circuits.

In some of the cases mentioned above, the information given by black-and-white television is sufficient. Often, however, colour can be vitally important; this is particularly so in medical applications [2]. However, this is just the sort of application where the considerable bulk of the ordinary colour camera makes things difficult. What is needed is a small and light-weight camera that will not get in the way during an operation. Such a camera would also be very suitable for use in conjunction with other instruments, and in television broadcasting a small portable camera for colour television could create a wealth of new possibilities.

With these applications in mind we have developed an experimental colour camera whose size and weight have been reduced to the minimum that present-day techniques will allow.

This camera, which weighs three kilogrammes and compares in size with a 16 mm cine camera, was demonstrated to professional users a few months ago [3]. *Fig. 1* shows the camera connected to a microscope for checking an integrated circuit. In this article we shall discuss a number of problems that were encountered in the development of the special pick-up tubes for this camera and in the construction of the camera itself.

The camera tube

A first requirement in the development of the new camera was that as far as possible it should have the same good characteristics as the larger camera with "Plumbicon" tubes which was described earlier in this journal [4]. However, the dimensions of a colour television camera containing three camera tubes are largely

Drs. P. H. Broerse, J. H. T. van Roosmalen and Ir. S. L. Tan are with Philips Research Laboratories, Eindhoven.

[*] Registered Trade Mark for television camera tubes.

[1] See Philips tech. Rev. **11**, 42, 1949/50.

[2] W. Holm and F. H. J. van der Poel, Colour television in medical teaching, Philips tech. Rev. **20**, 327-330, 1958/59.

[3] Philips tech. Rev. **29**, 188, 1968 (No. 6).

[4] See H. Breimer, W. Holm and S. L. Tan, A colour television camera with "Plumbicon" camera tubes, Philips tech. Rev. **28**, 336-351, 1967 (No. 11).

determined by the size of the tubes and their deflection coils. Before a miniature camera could be built it was therefore first of all necessary to make a "Plumbicon" camera tube which was much smaller than the tubes used previously, without sacrificing quality. The standard "Plumbicon" tube has a diameter of 30 mm and a length of 200 mm; in the miniature version these dimensions are 16 and 130 mm respectively. The image size of 8.4×6.3 mm (diagonal 10.5 mm) is scaled down by a factor of two from that of the standard tube. *Fig. 2* shows the two "Plumbicon" tubes with and without their deflection coils.

The design of the miniature "Plumbicon" tube is basically identical with that of the larger one described in reference [5]. The main problems presented by the reduction of dimensions were connected with the maintenance of a sufficiently high resolving power and with the accurate register of the images from the three tubes in the camera. We shall deal with these points at greater length in the following sections.

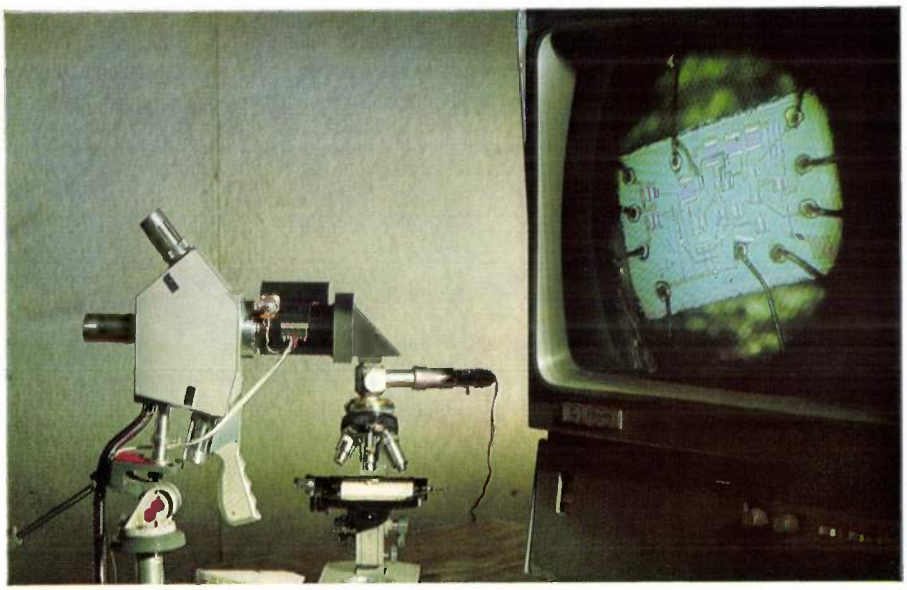


Fig. 1. The miniature camera connected directly to a microscope. The monitor is showing a picture of an integrated circuit which is placed under the microscope.

The electron beam is focused electrostatically. The advantages of electrostatic over magnetic focusing have already been explained in an earlier article describing a tube of standard dimensions [5]. Where space and weight are at a premium, it is obviously a great advantage to be able to dispense with bulky focusing coils. But this is not the only virtue of electrostatic focusing. Other advantages are that the picture is not rotated, magnified or reduced in size when the focus is adjusted, and moreover the deflection cannot affect the focusing, so that a sharp picture is easily obtained over the whole surface. Another important feature of electrostatic focusing is that it requires no power.

Fig. 3 shows a schematic cross-section of the tube. The electron beam is delivered by a *triode gun Tr*, which consists of an impregnated cathode *K* (rated at 2 W), a control electrode or "grid" *G* and an anode A_1. The divergent beam is made to converge on the photoconducting target *T* by a focusing lens, formed by cylinders A_1, A_2 and A_3; cylinder A_1 contains a dia-

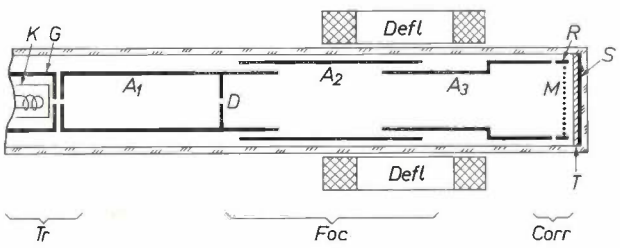


Fig. 3. Diagram of the construction of a "Plumbicon" tube with electrostatic focusing. *Tr* triode gun, consisting of the cathode *K*, the control electrode *G* and the anode A_1. The focusing lens *Foc* is formed by the electrodes A_1, A_2 and A_3. *D* diaphragm. The ring *R* and the mesh screen *M* together with A_3 form the correction lens *Corr*. *T* photoconducting layer. *S* signal plate. *Defl* deflection coils.

[5] J. H. T. van Roosmalen, Experimental electrostatically focused "Plumbicon" tubes, Philips tech. Rev. **28**, 60-66, 1967 (No. 2).

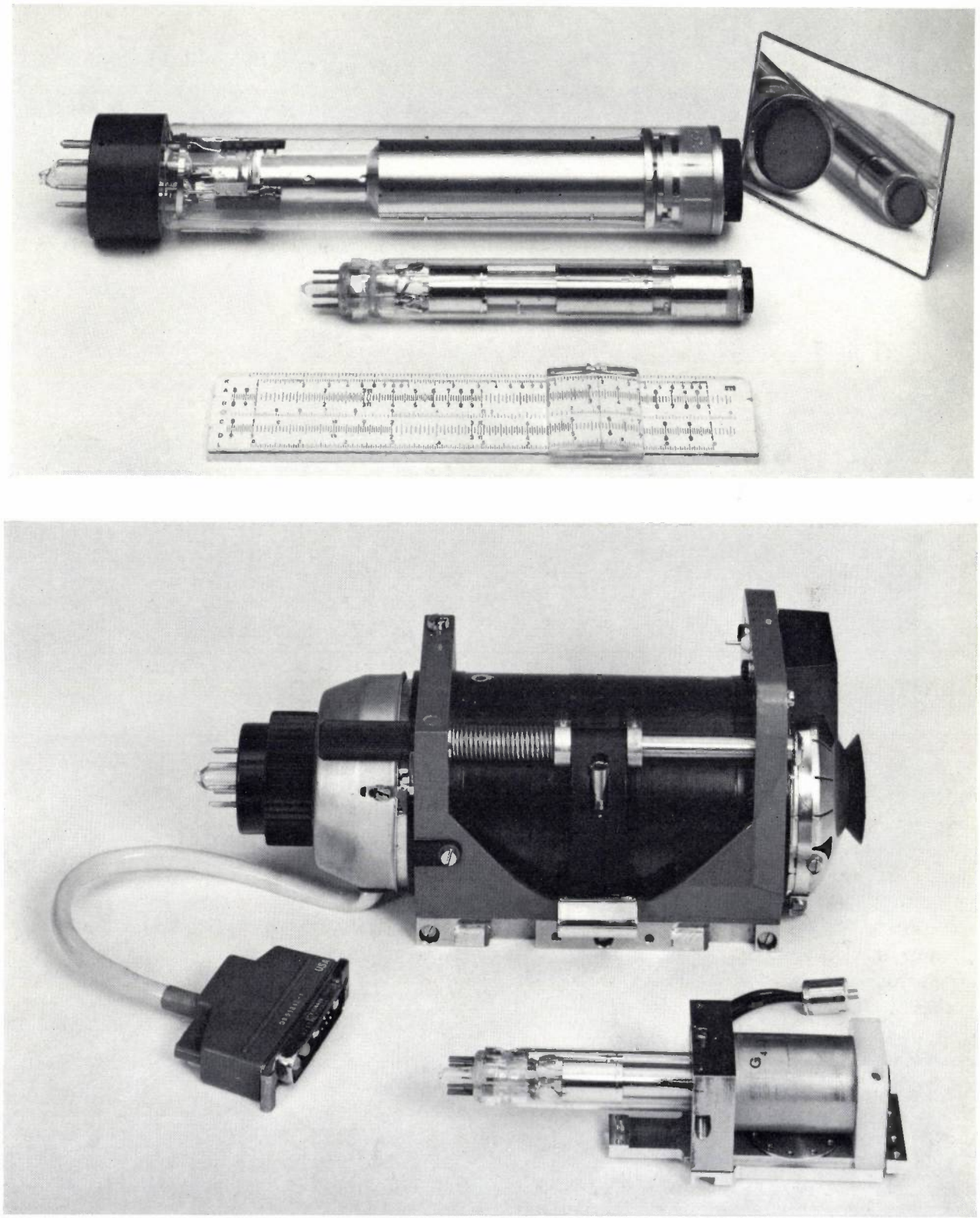


Fig. 2. *a*) Standard and miniature "Plumbicon" tubes. The dimensions of the tubes are 200 × 30 mm and 130 × 16 mm. *b*) The tubes mounted in their deflection coils.

phragm *D*. Following the focusing lens there is a *correction lens*. This consists of the right-hand part of A_3, a ring *R* and a mesh *M*. The potential of *R* and *M* is higher than that of A_3. This lens deflects the electron beam in such a way that the electrons arrive almost at right angles to the photoconducting layer. A description of the undesirable effects which occur when this condition is not met will be found in reference [5].

Focusing

The lens action of K, G and A_1 focuses the emitted electrons to the "crossover" and an image of the crossover area is formed on the photoconducting layer by the focusing lens. For good resolution this image must be small, and this requirement is of course accentuated when the image size of the tube is smaller.

Just as in the development of the large "Plumbicon" tube with electrostatic focusing, the dimensions, the mutual alignment and the potentials of the electrodes were fixed after calculating electron paths for different configurations. *Fig. 4* shows a number of equipotential surfaces for the final version, together with some electron paths for the undeflected beam.

deflection increases and there is a greater danger of breakdown between R and A_3. For a small d_M it follows from the equation that the beam current I_b has to be small, but if it is too small the photoconducting layer will not discharge fast enough, i.e. the speed of response will suffer. Moreover, the maximum permissible current density j_0 and the temperature T_k are fixed for the type of cathode employed. Finally, if we make α_M large, we introduce difficulties because of spherical aberration in the focusing lens and because of defocusing action of the deflection coils.

The effect of the *spherical aberration* of the focusing lens is that the electrons are scattered over a circle of confusion. The diameter Δd_M of this circle at the mesh

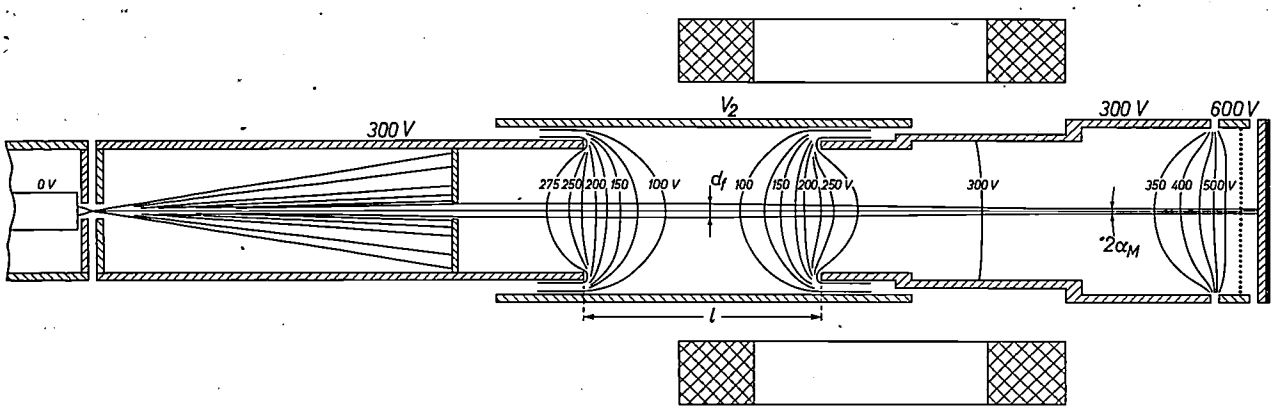


Fig. 4. Some equipotential surfaces and electron paths for an undeflected beam. The voltages at the electrodes A_1, A_3 and R (see fig. 3) are given with respect to the cathode. The voltage V_2 of A_2 is used for focusing the beam. d_f beam diameter at the centre of the focusing lens. $2\alpha_M$ aperture angle of the beam at the mesh. l spacing between A_1 and A_3.

The size of the image of the crossover area can be adversely affected by several factors connected with the electron-optical section. These are the thermal spread of the electrons, the spherical aberration of the focusing lens and the mutual repulsion of the electrons.

The effect of the *thermal spread* is that the beam has a diameter of

$$d_M = \frac{2}{\alpha_M}\sqrt{\frac{I_b k T_k}{\pi j_0 e V_M}}$$

at the mesh. Here α_M is half the aperture angle of the beam at the mesh, V_M is the potential of the mesh, I_b is the beam current beyond the diaphragm, j_0 is the current density at the centre of the cathode, T_k is the cathode temperature; k is Boltzmann's constant and e is the electronic charge. We wish to choose the quantities in this equation so as to make d_M as small as possible, but there is a limit for each quantity. If the voltage V_M is raised, the energy required for the deflection increases and there is a greater danger of breakdown between R and A_3. For a small d_M it follows from the equation that the beam current I_b has to be small, but if it is too small the photoconducting layer will not discharge fast enough, i.e. the speed of response will suffer. Moreover, the maximum permissible current density j_0 and the temperature T_k are fixed for the type of cathode employed. Finally, if we make α_M large, we introduce difficulties because of spherical aberration in the focusing lens and because of defocusing action of the deflection coils.

The effect of the *spherical aberration* of the focusing lens is that the electrons are scattered over a circle of confusion. The diameter Δd_M of this circle at the mesh is proportional to the cube of d_f, the diameter of the beam at the centre of the focusing lens:

$$\Delta d_M = B d_f^3.$$

Obviously, we want to keep the aberration coefficient B as small as possible. A configuration giving an acceptable spherical aberration was again found by calculating the electron paths. The results of a number of such calculations are summarized in *fig. 5*, where B is shown as a function of the distance l between A_1 and A_3. It can be seen that B decreases with increasing l. Since a larger l would require a longer tube, some limitation was necessary and we chose $l = 16$ mm.

It is easier to meet the various conflicting requirements when a diaphragm D is placed in the anode A_1 (see fig. 3 and fig. 4), since it is then possible to use a lower ratio between I_b and j_0. We shall return presently to the problem of deciding the position and size of this diaphragm.

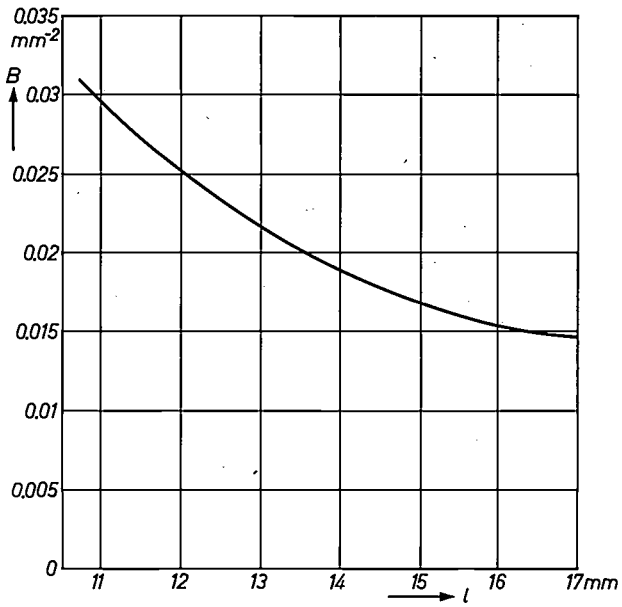


Fig. 5. Spherical-aberration coefficient B as a function of the spacing l between anodes A_1 and A_3.

The correction lens

The potential of the mesh M and the ring R is twice as high as that of the cylinders A_1 and A_3 (600 and 300 V respectively). If the deflection coils are positioned in such a way that the distance between the deflection point (the apparent point of origin of the deflected beam) and the mesh is about twice the diameter of the mesh, then the correction lens brings the deflected beam very nearly parallel again with the axis of the tube [6]. The barrel distortion which this causes is compensated by suitable design of the deflection coils. *Fig.6* shows the central trajectory of the beam in a number of deflected positions.

Because of the smaller image on the photoconducting layer, the mesh used in the standard tube, with a pitch of 36 μm, had to be replaced by a mesh with a pitch of only 18 μm. The wire thickness is 5 to 6 μm after an etching process. This finer mesh allows about 45% of the electron beam to pass through it.

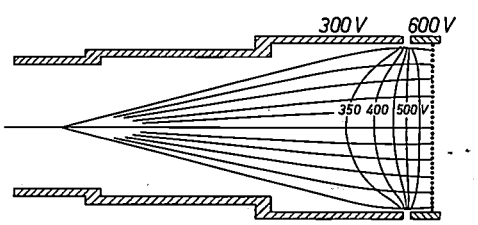


Fig. 6. Central trajectory of the beam in a number of deflected positions. The correction lens makes the electrons "land" at right angles to the photoconducting layer.

The aperture angle of the beam

An important point in the design was the establishment of the position and size of the diaphragm. Moving the diaphragm away from the crossover, while keeping its diameter constant, makes the aperture angle smaller and so makes the beam narrower. An advantage of a narrow beam is that it reduces the spherical aberration of the lens and the defocusing action of the deflection coils. On the other hand, at constant current it has the effect of increasing the mutual repulsion of the electrons. There should therefore be an optimum aperture angle of the beam which gives the best resolving power. This has been verified in experiments with a number of tubes in which a movable diaphragm could be located anywhere between 6 and 21 mm from the crossover. (The diameter of the aperture was 0.35 mm.) This made it possible to vary the beam diameter at the centre of the focusing lens between 2.4 and 0.7 mm. Some results of the measurements made with these tubes are collected in *fig. 7*. The figure shows the

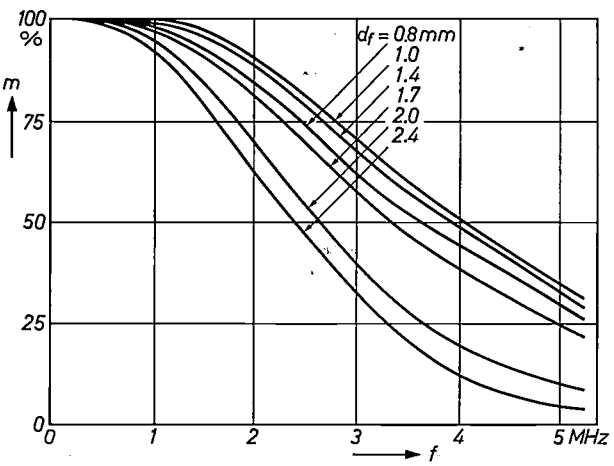


Fig. 7. Modulation depth m as a function of frequency f for some values of the diameter d_f of the beam at the centre of the focusing lens. An optimum value of m is achieved at $d_f = 1.0$ mm.

modulation depth m [7] as a function of frequency, with d_f the diameter of the beam at the centre of the lens as a parameter. We see that the value of d_f for which m is greatest is approximately 1.0 mm. At a frequency of 5 MHz the modulation depth is then 35%.

Fig. 7 shows the modulation depth at the centre of

[6] J. H. T. van Roosmalen, Dutch patent application No. 6513751.

[7] An account of the way in which this quantity expresses the resolving power is given in 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, in particular fig. 17.

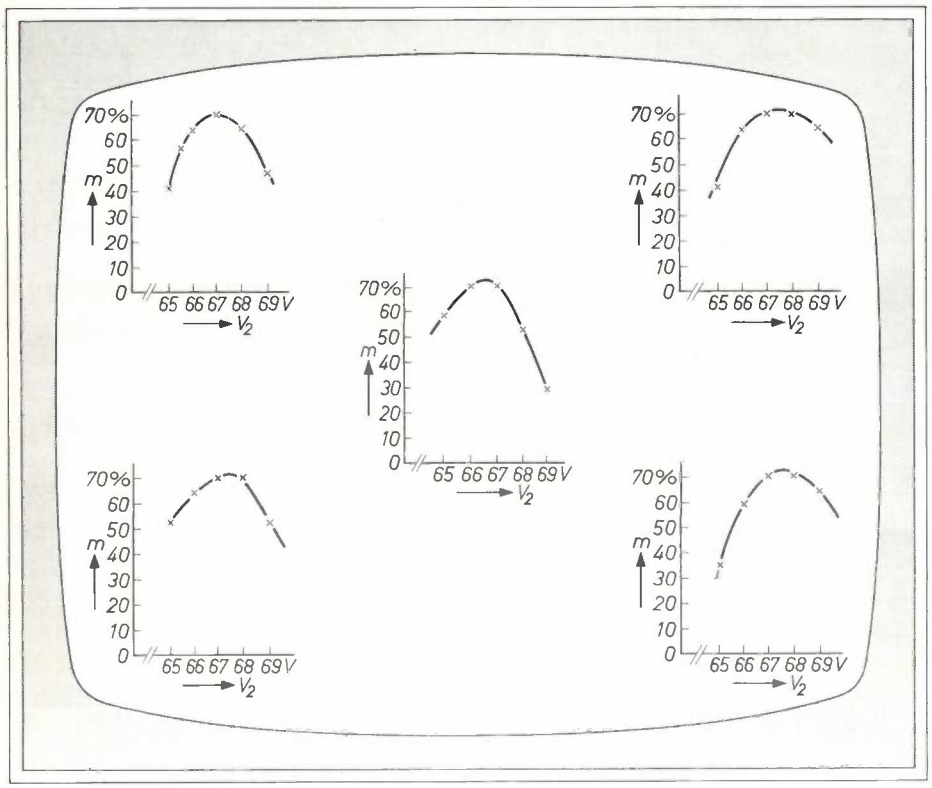


Fig. 8. Modulation depth m as a function of the voltage V_2 on electrode A_2, measured with white light, at the centre and at the four corners of the picture. The frequency used was 3 MHz and the beam current 0.2 μA. At $V_2 \approx 67$ V the resolving power is optimum over the whole area.

the screen. How good is the resolving power at other parts of the screen? This can be seen in *fig. 8* which gives plots of the modulation depth against V_2 (the voltage at A_2) for the centre and the corners of the screen; the modulation depth is measured with white light and at a frequency of 3 MHz. It can be seen that at the same voltage of about 67 V the optimum value $m = 70\%$ is almost reached at all places. This indicates that the pictures from the miniature tubes are equally sharp over the whole area.

The photoconducting layer

Reducing the image size and changing the thickness of the photoconducting layer affects the sensitivity, the resolving power and the lag of a camera tube.

The *sensitivity* and the *resolving power* depend on the wavelength of the light. This is due to the absorption of the light in the layer. In a "Plumbicon" tube blue light at a wavelength of 400 nm is fully absorbed in the first few microns of the layer, whereas only about 10% of red light at a wavelength of 620 nm is absorbed in the full 20 μm thickness of the layer. This makes the tube less sensitive to red light. Since the layer is polycrystalline there is scattering, and this is most pronounced for red light because of the lower absorption, thus giving a lower resolving power. Any reduction in image size must therefore be accompanied by a reduction in the thickness of the layer. Since this means that less red light is absorbed, however, the red sensitivity is reduced at the same time.

Recent developments have enabled us to improve both the resolving power and the red sensitivity by increasing the absorption of the layer for red light. Layers with increased red absorption have already been used in experimental "Plumbicon" tubes of standard size [8]: they have been made by incorporating a very small quantity of lead sulphide in the lead monoxide layer. Lead sulphide, which is also used in infra-red photocells, absorbs radiation at wavelengths extending to 3 μm. A small concentration of it in the lead monoxide layer increases the absorption of red light to such an extent that both sensitivity and resolving power are substantially improved. This made it possible to reduce the image area of the miniature tube by a factor of 4 and the thickness of the layer by a factor of 2. The sensitivity of a miniature "Plumbicon" tube at 620 nm is 0.11 μA/μW as compared with 0.04 μA/μW for a standard tube [9]. The resolving power of the miniature tube also compares very well with that of the standard tube: the modulation depth at a frequency of 5 MHz is 35% for the miniature tube as we noted earlier and 42% for the standard tube.

The *lag* of a "Plumbicon" tube is greatest at low light levels, mainly because of the beam-current lag. This lag occurs because the capacitance of the layer is charged via the beam, which has a certain effective resistance. Other factors involved here, apart from the capacitance of the photoconducting layer, are the signal current, the dark current and the cathode temperature. If a change in the incident luminous flux at the

[8] E. F. de Haan, F. M. Klaassen and P. P. M. Schampers, An experimental "Plumbicon" camera tube with increased sensitivity to red light, Philips tech. Rev. **26**, 49-51, 1965.

[9] The standard tube referred to is type XQ 1020, which has a lead-monoxide layer as described in reference [7]. This tube has been accepted universally for studio use.

[10] See reference [7], page 147.

time $t = 0$ suddenly causes the current through the layer to change from i_1 to i_2, the current i_a supplied to the photoconductor by the beam may be described as a function of time as follows [10]:

$$i_a(t) = \frac{i_2}{1 + \left(\frac{i_2}{i_1} - 1\right) \exp\left(-\frac{ti_2}{CV_0}\right)}.$$

In this expression C is the capacitance of the layer and V_0 is a constant which is proportional to the cathode temperature. At a sudden transition from dark to light i_1 is the dark current and $i_s = i_a - i_1$ is the signal current; at a transition from light to dark, i_2 is the dark current and $i_s = i_a - i_2$ is the signal current.

filament reaches the lead monoxide layer through the glass envelope of the tube, so that there is some illumination at the layer even when there is no image projected on it. This effect is more pronounced in the miniature tube than it is in the standard tube.

Fig. 9 shows the relative variations of the signal current as a function of time for an abrupt transition from dark to light. Curve *a* relates to the standard tube ($i_1 = 0.5$ nA) and curve *b* to the miniature tube ($i_1 = 1.5$ nA). These curves were not calculated with the same signal current for the two tubes. In fact, as we shall see when we come to discuss the camera, it has been found possible to reduce the noise level to such an extent that the miniature tube can be operated at a

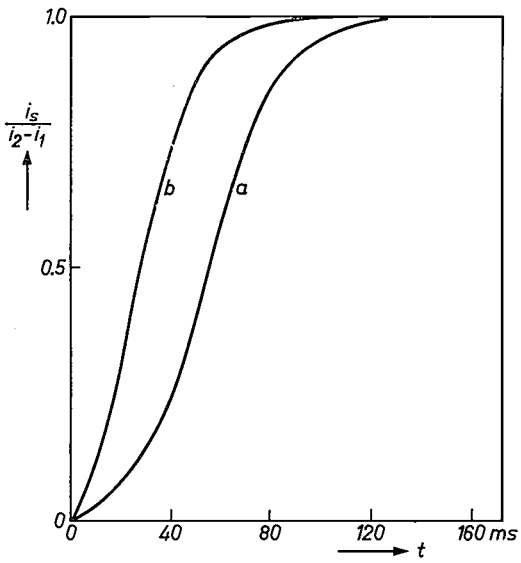


Fig. 9

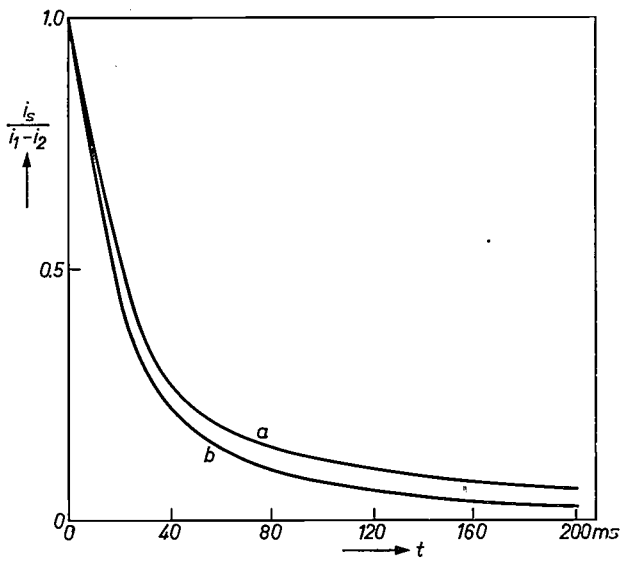


Fig. 10

Fig. 9. Relative variation of the signal current i_s with time t at a sudden transition from dark to light. The current through the photoconducting layer changes from i_1 (dark current) to i_2. The curve *a* refers to the standard "Plumbicon" tube (dark current 0.5 nA, capacitance of the photoconducting layer 1 nF), the final value of the signal current being taken as 20 nA. Curve *b* refers to the miniature tube (dark current 1.5 nA, capacitance 0.5 nF) and has been drawn for a final signal current of 10 nA.

Fig. 10. Relative variation of the signal current i_s with time t at a sudden transition from light to dark, *a* for the standard "Plumbicon" tube, *b* for the miniature tube.

It follows from the above expression that if i_1, i_2 and C are decreased by the same factor, the relative variation of i_a as a function of time does not change and neither therefore does the beam-current lag. Now in view of the above-mentioned scaling-down of the surface area and thickness of the photoconducting layer, one might expect that both the capacitance and the dark current of the miniature "Plumbicon" tube would be twice as small as in the standard tube. It turned out, however, that although the capacitance of the miniature tube was a factor of 2 smaller, the dark current was a factor of 3 higher (1.5 nA compared with 0.5 nA). The reason for this is that light from the signal current which is about half that of the standard tube. We have therefore assumed a signal current of 20 nA for the standard tube and 10 nA for the miniature tube. From fig. 9 it can be seen that the miniature tube has a smaller lag than the standard tube. This is partly due to the smaller capacitance and partly to the higher dark current.

Fig. 10 shows the results of corresponding calculations for an abrupt transition from light to dark. The smaller lag of the miniature "Plumbicon" tube is even more apparent here than in fig. 9. (The horizontal distance between the curves is particularly large in the lower part of the figure.)

The camera

During the first phase in the development of the small camera described here, we wanted to make use of the existing control unit of the studio camera [4]. To keep down the dimensions of the camera unit, however, it was decided that only the circuits which of necessity had to be near to the camera tubes would be included in it. The rest of the circuits were accommodated in a separate unit, the "intermediate unit" (see *fig. 11*), which is connected to the camera unit by a short cable. The intermediate unit is connected to the control unit by the same cable as is used with the normal studio camera.

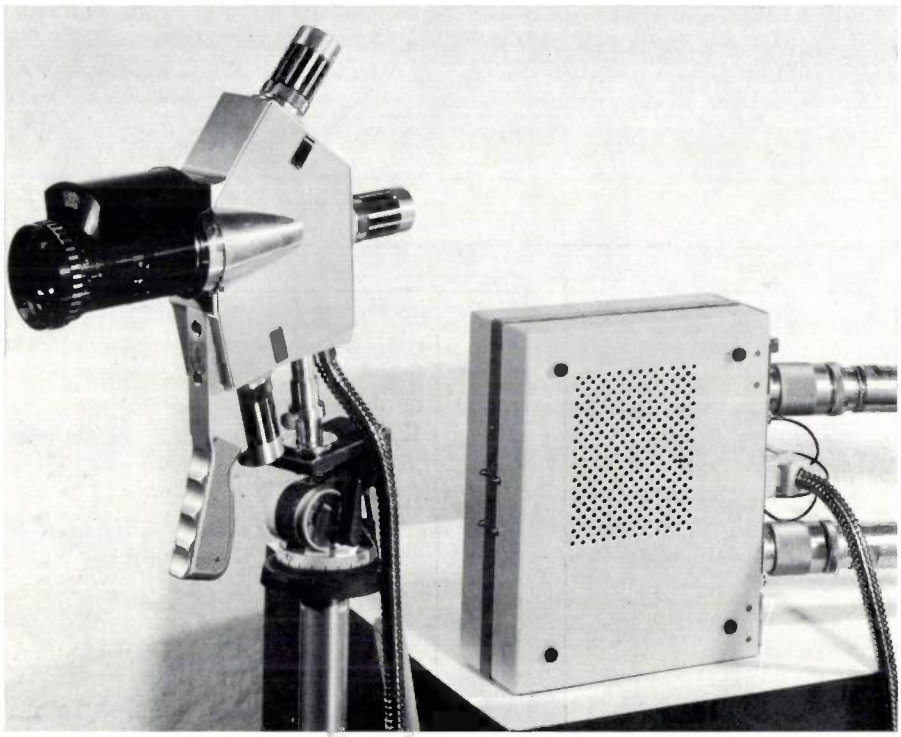


Fig. 11. The miniature camera with the "intermediate unit", which measures 28 × 24 × 10 cm.

Fig. 12 shows one side of the opened camera. The prismatic colour-separation system and the three camera tubes with their deflection coils can be seen in the picture; the three preamplifiers, the blanking amplifier and other circuits required for the operation of the "Plumbicon" tubes are on the other side of the camera.

The optical system

The optical section of the camera is basically the same as that of the studio camera. As in the larger camera the three camera tubes are not mounted with axes parallel. This arrangement requires good screening to keep out external magnetic fields, and this is much more easily achieved with the smaller dimensions of the camera.

The *camera lens* (a zoom lens) was specially developed for the small image format; the focal distance can be varied from 13 to 65 mm. The relative aperture is f/1.6. Because of the smaller image size, the depth of focus at the widest aperture is just as large as in the studio camera, whose lens is set to a relative aperture of f/3.2. The prismatic colour-separation system is adapted to the small dimensions and to the large lens aperture; the angles which the reflecting surfaces make with the optical axis have been made a little larger to avoid the edges of the picture being cut off (vignetting effect).

Image register

In the images which the colour-separation system projects on to the three "Plumbicon" tubes, the points corresponding to one point in the scene must be scanned simultaneously. The accuracy with which this is done depends on many factors. The requirements which this sets for the accuracy and stability of the mechanical construction are much tighter with the smaller image size than for the standard tubes in the studio camera. The camera housing was therefore milled from a single piece of metal. The deflection coils were carefully matched with one another and, as in the large cameras, they can be accurately aligned with respect to the optical axis by means of adjusting screws. To ensure that no errors are introduced by the camera-

tube tolerances, this alignment is carried out with accurately machined acrylic-resin tubes mounted in place of the "Plumbicon" tubes during the alignment procedure. The "Plumbicon" tubes are then placed in the camera and adjusted with screws until their electron-optical axes coincide with the axes of the optical system. If this were not done, differences in the geometry of the three tubes would lead to difficulties in the simultaneous scanning of the three images. Moreover, the three images might be displaced with respect to one another, resulting in less effective use of the surface of the sensitive layer.

Several other precautions have been taken to improve the registration of the three images, such as the provision of variable inductors and resistors to correct any differences in the impedance of the three deflection coils, which are connected in parallel.

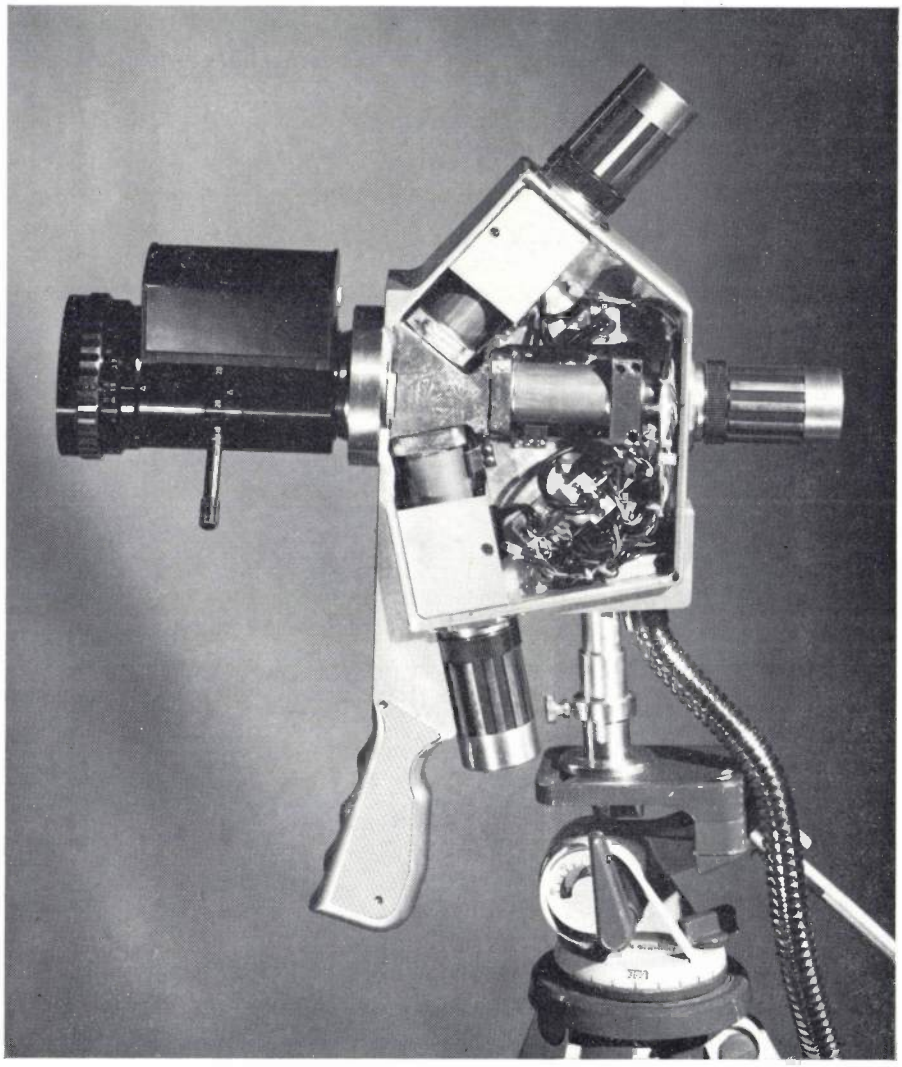


Fig. 12. The camera with a side-plate removed, showing the prismatic colour-separation system and the three "Plumbicon" tubes with their deflection coils. Some of the electronic circuits are on the other side.

Sensitivity

The sensitivity of the "Plumbicon" camera is limited on one hand by the amount of lag that can be allowed with moving pictures and on the other hand by the noise. In the studio camera the use of a field-effect transistor in the input stage of the signal amplifier gives such a low noise level that in practice the sensitivity of the camera is determined by the lag. The lag of the miniature "Plumbicon" tubes, however, is so much smaller that the noise is practically the only limiting factor here.

A large signal-to-noise ratio requires a high input impedance in the signal amplifier. This impedance is formed by the parallel connection of the signal resistance and the capacitances of the signal plate, the amplifier and the wiring. Because of the small dimensions of the signal plate and an extremely short connection

between signal plate and amplifier, the total capacitance has been kept smaller than 10 pF. This low capacitance, coupled with the use of a selected field-effect transistor with a particularly low noise contribution for the input stage of the amplifier, has enabled a signal-to-noise ratio 4 to 6 dB greater than that of the studio camera to be obtained at the same signal current.

Partly because of their reduced lag the tubes in the small camera can therefore be operated with a signal current two or three times lower than the standard "Plumbicon" tubes in the larger camera. A scene illumination of 400 to 500 lux is sufficient to obtain good quality pictures at the maximum lens aperture of f/1.6.

If a camera is to be used in conjunction with other optical instruments, the problem of the optical coupling of the instruments assumes vital importance. In such applications there may often be a considerable loss in luminous intensity and hence in sensitivity. This is the case, for example, when a large camera is to be used with an endoscope, necessitating the use of a flexible optical coupling. The fact that the miniature camera can be connected *directly* to the endoscope thus allows a significant gain in sensitivity.

Image definition

Because of the good resolving power of the miniature "Plumbicon" tubes, the definition is surprisingly sharp over the whole picture surface. The performance of the studio camera and the miniature camera can be compared by means of the experiment illustrated in *figs. 13* and *14*. As in the studio camera, the definition of contours in the picture can be further improved by means of contour signals derived from the green colour signal[11].

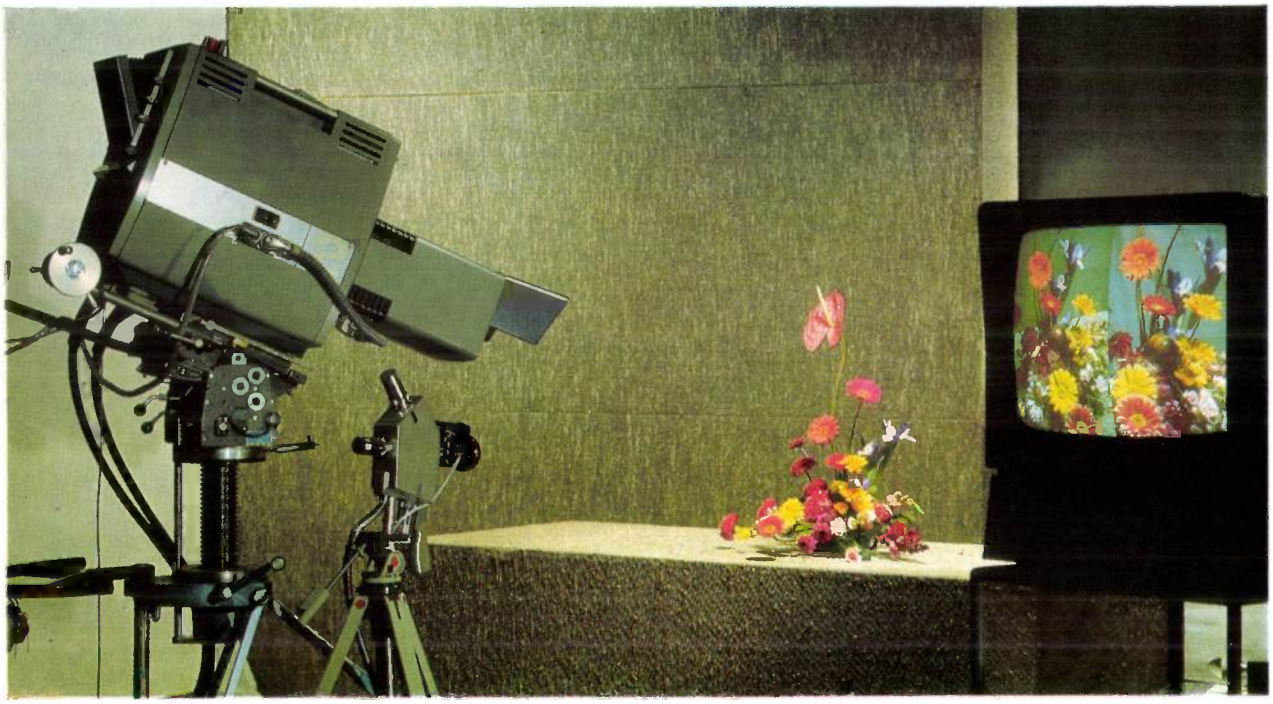


Fig. 13. The studio camera described in reference [4], and the miniature camera, both equipped with "Plumbicon" tubes. The "intermediate unit" (fig. 11) associated with the miniature camera is not shown. Both cameras are taking pictures of the same object to give a comparison of the picture quality: an electronic switching circuit permits the two pictures to appear side by side on the monitor screen at the right.

It is not possible to reproduce the colours of the scene and the monitor picture correctly on the same type of colour film. A better reproduction of the monitor screen is shown in fig. 14.

Even with the lower signal current, the smaller lag of the miniature "Plumbicon" tube gives better definition for moving subjects than is obtained with the studio camera.

When the miniature camera is used directly coupled to other optical instruments such as a microscope, a further improvement in definition can usually be obtained, since optical instruments generally have a circular image field. The deflection currents can then be increased so as to make use of the whole of the circular

[11] See reference [4], page 345 *et seq.*

photoconducting layer. Since less contour correction is necessary with this enlargement of the image size, this also gives a further improvement in signal-to-noise ration. (Of course, this enlargement of the image size also affects the sensitivity and the lag.)

with the camera tubes and with the camera itself do remain. For example, the advantages of the camera could be more fully utilized in outside broadcast work if the electronic circuits now housed in the intermediate unit could be mounted in the camera unit. This could

Fig. 14. The monitor screen in the experiment shown in fig. 13. The left-hand half shows the picture from the studio camera, which for this experiment was fitted with "Plumbicon" tubes of increased red sensitivity. The right-hand half of the screen shows the picture from the miniature camera.

The results obtained with the experimental camera already indicate that it would be very useful wherever it is difficult to install, mount, or connect up a large camera. However, one or two problems connected then produce a complete colour signal, and the camera could be connected by a thin cable to the mobile broadcast unit. An experimental camera of this type is at present under development.

Summary. Description of a small experimental colour television camera, fitted with specially developed miniature "Plumbicon" camera tubes with electrostatic focusing. The dimensions of the image on the photoconducting layer are only 8.4×6.3 mm. In spite of the small dimensions, the camera gives an image quality that compares very favourably with that given by a studio camera. Apart from applications in studios and for outside broadcasts, this camera also offers great advantages if images have to be viewed that have been produced by another optical instrument, for example a microscope. In such cases this small camera can be directly coupled to the relevant instruments. This reduces the loss of sensitivity and picture definition caused by the coupling elements needed when a large television camera is used for such purposes.

Magnet material with a $(BH)_{max}$ of 18.5 million gauss oersteds*

The usefulness of a material for permanent magnets is primarily determined by its demagnetization curve (which gives the magnetic flux density B as a function of the demagnetizing field strength H) and depends in particular on the maximum value which the product BH attains as this curve is traversed. Stages in the race to reach ever higher values of $(BH)_{max}$ were the development of tungsten steel in 1900 (with a $(BH)_{max}$ value of 0.34×10^6 gauss oersteds), cobalt steel in 1917 (0.90×10^6), "Ticonal" II [**] in 1936 (1.80×10^6), "Ticonal" G in 1937 (5.0×10^6), "Ticonal" GG in 1949 (8.3×10^6), platinum cobalt (gradually raised since 1936 to 10.5×10^6 in 1960) and "Ticonal" XX in 1956 (11.0×10^6) [1]. Later it was found possible to increase the $(BH)_{max}$ value of the last-mentioned material to 13.4×10^6 G Oe [2].

The search for still better magnetic materials is guided by the consideration that such materials should combine high values of saturation magnetization with high coercive force. This combination can be found in intermetallic compounds of cobalt and the rare earths. A particular material that has attracted attention in recent times is the samarium-cobalt compound $SmCo_5$ [3][4]. We shall mention here some results of recent experiments with this material in our laboratories.

The atoms of most of the rare-earth metals (general symbol R) have a magnetic moment of their own because of the unpaired electrons in the incompletely filled 4f shell, and this moment is generally a multiple of that of a cobalt atom. An intermetallic compound like RCo_5 may therefore be expected to have a high saturation magnetization if there is "ferromagnetic coupling" between the magnetic moments of the R and the Co atoms, i.e. if these align themselves parallel to one another. In the lighter rare-earth metals this does in fact happen.

Furthermore, the compounds RCo_5 have a hexagonal crystal structure with an exceptionally high crystalline anisotropy; the preferred direction of the magnetic moments of R and Co is generally the c-axis. The anisotropy field H_A, i.e. the field strength needed to align the magnetic moments perpendicular to their preferred direction, is as much as 290×10^3 oersteds in the case of $SmCo_5$. Provided the grains of the material are small enough, so that they contain no mobile domain walls whose displacement permits relatively easy demagnetization, the coercive force ($_IH_c$) could in theory reach the exceptionally high value mentioned above.

This combination of properties should theoretically enable permanent magnets with $SmCo_5$ as the base material to reach a $(BH)_{max}$ of more than 23×10^6 gauss oersteds, which is much higher than is obtainable with most other compounds from the RCo_5 series.

In our first experiments with this compound we were able to reach values of 8.1×10^6 G Oe [4]. The starting material for the magnets was obtained by melting together Sm and Co in a protective gas atmosphere, powdering the product of the melt and then subjecting the $SmCo_5$ powder to prolonged grinding (chiefly to reduce the size of the particles; see above). After the grinding, the powder was mixed with a binder, and then grain-oriented in a magnetic field and compacted. One purpose of the binder is to ensure that the orientation of the $SmCo_5$ particles — and hence their magnetic moment — remains fixed.

The fact that the $(BH)_{max}$ product obtained in this way was still far below the theoretically attainable value was partly attributable to the high porosity of the compacted magnet body: the relative bulk density (the ratio of the density to that of solid material) was only 70%. The familiar methods of increasing the bulk density of hard powders cannot be used with the RCo_5 compounds. In sintering, for example, the compacted powder has to be heated, which causes the permanent magnet properties to deteriorate.

By using a new pressing technique we have now succeeded in compacting the $SmCo_5$ particles into a body with a relative bulk density of 95%. It has also been found possible to improve the properties of the powder itself. The product $(BH)_{max}$ of the magnet obtained in this way has the unprecedented high value of 18.5×10^6 gauss oersteds.

The technique used may be described in general terms as follows. The starting material is again the ground $SmCo_5$ powder. After the grinding, however, the grains are now subjected to a treatment in a nickel bath, as a result of which better values for the coercive force and the remanence are obtained. The coercive

[*] This article was presented in a slightly different form as a paper at the 14th Annual Conference on Magnetism and Magnetic Materials, New York, November 18-21, 1968.

[**] "Ticonal", Registered Trade Mark.

[1] A. I. Luteijn and K. J. de Vos, Philips Res. Repts. **11**, 489, 1956. For a survey see H. J. Meerkamp van Embden, The evolution of the permanent magnet: a brief review, Philips tech. Rev. **18**, 358-360, 1956/57.

[2] P. A. Naastepad, Z. angew. Phys. **21**, 104, 1966.

[3] K. Strnat, G. Hoffer, J. Olson, W. Ostertag and J. J. Becker, J. appl. Phys. **38**, 1001-1002, 1967 (No. 3).

[4] W. A. J. J. Velge and K. H. J. Buschow, J. appl. Phys. **39**, 1717-1720, 1968 (No. 3).

force increases to 15.8×10^3 Oe. This powder is grain-oriented in a magnetic field of 30×10^3 Oe and then compacted to a small block which undergoes a further compaction under a hydrostatic pressure of 10 kilobars, giving a relative bulk density of 70%. The block is now packed in a lead container and subjected to a hydrostatic pressure of 20 kilobars, which raises the bulk density to 82%. Increasing the pressure has very little further effect, but when the block inside the pressure vessel is subjected to a slight uniaxial plastic deformation, the bulk density again rises steeply to 95% or more. An essential point here is that the parallel orientation of the powder particles remains virtually unaffected by these processes.

The demagnetization curve of a magnet produced in this way is shown in *fig. 1*. For comparison the curves for "Ticonal" XX, for the ceramic magnetic material ferroxdure 330k and for platinum cobalt are also shown. It can be seen that the new material is remarkable in that both the remanence B_r and the coercive force are very high. Previously ferroxdure had been noted for its high coercive force, which permits the design of unusually flat devices. The coercive force of the new magnetic material described here is even higher, so high in fact that it takes a strong laboratory magnet to demagnetize this material.

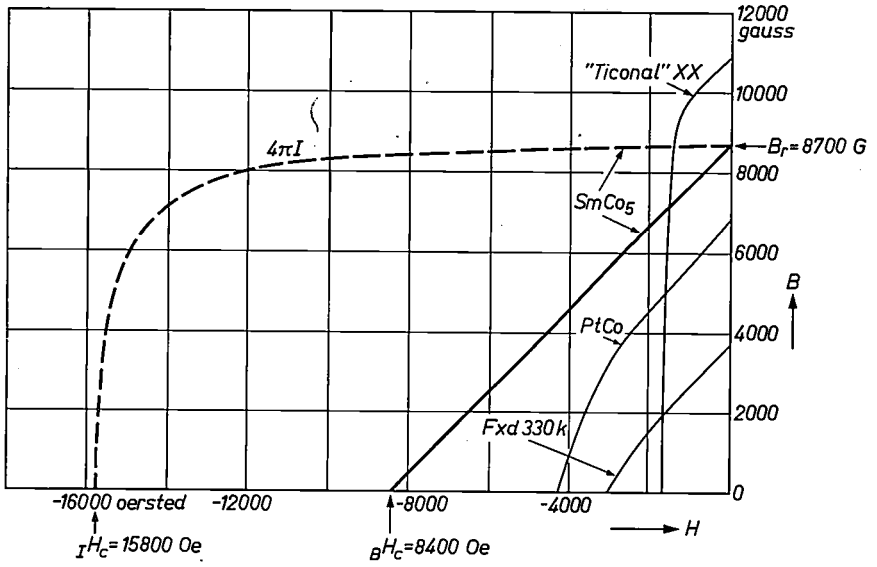


Fig. 1. Flux density B (solid line) and magnetization $4\pi I$ (dashed line) of the magnet of $SmCo_5$ as a function of the demagnetizing field strength H. For comparison the B-H curves for "Ticonal" XX, ferroxdure 330 k and platinum cobalt are also drawn (thinner curves).

Little can be said on the practical usefulness of the material so far. We have yet to see to what extent the other properties of the new material, in particular its stability, can be matched to the various requirements.

K. H. J. Buschow
W. Luiten
P. A. Naastepad
F. F. Westendorp

Dr. K. H. J. Buschow, Ir. W. Luiten and Ir. F. F. Westendorp are with Philips Research Laboratories; Ir. P. A. Naastepad is with the Metallurgical Laboratory of the Philips P.M.F. Division (Philite and Metalware Factories), Eindhoven.

Computer programming

L. Crousel

With the coming to maturity of computers, increasing attention is being paid to the ways in which they may best be used. In recent years the interest of most computer users has shifted from machine language to symbolic assembly languages and from these to statement-type languages. The advantages of the latter are that they are better adapted to the problems they are required to describe, they are not dependent on special machines, and they relieve the programmer of a great deal of programme organization. In addition, programme specialists find statement-type languages an excellent form in which to present procedures or "algorithms" to other users. In this article the author discusses the principal stages in computer programming for scientific problems. Simple examples illustrate some of the features of statement-type languages, symbolic assembly languages and machine languages, and some of the relationships between them.

Current developments in computer systems are making them much more accessible for the user. Remote computer terminals with typewriters and displays are being installed in user's offices; standard programmes for frequently recurring problems are being developed to assist designers; statement-type programming languages are being made available to express algorithms in a readily understandable form. To reap the benefit of these developments it is imperative that scientists and engineers keep up to date with the services offered by the computer centres to which they have access. This implies, among other things, that they should become familiar with direct access to the computer via the remote terminals and learn to use the general-purpose programmes. Moreover, computer centres are concentrating increasingly on the development of basic software, so that the burden of developing special-purpose programmes is being shifted more and more on to the users. Our aim here is to give some general guidance on how to meet this new situation. The processes involved in setting up a programme are often sufficiently complicated to justify a systematic approach.

Separation into stages

It is often not at all easy to see whether the results obtained from a programme-run are meaningful. All sorts of things can go wrong. It may be that the mathematical model is inadequate; the numerical procedure may turn out to be unsuited for a particular case because of loss of accuracy; some part of the programme may have undesirable side-effects; the computer itself may fail. It is a daunting task to trace the source of error once the programme has been run. Such difficulties must be foreseen and avoided by breaking down the development of the programme into separate stages. The advantage of such a separation is that the work can be distributed among several people each of whom looks for the best solution to his particular piece of the work. This cooperation stimulates creative thinking and each worker has to devise his own special checks for his part of the work.

The breakdown into stages does not necessarily imply that different people are involved with each. Neither does it imply that the boundaries between the stages are rigid. Indeed communication across these boundaries is essential if the desired end is to be attained. But this communication can only be efficient if its object is clearly defined. The stages involved in developing a programme are as follows:

Stage 1. Preparation of a programme specification in the form of an analytical definition of output as a function of input.

Stage 2. Design of a procedure which sets out the sequence of operations for implementing the function defined in the programme specification.

Stage 3. Coding of the procedure in a language acceptable to the computer system.

L. Crousel is with MBLE Laboratoire de Recherches, Brussels.

Stage 1. Programme specification

The first step in the development of a programme for solving a given problem is to collect relevant information, find an analytical formulation, define the input parameters and the output results and outline the method of solution. This stage is called systems analysis [1] by analogy with system design [2] in the development of hardware. Its object is to draft a comprehensive specification of what the system (computer and programme) should do.

Systems analysis is an intrinsic part of problem solving, whether a computer is used or not. The computer is of little help at this stage. Indeed it compels the systems analyst to adopt an unusually rigorous approach because of the almost complete lack of surveillance of intermediate results in automatic computation. An oversight in the analysis carried through to a programme may give rise to erroneous results when the programme is run on a computer, whereas conventional means of solution would have revealed the error. A typical example is that of a quotient of two functions which takes the indeterminate form 0/0 for some values of the arguments. If this has been overlooked, it is quite possible that a loss in accuracy may cause the denominator and numerator to assume values differing slightly from zero. This can lead to a completely wrong value of the quotient, and the error might not be noticed.

More often than not, systems analysis is the most important stage in programme development. Yet, because of the variety of problems amenable to computer solution, it is not feasible to list all the processes involved. The comments which follow are therefore neither exhaustive nor universally applicable.

An underlying similarity in the analysis of problems which originate in various scientific fields (physics, chemistry, engineering, economics, etc.) is the characterization of concepts by variables and the formulation of mathematical relations between them. Such a relationship is known as a mathematical *model* and is a prerequisite to further analysis and numerical computation. As an illustration, consider the computation of electron trajectories in a cathode-ray tube for given parameters [3]. Definition of the type of tube defines the problem without ambiguity. The design of a model consists in the definition of the relevant variables (potential, electric field, forces, etc.) and the expression of the appropriate laws of physics and mechanics in the form of differential equations.

Systems analysis does not end with a mathematical description of the problem. It must also set out the main structure of a solution. Every possible method from algebra, calculus, numerical analysis etc., may be brought to bear on the problem but the characteristics of machine computation must be kept in mind to guide the final choice of method. One guiding principle is the trade-off between computation time and programming time. Thus to find the minimum of a function within an interval it may be better to use tabulation with an adjustable step rather than a method which determines where the derivative is zero. Programming may be much faster in the first case while the difference in computation time may not even be noticeable. Another feature of machine computation is the benefit that may be derived from previous developments. A computer centre generally offers a collection of standard programmes for solving problems of common interest. Typical examples are the computation of standard mathematical functions, the solution of a system of linear equations, matrix inversion, various numerical integration procedures, etc. Programmes like these provide the building blocks from which larger programmes can be assembled.

The foregoing shows the need for familiarity with the peculiarities of computation in the formulation of solutions to be calculated on a computer. This does not mean that a computer analyst requires no help from the scientist who has set up the mathematical model. Knowledge of the theory behind the model often provides insight into methods for solution which are not apparent from abstract mathematical analysis.

While the use of a computer for problem solution may provide considerable saving in labour, it does not reduce the need for clear thinking. A programme specification of extreme precision must be drafted and this requires a thorough systems analysis. It is not really possible to think the problem through and take action as parts of the solution are unfolded, as one might do when using a slide rule or a desk calculator. True, it is possible and sometimes desirable to break up the problem into simpler subsections. But each "segment" of the problem must itself be accurately specified.

The programme specification is the point of departure for the development of a procedure layout. It is generally the key element at the interface between the author of the problem and the person who solves it, as when the programming is taken over by a computer centre. This does not imply one-way transmission since the method of solution may suggest that the specifications should be altered. Consequently the programme specifications may undergo successive drafts until both parties agree on its adequacy and comprehensiveness.

[1] C. Strachey, System analysis and programming, Sci. Amer. **215**, No. 3, 112-124, 1966.

[2] G. M. Amdahl, G. A. Blaauw and F. P. Brooks Jr., Architecture of the IBM System/360, IBM J. Res. Devel. **8**, 87-101, 1964.

[3] C. Weber, Calculation of potential fields and electron trajectories using an electronic computer, Philips tech. Rev. **24**, 130-143, 1962/63.

Stage 2. Procedure layout

Once an adequate programme specification has been obtained the next stage is to define a procedure for solving it on a computer. The analyst who performs this task must be able to select an approach inspired by the appropriate disciplines (including numerical analysis) and tailored to computer processing.

As an illustration of the constraints imposed by computer structure, consider the "travelling salesman problem". He starts from base, visits m places and returns to base. We want to find the shortest route. For lack of a better strategy, we propose to list all possible routes, compute their length and retain the shortest. This method requires a procedure for generating the permutations of m elements identifying the m places. There are P_m permutations, where P_m is given by the recurrence formula:

$$P_m = m\,P_{m-1}.$$

All the permutations may be produced by means of the iterative procedure commonly used for deriving this formula. It requires the memorization of $(m-1)!$ sequences of $m-1$ elements and this may be prohibitive when m is large. What is called for is a better strategy which produces the permutations in sequence without having to retain a large number of permutations previously generated. There are in fact procedures which, apart from some simple "book-keeping" like making a number of counts, use only the last permutation produced. For example, a method is given by Langdon [4], based on circular permutations of subsets of the sequence. Though this procedure is straightforward, its formulation was stimulated by computer programming without which it probably would not have been conceived.

As another illustration, consider the tabulation of the function:

$$A(\gamma) = \ln \left| \prod_{i=1}^{n} \frac{e^{\gamma}/e^{\gamma_i} + e^{\gamma_i}/e^{\gamma}}{e^{\gamma}/e^{\gamma_i} - e^{\gamma_i}/e^{\gamma}} \right|. \quad \ldots \quad (1)$$

This is used in filter design as an approximate expression for attenuation as a function of frequency. γ is a suitable transform of frequency and the n values of γ_i are fixed frequencies at which the attenuation is infinite. Using auxiliary variables

$$w_i = e^{\gamma_i}$$

and

$$v = e^{\gamma},$$

we have:

$$A(\gamma) = \ln \left| \prod_{i=1}^{n} \frac{v/w_i + w_i/v}{v/w_i - w_i/v} \right|. \quad \ldots \quad (2)$$

The values of w_i are constants which may be determined by a preliminary computation. Consequently, a programme based on form (2) requires only one exponential (e^{γ}) and one logarithm for each value of γ (apart from elementary operations which are relatively fast). The function (1) is usually given in the form:

$$A(\gamma) = \sum_{i=1}^{n} \ln \coth \left|(\gamma - \gamma_i)\right|. \quad \ldots \quad (3)$$

This form is convenient for a computation using graphical methods or tabulated values. With the aid of a table of $\ln \coth |x|$, all that is required is n searches in the table and n additions. In automatic computation however, standard functions such as logarithm and exponential are computed from series expansions rather than stored in table form which is a waste of memory space. The evaluation of such functions requires a fair number of arithmetic operations. Contrasting computation using (2) and (3) we see that in (3) the number of standard functions is n times greater, for there are n logarithms and n hyperbolic functions (which can be reduced to exponentials). Thus, the time required for running the programme based on (3) will be larger by a factor of approximately n. Such is the penalty for overlooking the characteristics of automatic computation.

These examples illustrate how a basic knowledge of automatic computation guides an analyst in deciding on the layout of the procedure. The mental processes which lead to the solution of the problem vary with the temperament and background of the analyst. But an outline of the method of attack is not enough to provide a secure basis for coding. The steps must be assembled in a systematic procedure layout which is a complete description of the sequence of operations which have to be performed to obtain a specified result. A formalized presentation known as the "flow chart" has been developed for this purpose.

The distinctive feature of flow charts is the unambiguous definition of the order in which operations have to be performed. In a typical representation, each operation is framed in a block and the sequence is defined by arrows between blocks. A simple example is given in *fig. 1* which exhibits the flow chart of a

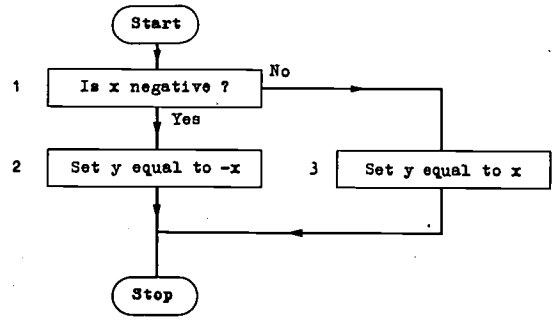


Fig. 1. Layout of procedure that evaluates y, the absolute value of a given variable x.

procedure to compute y, the absolute value of x. It consists of three operations. The first is a branch, which directs the operation sequence along the "Yes" or "No" path depending on the answer to the question.

The operation in a block may be quite a general one, requiring a more detailed flow chart to describe it. Outlining the procedure in stages of increasing complexity allows the problem to be divided into subproblems and limits the size and complexity of the elements which must be considered at one time. It corresponds to the design of a system with block diagrams.

The merit of flow charts is that attention is focused on the topology of the blocks. Different ways of organizing the steps may be compared and related to the overall pattern. As an illustration, consider the computation of a matrix B equal to the n^{th} power of a square matrix A (n is an integer). A procedure is given in the flow chart *fig. 2*. Operation 6 is the computation of a matrix product, a problem whose solution is given in a separate flow chart, the details of which are of no concern here. The flow chart includes 0 as a legitimate power with the unit matrix as a result. This is taken care of by the branch instruction at the beginning. The result could have been achieved without a branch simply by changing operations 1, 3, 4 etc. into:

3 Set k equal to 0
4 Set B equal to the unit matrix,

and keeping the core, which is the matrix multiplication, the same. The drawback, however, is clear — for all $n \neq 0$ the process takes one matrix multiplication more.

The examples given above point out the desirability of developing a concise notation to describe detailed operations in procedure descriptions. Operation 3 in fig. 1 is an assignment statement which we shall represent by:

$$y := x,$$

using a notation borrowed from ALGOL 60. (This is read as "y becomes equal to x".) The operator := is designed to draw attention to the asymmetry of the operation, a concept which is lacking in the common notation

$$y = x,$$

which merely states a relation between two variables without specifying that one is determined from the other. The operation defined by the assignment statement $y := x$ implies that the variable y is assigned the value x, and such assignments can be repeated with different values in the course of the procedure. We stress this property by stating that procedure variables are *dynamic*. By contrast, the variables in a mathematical description may be termed static because their relationship with other variables as expressed by equations is assumed to hold throughout the body of a mathematical description. To illustrate this feature, consider the Newton-Raphson iteration formula for improving the approximation x_n of a zero of the function $f(x)$:

$$x_{n+1} = x_n - f(x_n)/f'(x_n). \quad . \quad . \quad . \quad . \quad (4)$$

The assignment statement which implements this rule is:

$$x := x - f(x)/f'(x). \quad . \quad . \quad . \quad . \quad . \quad (5)$$

In equation (4) successive approximations are associated with a sequence of static variables whereas in statement (5) they are assigned successively to the unique dynamic variable x.

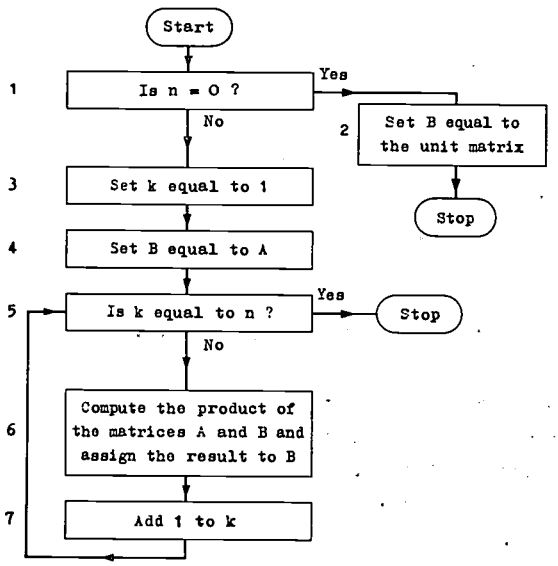


Fig. 2. Procedure that computes B, the nth power of a square matrix A.

Similarly the assignment statement

$$i := i + 1$$

increases variable i by unity, each time it is carried out. A statement like this is usually used for counting, in which case it has to be preceded by a statement assigning an initial value to i, e.g. $i := 0$.

The dynamic nature of flow-chart variables reduces the number of symbols required for expressing a procedure and hence the number of memory cells. It also makes it difficult to describe or annotate a procedure layout, especially if the same variable is used for a variety of purposes in a procedure body.

Other notations may be introduced as the need arises, although a number of them are more or less standard.

[4] G. G. Langdon, Jr., An algorithm for generating permutations, Comm. ACM **10**, 298-299, 1967 (No. 5).

Thus branch operation 1 in fig. 1 is represented as shown in *fig. 3*, the distinctive block form (a diamond in this case) being used to distinguish it from the operations which do not control the sequence.

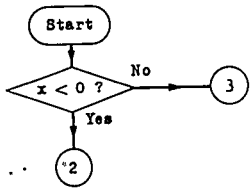


Fig. 3. Flow diagram symbolism representing a branch operation.

We shall now look at the designing of a procedure in some detail. As an illustration, let us consider the computation of the product C of two matrices A and B represented by:

$$C_m^p = A_m^n B_n^p,$$

where subscripts and superscripts specify the number of rows and columns respectively as shown schematically in *fig. 4*. The problem can be defined by the equations:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \quad \ldots\ldots \quad (6)$$

for $i = 1, 2, \ldots m,$
$j = 1, 2, \ldots p,$

where a_{ik} designates the element of row i and column k of matrix A. Similar definitions hold for b_{kj} and c_{ij}.

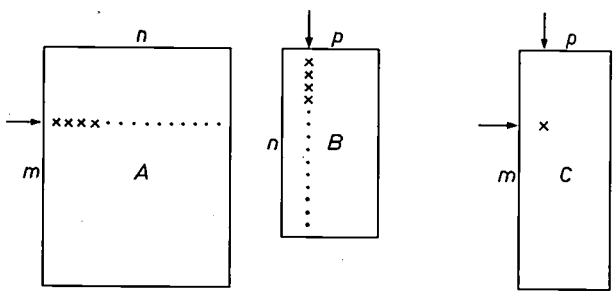


Fig. 4. Illustration of the relation between the dimensions of the matrices in the product $C_m^p = A_m^n B_n^p$.

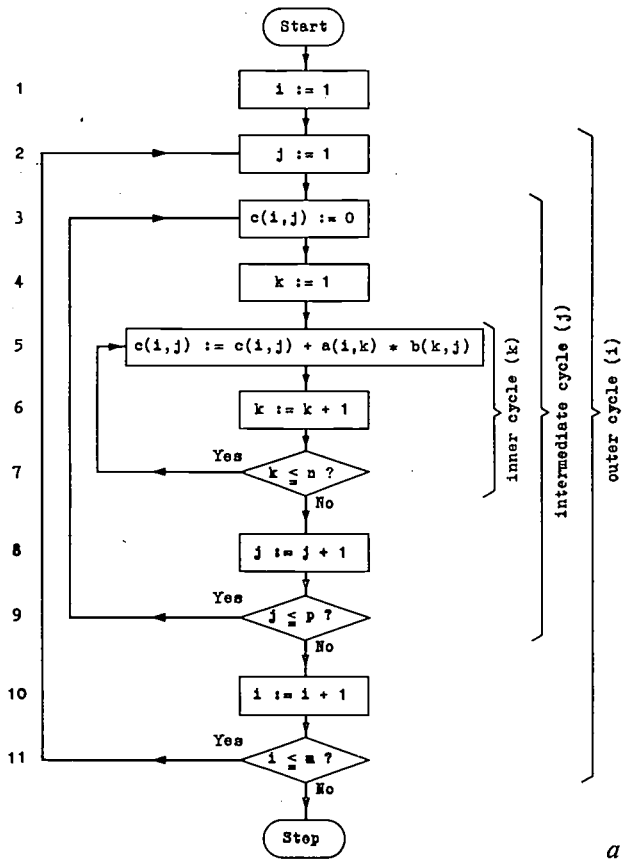
A procedure that solves this problem is given in *fig. 5a*. For ease of reference each operation is labelled with a number on the left of the block. The significance of the variables is obvious in this case. Even so, it is good practice to complement flow charts with a table of symbols such as given in fig. 5*b*.

The core of the procedure is operation 5 which accumulates the products defined by equations (6). While the latter merely state what results are desired, the procedure embodies a unique choice of the order in which the elements of C are computed (row by row) and the order in which the summation of terms is performed (scanning a row of A from left to right and a column of B from top to bottom).

To visualize the mechanics of the procedure and at the same time check its validity, let us apply it to a simple example. Let

$$A = \begin{vmatrix} 1 & 2 \end{vmatrix} \quad \text{and} \quad B = \begin{vmatrix} 1 & 3 \\ 2 & 4 \end{vmatrix},$$

so that $m = 1, n = 2, p = 2$. The results are presented



a

Symbol	Designation
i	row index of matrices A and C
j	column index of matrices B and C
k	row index of matrix B column index of matrix A
m	number of rows in matrices A and C
n	number of columns in matrix A number of rows in matrix B
p	number of columns in matrices B and C
$a(i,k)$	element of matrix A
$b(k,j)$	element of matrix B
$c(i,j)$	element of matrix C

b

Fig. 5. *a*) Basic flow chart of procedure that computes the matrix product $C = A \times B$. In FORTRAN multiplication is denoted by * to avoid confusion with the letter x, and we have adopted this convention in all our flow charts.
b) Table of variables in flow chart of (*a*). A table like this improves the readability of the flow chart.

in tabular form (*fig. 6*). The first column contains the labels of the operations in the order of their execution. The following columns give the values of the flow-diagram variables. The last column lists the answer ("Yes" or "No") to the question raised in a branch instruction. The first line in the table represents the conditions at the start which are unknown and irrelevant. Successive line entries are obtained by determining from the flow chart which operation is to be carried out, performing the operation found there and entering the result in the appropriate variable or branch condition column. The value of a variable at any one time (corresponding to a particular line entry) is the last one listed in the corresponding column. Final values are repeated in the last (Stop) line and give the result

$$C = |5 \quad 11|.$$

Operation label	Variables					Branch condition
	i	*j*	*k*	*c*(1,1)	*c*(1,2)	
start	—	—	—	—	—	—
1	1					
2		1				
3				0		
4			1			
5				1		
6			2			
7						Y
5				5		
6			3			
7						N
8		2				
9						Y
3					0	
4			1			
5					3	
6			2			
7						Y
5					11	
6			3			
7						N
8		3				
9						N
10	2					
11						N
stop	2	3	3	5	11	N

Fig. 6. Successive steps in the matrix multiplication procedure given in fig. 5.

The check thus obtained is practically complete. It is too cumbersome to be applied in general but it is often valuable for testing delicate points in the logical organization such as whether a cycle has been carried out the correct number of times or whether the procedure is valid for critical values of the parameters.

The notation developed for flow charts obviously owes much to computer structure and languages. The dynamic variables which occur in consecutive assignment statements are associated with computer memory cells in which successive values are stored during the running of the programme. Branching operations are the equivalent of computer jump instructions, operation labels are associated with instruction addresses. The table in fig. 6 can be thought of as a "film" of the contents of the memory cells during the running of the programme. A flow chart may be considered as a shorthand notation for a computer programme. It constitutes the detailed specification from which a programme is written, and yet, in its form considered so far, it is not tailored to a specific computer. Indeed the "computer" can be a person as the check of the matrix multiplication procedure has shown. The general nature of the flow chart is most valuable. The procedure layout can be read without reference to any particular programming language. It is a clear specification by which the performance of a programme may be assessed and coding errors distinguished from logical errors.

The design of a procedure layout is a process which evolves in a number of steps. The function defined by the programme specification is broken down into a sequence of operations whose interrelationship is mapped in a first-level flow chart. Any operation which is not elementary constitutes a subproblem which is handled in the same way. This gives rise to second-level flow charts and the process continues until the problem is finally reduced to elementary operations. The definition of what is elementary depends on the computer language and the available software, as will be illustrated below.

A good procedure layout can seldom be obtained at a first attempt. The subdivision of the problem chosen at one level may prove to be awkward when the subproblems are worked out. Similarities between subproblems may even suggest a complete reappraisal. Not the least merit of the successive levels of description allowed for by the flow-chart formalism is that this reappraisal can be carried out without the programmer being burdened by the details of the syntax of a computer language.

Stage 3. Coding

The next stage in programme development consists in coding the procedure layout in a language acceptable to the computer system. Perhaps the easiest way to introduce the subject of computer languages is to start with statement-type languages although historically they represent the latest development. The justification for this reversed order of presentation is that computer system development has quite naturally been geared to simplifying the task of programming at the expense of an increase in the complexity of the basic software for translating and running the programme on the machine.

Statement-type language coding

Statement-type languages are designed to express operations by means of formalized statements which look very much like standard mathematical expressions. They are, to a large extent, computer-independent and their use does not require the knowledge of a specific computer code, although this may contribute to programme efficiency as will be illustrated below.

In order that a programme written in a statement-type language may be run on a computer it is necessary to translate it into computer instructions. This translation is performed by special programmes called compilers, and the development of suitable compilers has markedly influenced the evolution of the design of statement-type languages. Thus FORTRAN [5] (formula translation) which came first was designed with obvious attention to the difficulty of translation, while more recent languages such as ALGOL 60 [6] (algorithmic language) have been given a more systematic structure.

As an illustration, *fig. 7* and *fig. 8* show the main body of a FORTRAN and an ALGOL 60 programme respectively for computing the matrix product. Both are based on the procedure laid out in the flow chart fig. 5*a*. Operations in the flow chart are translated into statements which obey rigid syntactical rules. Thus operation 5 in the flow chart is expressed by statement 6 in the ALGOL 60 programme. A semicolon separates the statements and can be thought of as equivalent to the block which frames the operation in the flow chart. The corresponding FORTRAN statement 5 differs in the notation for operators and the compulsory use of capitals in the names of the variables. The statement termination is not explicit but implied by the transition to the next line.

```
1        DØ 10, I = 1(1)N
2        DØ 10, J = 1(1)P
3        C(I,J) = 0
4        DØ 10, K = 1(1)N
5   10   C(I,J) = C(I,J) + A(I,K) * B(K,J)
```

Fig. 7. Main body of a FORTRAN programme that computes the matrix product $C = A \times B$. In FORTRAN the character Ø is used to represent the letter O and thus avoid confusion with the digit 0.

```
1  for i := 1 step 1 until m do
2  for j := 1 step 1 until p do
3      begin
4          c[i,j] := 0;
5          for k := 1 step 1 until n do
6          c[i,j] := c[i,j] + a[i,k]×b[k,j]
7      end
```

Fig. 8. Main body of ALGOL 60 programme for computing the matrix product $C = A \times B$.

Some sets of flow-chart operations are reorganized in the statement languages. Thus the intermediate cycle (j) laid out in operations 2, 8, 9 of the flow chart is represented by the FORTRAN statement 2 associated with the label 10 of statement 5. Its meaning is, "repeat the sequence of statements starting with the next (3) up to and including 5 for integer values of index J ranging from 1 to P". In ALGOL 60 the intermediate cycle is represented by statement 2 and its extent is specified by the pair **begin** (line 3) and **end** (line 7). Such pairs are used in a similar way to brackets in an arithmetic expression.

We should note that the matrix-multiplication procedures given above can be coded without the benefit of the detailed procedure layout shown in fig. 5. This does not mean that flow charts are superfluous when statement-type languages are used for coding. They remain a powerful means of expressing and analysing the structure of complex procedures.

Assembly-language coding

It is not always desirable or even feasible to code a procedure in a statement-type language. In particular, a compiler must be coded in a language for which a translator already exists. More generally, a statement-type language may not provide appropriate statements for efficient organization and processing of some types of data. As a result of their independence of type of computer statement-type languages do not give access to the complete range of instructions of a particular computer. This access is only given by languages which provide a symbolic formalism in order to express individual computer instructions. Such a language is called an assembly language [7] and is associated with a specific computer.

Assembly-language coding requires detailed knowledge of computer instructions, internal data representation, index registers etc. It involves the design of flow charts in greater detail. In practice, assembly-language coding is used only by programmers at a computer centre, e.g. for programmes where exceptionally high efficiency is required. Nevertheless, some insight into assembly-language coding may be useful to indicate the differences likely to be encountered between the result of a compilation from a statement-type language and a programme written directly in assembly language.

As an illustration consider again the matrix multiplication outlined in fig. 5. With few exceptions computers are equipped with instructions which provide direct access to variables with one index (vectors) but not with two (matrices). To overcome this restriction matrices are represented by vectors (for instance by assembling the rows of a matrix in sequence). Thus the matrix C defined above, which has m rows and p col-

umns, is represented by a vector C_1 with $m \times p$ components. The vector and matrix elements are related by:

$$c1(r) = c(i,j),$$

where the relation between the indices is:

$$\begin{aligned} & r = (i-1)\times p + j \\ \text{for} \quad & i = 1, 2, \ldots m \\ & j = 1, 2, \ldots p \\ \text{so that} \quad & r = 1, 2, \ldots m \times p. \end{aligned}$$

Similarly matrices A and B are represented by vectors $\mathbf{A}_1$ and $\mathbf{B}_1$, where:

$$\begin{aligned} & a1(s) = a(i,k) \\ \text{with} \quad & s = (i-1)\times n + k \\ \text{for} \quad & i = 1, 2, \ldots m \\ & k = 1, 2, \ldots n \\ \text{so that} \quad & s = 1, 2, \ldots m \times n, \end{aligned}$$

and similarly:

$$\begin{aligned} & b1(t) = b(k,j) \\ \text{with} \quad & t = (k-1)\times p + j \\ \text{for} \quad & k = 1, 2, \ldots n \\ & j = 1, 2, \ldots p \\ \text{so that} \quad & t = 1, 2, \ldots n \times p. \end{aligned}$$

Substituting the vector notation and the index relations which it requires into the procedure shown in fig. 5 yields the flow chart *fig. 9*. The position of the index conversion operations in the flow chart requires careful consideration. Thus operation 3 which computes index r:

$$r := (i-1)\times p + j$$

must be performed after any change of i or j and before any operation which involves $c1(r)$.

The flow chart in fig. 9 exhibits all the variables and operations which are required for coding in assembly language. The index conversion operations 3, 6 and 7 are similar to those which can be expected in the object programmes obtained by translation (compilation) of the FORTRAN and ALGOL 60 programmes given in figs. 7 and 8. These index-conversion operations will take up a significant portion of the running time. This is apparent from fig. 9 if we observe that operations 6 to 10 which constitute the inner cycle (k) are repeated $m \times n \times p$ times, i.e. a multiple of the number of times the other operations are performed (except in the trivial case where $n = 1$). To get an approximation we count the elementary operations in the inner cycle and find that while the operation on the matrix variables (8) involves only one addition and one multiplication, the operations on the indices call for five additions and two multiplications. Since in most computers multiplication time is far longer than addition time, it is worth while to devise a method of index computation which does not require multiplication. The objective is to find the laws which govern the progression of the vector indices r, s and t and to implement these laws explicitly rather than their dependence on the matrix indices i, j and k. The result of this change of index computation is shown in *fig. 10*. Inspection of the inner cycle (k) shows that operations on the indices have been reduced to only three additions. The important advantage of the flow chart of fig. 10 compared to that of fig. 9 is that two multiplications have been eliminated. For example, on the Philips P3 computer (a transistorized version of

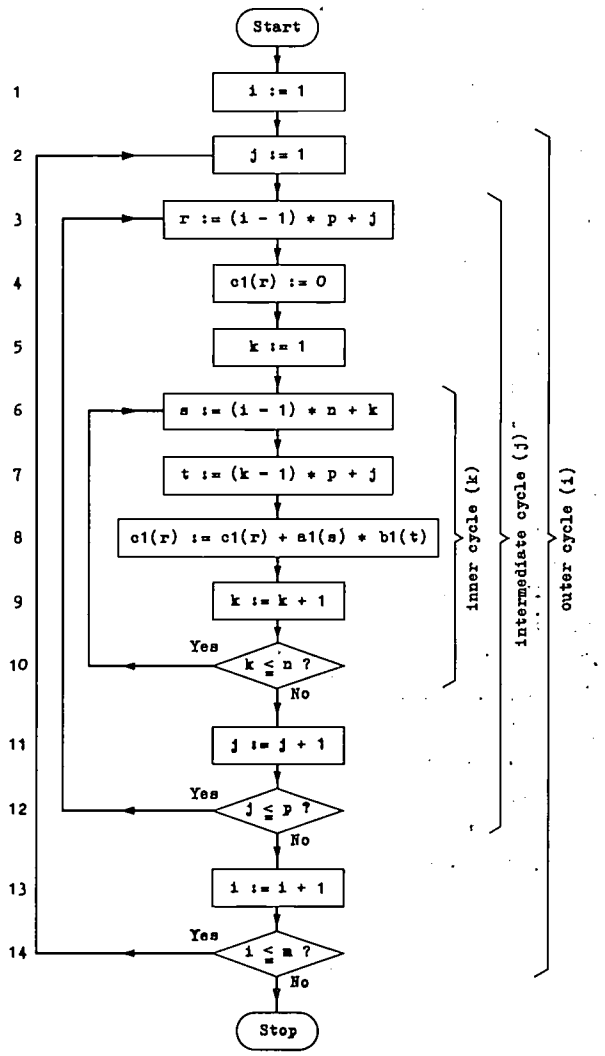


Fig. 9. Flow chart for computing the matrix product in which the variables have at most one index. This is suitable for assembly language coding.

[5] D. McCracken and W. Dorn, Numerical analysis and FORTRAN programming, Wiley, New York 1964.

[6] Revised report on the algorithmic language ALGOL 60, editor Peter Naur, A/S Regnecentralen, Copenhagen 1964. See also E. W. Dijkstra, A primer of ALGOL 60 programming, Academic Press, London 1962.

[7] P. Wegner, Introduction to symbolic programming, Griffin, London 1963.

PASCAL described in [8]) it accounts for a 40 % reduction in running time. This may not be significant for programmes requiring only a few runs but a programme such as this is obviously designed to be a standard, that is, a procedure that will be entered as a block in diverse programming applications. The effort required to transform flow chart fig. 9 to fig. 10 is therefore worth while.

It must be emphasized that the flow charts are the appropriate stage of programming at which to analyse and implement such procedure modifications. It must further be observed that the coding of the flow chart in fig. 10 requires no more effort than the coding of the flow chart in fig. 9.

So far we have been concerned with the transformation and refinement of flow charts to provide the more detailed procedure layout required by assembly-language coding. We now proceed with the work of coding proper, using a different example for the sake of brevity. Consider the evaluation of a polynomial:

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n.$$

Written in the form:

$$P(x) = \{ \ldots [(a_0x + a_1)x + a_2]x + \ldots a_{n-1}\}x + a_n,$$

the evaluation can be implemented by the procedure laid out in flow chart *fig. 11*. Coding of this procedure will now be discussed.

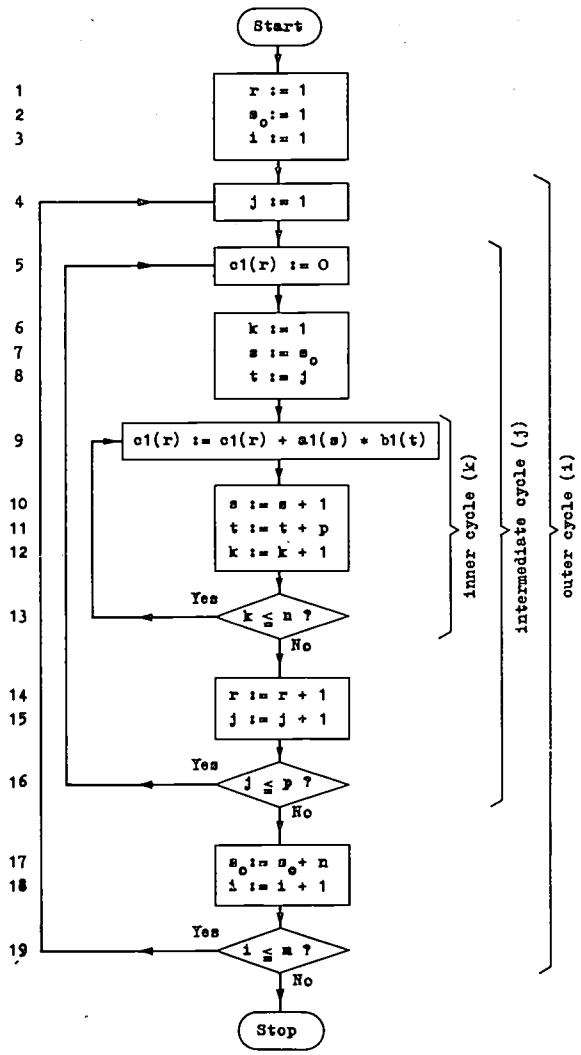
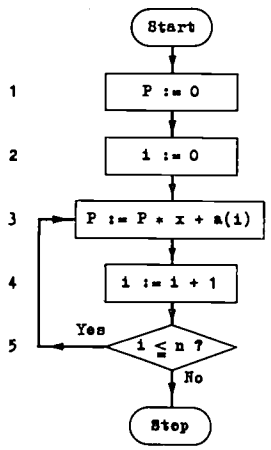


Fig. 10. Flow chart of matrix multiplication with explicit sequencing of indices r, s, t; this is more efficient than the flow chart in fig. 9. The index r of the elements of C_1 increases in the natural order. Operation 1 gives r its initial value, and 14 increases it by one with each termination of the inner cycle k. To scan column j of matrix B, index t of vector B_1 must start from j and step by p with each step in k. This process is independent of row index i and is performed by operations 8 and 11. The progression of index s of vector A_1 is more involved. Matrix A is scanned row by row and hence s increases by one with k (operation 10). Each j cycle maintains the same row in A and therefore resets s to its initial value s_0 (operation 7). This is the index in A_1 of the first element in row i of A. Thus it starts at 1 (operation 2) and steps by n with each step in i (operation 17).



Flow-chart symbol	Programme symbol	Designation
P	PØL	polynomial value
x	X	argument
n	N	polynomial degree
a	AV	coefficient vector
i	IRI	component index

Fig. 11. Flow chart of polynomial evaluation and table of symbols for assembly-language programming.

A preliminary task in coding consists in the selection of names for the variables which satisfy the syntax and associating them with flow-chart variables. Similarly, computer index registers are selected to represent flow-chart indices as far as these are available (otherwise instructions for *alternate* index saving must be added). The associated symbols are shown in the table; this is necessary to make the programme understandable when the syntax does not permit an obvious correspondence.

Assembly-language coding proceeds by breaking down each flow chart operation into a sequence of

instructions as shown in *fig. 12a*. Column 1 is an order number used only for convenience of description. Column 2 gives the sequence of operations chosen from those available to the computer, but represented in flow-diagram notation. The names A and S designate computer registers which participate in the built-in (micro-programmed) arithmetic. Column 3 shows the actual assembly code. It consists of the sequence of instructions 1 to 13 followed by specifications of variables 14 to 19. The instructions have three components: a label (optional) which provides a symbolic reference

programme (fig. 12) will now be considered for one or two of the operations.

Operation 3 of the flow chart:

$$P := P \times x + a(i),$$

is expressed by the four instructions 4 to 7. Instruction 4 transfers the value of x to the S register. Instruction 5 takes P, multiplies it by the contents of the S register and the product is put into the A register. Operand AV,1 of instruction 6 (the 1 after the comma refers to an index register) designates $a(i)$, the component of vector

a

1 Order number	2 Operation	3 Instruction			4 Memory
		label	operator	operand	
1	$A := 0$		CAA	ZERØ	512
2	$P := A$		STØ	PØL	513
3	$i := A$		STØ	IR1	514
4	$S := x$	CYCLE	CAS	X	515
5	$A := x \times P$		MPY	PØL	516
6	$A := A + a(i)$		ADA	AV,1	517
7	$P := A$		STØ	PØL	518
8	$A := i$		CAA	IR1	519
9	$A := A + 1$		ADA	ØNE	520
10	$i := A$		STØ	IR1	521
11	$A := A - n$		SBA	N	522
12	if $A \leqq 0$ go to 4		JAN	CYCLE	523
13	stop		STP		524
14		ZERØ	FLN	0	525
15		ØNE	FLN	1	526
16		PØL	ØPN		527
17		X	ØPN		528
18		N	ØPN		529
19		AV	ØPN	50	530

b

Operation code		Description of operation
numeric	symbolic	
00	CAA	Clear A and add operand to A register
01	CAS	Clear S and add operand to S register
02	STØ	Store A register content in operand address
03	ADA	Add operand to A register
04	SBA	Subtract operand from A register
05	MPY	Multiply operand with S register
06	JAN	Jump to instruction: label given in the operand if A register is zero or negative
07	STP	Stop

Fig. 12. *a*) Assembly-language programme for polynomial evaluation. In column 2 operations are designated by flow-diagram notation. Column 3 gives the actual assembly-language code. *b*) List of operation codes used in the assembly-language programme (*a*).

for the instruction, a three-letter code (see fig. 12*b*) for designating the operation and an operand which is the label of a variable or an instruction. The specifications of the variables have a similar structure and will be explained in connection with the translation of a programme written in assembly language.

The correspondence between flow chart (fig. 11) and

AV whose rank is defined by the value i contained in index register $IR1$. Thus instruction 6 adds this component to the product $P \times x$ present in A, while instruction 7 stores the result in P.

Operation 5 of the flow chart is carried out by the sequence of instructions 10 to 12, where 12 is a branch instruction dependent on the sign of A. The sequence is adequate if the zero result of instruction 11 for $i = n$ has a computer representation whose sign is negative, as is the case with many computers.

Coding in assembly language thus requires consider-

[8] W. Nijenhuis, The PASCAL, a fast digital electronic computer for the Philips Computing Centre, Philips tech. Rev. **23**, 1-18, 1961/62. See also H. J. Heijn and J. C. Selman, The Philips computer PASCAL, IRE Trans. on Electronic Computers **EC-10**, 175-183, 1961.

able attention to detail. It is therefore essential for the programmer to be able to rely on an adequate procedure layout which frees him from all questions of logical organization. Even so, it is unrealistic to expect that a large programme will be correct in its first draft. A reasonable approach is to anticipate errors and to take steps to discover and correct them. This is known as "debugging". The techniques include programme subdivision (segmentation), the development of special-purpose test programmes, output of intermediate results, etc.

It should be stressed again that there is no rigid boundary between the three stages. Efficient coding may require modifications of the layout in the detailed flow charts. The efficiency considered here relates to the complete process of coding and debugging and should not be confused with the reduction of programme length by the use of tricks. A common trick consists in altering an instruction sequence during the running of the programme by using these instructions as operands. Such games are left over from the pioneering era when computer index registers were not available. They still amuse some programmers who thereby succeed in making their programmes practically unreadable.

Machine-language coding

What has been said in the foregoing with regard to assembly languages applies directly to machine-language coding. Although scarcely anyone will use this nowadays, it seems appropriate to consider it here to give some insight into the problem of translation. The programme in machine language is of course the ultimate result of every translation, since it is the only language which the computer itself understands.

A programme in assembly language is a faithful image of the code which is in the computer memory during the actual run but its introduction into the computer still requires a translation into machine code. The programmes which perform this translation are called assemblers. Their main purpose is to map the programme into the computer memory. This mapping process produces a list of computer memory addresses associated with labels of instructions and variables. The list is used to substitute numerical for symbolic addresses in the operands of the instructions. The specification of the variables used as operands is performed by means of special commands in the programme text which direct the assembler but do not generate instructions in final machine code. Thus lines 14 to 19 in fig. 12 are not ordinary instructions but commands which assign memory cells to the programme variables, and either fill these cells with initial values, such as the constants 0 and 1 (in floating-point notation), or leave these values undefined. Thus, line 19 assigns the 50 memory locations 530 through 579 to the coefficients $a(i)$.

The computer code is illustrated by *fig. 13* which lists the computer instructions for the polynomial evaluation in a hypothetical computer code. The programme can be obtained from the assembly programme fig. 12 by simulating the task of the assembler, as follows. First consecutive addresses are assigned to the instructions as shown in the right-hand column of fig. 12*a*. The address of the first instruction (512) is arbitrary, but not the address of the following instructions since the computer control unit is designed to extract and carry out instructions from consecutive memory cells — except when directed otherwise by the instruction itself (branch). The correspondence between labels and numerical addresses thus obtained is summarized in a dictionary (*fig. 14*). This is used for substitution of the symbolic operands in the programme shown in fig. 12*a*.

Instruction address	Operation code	Index register	Operand address	
512	00	0	0525	
513	02	0	0527	
514	02	0	0032	
515	01	0	0528	
516	05	0	0527	
517	03	1	0530	
518	02	0	0527	
519	00	0	0032	
520	03	0	0526	
521	02	0	0032	
522	04	0	0529	
523	06	0	0515	
524	07	0	0000	
525	..	.	....	= "0"
526	..	.	....	= "1"
527	..	.	....	= "*P*"
528	..	.	....	= "*x*"
529	..	.	....	= "*n*"
530	..	.	....	= "*a*(I)"
.				.
.				.
.				.

Fig. 13. Machine-code programme for polynomial evaluation. The programme results from the translation of the assembly-language programme of fig. 12*a*. Column 3 designates which index register (if any) has to be used to modify the address part. A zero indicates "no modification". Addresses 525 etc. contain numerical values of the variables given in inverted commas.

Label	Address
AV	530
CYCLE	515
IR1	32
N	529
ØNE	526
PØL	527
X	528
ZERØ	525

Fig. 14. Dictionary of memory addresses associated with labels in assembly programme fig. 12*a*. IR1 has the fixed address 32.

Finally the substitution of numerical operation codes (fig. 12*b*) for the symbolic operators yields the programme fig. 13 which specifies in numerical code the contents of each memory cell used in this programme. Except for a possible conversion to binary code this is what actually resides in the memory when the programme is run. To enter it requires no more than an elementary input programme to read tape or cards on which the text of fig. 13 would be punched. This elementary input programme is usually present as a wired-in programme.

Programming languages and translators

At the end of this brief survey of coding in various languages it is interesting to retrace our steps and view programming languages in historical perspective [9]. The labour involved in computer language coding encouraged the early development of assembly languages. The assembly language provides access to every computer instruction and is thus almost as flexible as the computer language itself. Consequently, coding in computer language by explicit designation of computer addresses and instruction codes has practically vanished, except for fundamental programmes and subroutines for internal use by the system.

Statement-type languages evolved as an attempt to provide a means of writing programmes which were independent of any particular computer. Their development originated from the observation that the translation of a statement such as

$$z := x + y$$

into a sequence of instructions

$$A := x$$
$$A := A + y$$
$$z := A$$

is a logical operation which can be expressed as an algorithm written in computer instructions. The translation can therefore be done by the computer itself.

FORTRAN was devised as a language which would allow the writing of parenthesized arithmetical expressions. The main problem was the design of the compiler which would provide the translation into a corresponding sequence of computer instructions — called the object programme. The imperfections and inadequacies of FORTRAN together with the experience gained in compiling techniques encouraged the study of the structure and description of programming languages. ALGOL 60 is one of the results of these studies and newer versions are being devised. At the same time, the advances made in syntactic analysis allow part of the process of compiler writing to be automated [9].

As programmes are available to do the conversion for him the programmer no longer needs a detailed knowledge of the machine code. There is of course an increase in computer running time since a programme in a statement-type language must first be translated by a compiler and the object programme produced can hardly be as efficient as a well-written assembly programme. This loss of efficiency is generally unimportant compared with the time saved in programme writing and debugging. Nevertheless it is desirable to be aware of the sequence of computer instructions that results from the translation of certain statements.

As an illustration the ALGOL 60 statement 6 in fig. 8:

$$c[i,j] := c[i,j] + a[i,k] \times b[k,j],$$

might be compiled into an object programme represented by the sequence of operations:

1	$r := (i-1) \times n + k$
2	$y := a1(r)$
3	$r := (k-1) \times p + j$
4	$y := b1(r) \times y$
5	$r := (i-1) \times p + j$
6	$y := y + c1(r)$
7	$r := (i-1) \times p + j$
8	$c1(r) := y,$

in which the variables are defined as follows:

r	index of the vector representing the matrix in computer representation,
$a1, b1, c1$	vector components corresponding to matrix components a, b, c as described above,
y	intermediate variable.

The index conversion

$$r := (i-1) \times p + j$$

which establishes the correspondence between the matrix C and its vector representation C_1 is repeated because the compiler is not expected to scan the ALGOL 60 statement to discover whether the same component $c(i,j)$ has been designated twice. Moreover, the statement is part of the inner cycle (k) of the procedure (fig. 5*a*) and the index conversion is therefore repeated $m \times 2n \times p$ times in the programme even though it does not vary within the inner cycle.

Attention to such features may improve programme efficiency with little effort. In the example given, a slight modification of the flow chart of fig. 5*a* and the ALGOL 60 programme of fig. 8 produces a considerable decrease in running time (almost 50 % on the Electrologica X8-ALGOL 60 system). The modification consists in the use of an auxiliary variable x to accumulate

[9] J. Loeckx and M. Sintzoff, Les langages de programmation et leur traitement automatique, Rev. MBLE **9**, 126-146, 1966.

the partial products within the k cycle (*figs. 15* and *16*). The increase in programme length is merely an assignment of the value of x to $c(i,j)$ upon exit from the k cycle.

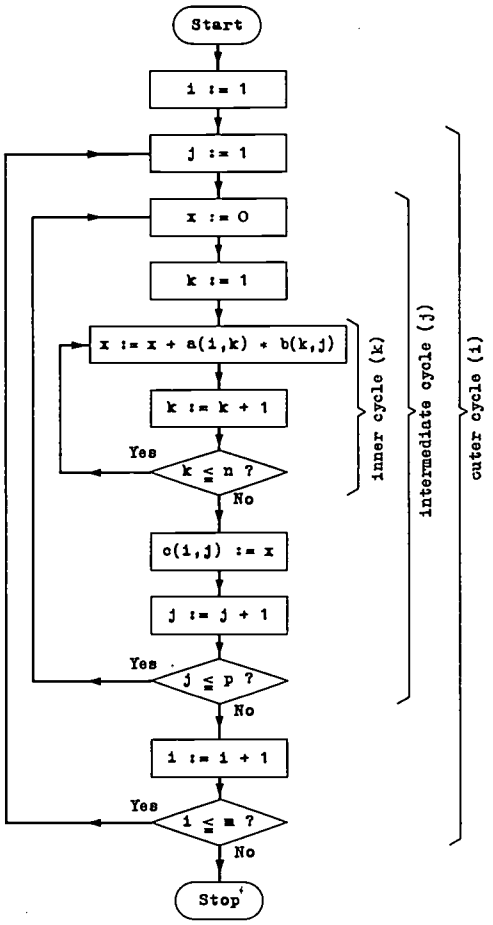


Fig. 15. Flow chart of improved procedure for matrix product computation in statement-type language.

```
for i := 1 step 1 until m do
for j := 1 step 1 until p do
    begin
      x := 0;
      for k := 1 step 1 until n do
        x := x + a[i,k]×b[k,j];
      c[i,j] := x
    end
```

Fig. 16. Main body of improved ALGOL 60 procedure to compute a matrix product $C = A \times B$.

Conclusion

The development of computer programmes to solve problems arising from scientific investigations is a lengthy process and the main difficulties usually arise in unexpected areas. The services offered by computer centres and the use of statement-type languages which are independent of a particular computer have made computers fairly accessible. Nevertheless there is no substitute for the hard work involved in formulating the problem and in thinking out the solution. These considerations may sound trivial, but the invention of the computer has given rise to a myth which is still widely held. It is tempting to regard a computer centre as a universal problem solver that will accept vaguely defined problems and produce solutions at tremendous speed. The exact opposite is probably the case. True, a computer provides the means of carrying out instructions speedily and reliably, but before a programme becomes operational a lengthy and painstaking development must be undertaken. We have attempted to highlight some of the facets of this development.

Dividing the development into stages is not done just for convenient presentation. It also gives a series of staging-posts at which specific results have been achieved, and responsibilities can be transferred. The author of the problem deals with the programme specification; the analyst defines the procedure layout; the coder checks programme validity. Just who does what is of no concern here: a programmer may assume responsibility for both analysis and coding. The most vital interface is the programme specification. Any residual inaccuracies at this point give rise to conscious or unconscious decisions in the programming which result in a programme which will very likely not produce the required results.

Summary. The article analyses the stages involved in the development of computer programmes for the solution of scientific problems. It is particularly directed to scientists and engineers to whom computing centres offer increasingly powerful facilities and from whom in turn an increasing participation is required. A methodical approach is advocated in which the development of the programme is tackled in three stages: systems analysis, procedure lay-out, and coding. This compels the exercise of rigour at each step. Coding by the user is usually restricted to statement-type languages. Symbolic and machine languages are briefly touched upon in order to explain the function of translators and thus to indicate what happens to a programme when run on a computer.

Recent scientific publications

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

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

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

G. A. Acket: Galvanomagnetic effects of hot electrons in gallium arsenide.
Philips Res. Repts. **23**, 317-331, 1968 (No. 4). *E*

F. T. Backers: Ultrasonic delay lines for the PAL colour-television system.
Thesis, Eindhoven 1968. *E*

W. J. H. Beekman: On a new type of X-ray generator.
Thesis, Eindhoven 1967. *E*

V. Belevitch: Conference networks and Hadamard matrices.
Ann. Soc. Scient. Bruxelles **82**, 13-32, 1968 (No. 1). *B*

G. Blasse: Fluorescence of niobium-activated antimonates and an empirical criterion for the occurrence of luminescence.
J. chem. Phys. **48**, 3108-3114, 1968 (No. 7). *E*

G. Blasse: Some considerations and experiments on concentration quenching of characteristic broad-band fluorescence.
Philips Res. Repts. **23**, 344-361, 1968 (No. 4). *E*

G. Blasse & A. Bril: Fluorescence of Eu^{3+}-activated sodium lanthanide titanates ($NaLn_{1-x}Eu_xTiO_4$).
J. chem. Phys. **48**, 3652-3656, 1968 (No. 8). *E*

G. Blasse, W. L. Wanmaker (Philips Lighting Division, Eindhoven) **& J. W. ter Vrugt** (Philips L.D.): Some new classes of efficient Eu^{2+}-activated phosphors.
J. Electrochem. Soc. **115**, 673, 1968 (No. 6). *E*

G. Bosch: A Hall device in an integrated circuit.
Solid-State Electronics **11**, 712-714, 1968 (No. 7). *E*

L. K. Brundle & N. J. Freedman: Magnetostatic surface waves on a y.i.g. slab.
Electronics Letters **4**, 132-134, 182 (erratum), 1968 (Nos. 7, 9). *M*

K. H. J. Buschow: Crystalline Stark splitting in $TmAl_3$.
Z. phys. Chemie Neue Folge **59**, 21-26, 1968 (No. 1-4). *E*

H.-J. Butterweck: Mode filters for oversized rectangular waveguides.
IEEE Trans. **MTT-16**, 274-281, 1968 (No. 5). *E*

T. Chisholm: Influence of space charge on the Langmuir limit for focused electron beams.
IEEE Trans. **ED-15**, 374-381, 1968 (No. 6). *M*

A. Dekker & A. Middelhoek (Philips Electronic Components and Materials Division, Development Dept. Electrolytic Capacitors, Zwolle): The anodic oxidation of aluminium.
Philips Res. Repts. **23**, 342-343, 1968 (No. 4).

F. C. M. Driessens, G. D. Rieck (Technical University Eindhoven) **& H. N. Coenen** (T.U.E.): Phase equilibria in the system cobalt oxide/copper oxide in air.
J. inorg. nucl. Chem. **30**, 747-753, 1968 (No. 3). *E*

W. F. Druyvesteyn, A. J. Smets & C. A. A. J. Greebe: Experiments on Alfvén-wave resonances in liquid sodium.
Philips Res. Repts. **23**, 332-341, 1968 (No. 4). *E*

A. S. Fruin & J. Jackson: The effect of heat treatment on the conductivity of copper foil at low temperatures.
Cryogenics **8**, 254, 1968 (No. 4). *M*

G. Groh: Multiple imaging by means of point holograms.
Appl. Optics **7**, 1643-1644, 1968 (No. 8). *H*

J. Jackson & A. S. Fruin: Heat transfer to liquid helium from bare and enamelled wires.
Proc. int. Conf. on magnet technology, Oxford 1967, p. 494-495. *M*

J. E. Knowles: The magnetostatic energy associated with a polycrystalline ferrite.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 987-994, 1968 (No. 8). *M*

J. Loeckx & M. Sintzoff: Génération mécanique de traducteurs pour langages formels.
Actes 5ème Congrès AFIRO, Assoc. Franç. Inform. Paris, 1967, p. 145-148. *B*

F. K. Lotgering & R. P. van Stapele: Cation-anion distances in chalcogenide spinels.
Mat. Res. Bull. **3**, 507-511, 1968 (No. 6). *E*

J. M. van Nieuwland, H. L. Hagedoorn, N. Hazewindus & P. Kramer: Magnetic analogue of electric field configurations applied to the central region of an A.V.F. cyclotron.
Rev. sci. Instr. **39**, 1054-1055, 1968 (No. 7). *E*

W. C. Nieuwpoort & G. Blasse: Remark on the calculation of cohesion energies using the ionic model.
J. inorg. nucl. Chem. **30**, 1635-1637, 1968 (No. 6). *E*

W. J. Oosterkamp, A. P. M. van 't Hof, W. J. L. Scheren & P. G. A. Teunissen: New methods for television display of roentgenological information in black and white and in color.
103rd Tech. Conf. SMPTE, Los Angeles 1968, preprint No. 103-114, 8 pp. *E*

P. Penning & A. H. Goemans: Contrast in X-ray topographs of thin crystals.
Phil. Mag. **18**, 297-300, 1968 (No. 152). *E*

R. Perrin: Contribution à l'étude des propriétés photoélectriques du rubidium pur, préparé et conservé dans l'ultravide.
C.R. Acad. Sci. Paris **267B**, 58-60, 1968 (No. 1). *L*

A. Rabenau & H. Rau: Hydrothermale Züchtung von Goldkristallen.
Naturwiss. **55**, 336-339, 1968 (No. 7). *A*

E. Schwartz: Broadband matching of resonant circuits and circulators.
IEEE Trans. **MTT-16**, 158-165, 1968 (No. 3). *A*

J. G. Siekman & R. E. Morijn: The mechanism of welding with a sealed-off continuous CO_2-gas laser.
Philips Res. Repts. **23**, 367-374, 1968 (No. 4). *E*

J. G. Siekman & R. E. Morijn: A simple power-output meter specially designed for continuous laser beams.
Philips Res. Repts. **23**, 375-387, 1968 (No. 4). *E*

F. L. H. M. Stumpers, N. van Hurck & J. O. Voorman: A receiver for AM and for single sideband AM with a partially suppressed carrier.
E.B.U. Rev. 108-A, 93-94, 1968. *E*

C. van Trigt: Asymptotic expansions of the integrated absorptance for simple spectral lines and lines with hyperfine structure and isotope shifts.
J. Opt. Soc. Amer. **58**, 669-678, 1968 (No. 5). *E*

A. G. van Vijfeijken: On the theory of vortices in type-II superconductors.
Thesis, Amsterdam 1967. *E*

W. L. Wanmaker & D. Radielović (Philips Lighting Division, Eindhoven): Luminescence of phosphates.
Bull. Soc. Chim. France 1968, 1785-1791 (No. spécial).

W. L. Wanmaker & J. W. ter Vrugt (Philips Lighting Division, Eindhoven): Luminescence of calcium orthophosphates activated with divalent europium.
Philips Res. Repts. **23**, 362-366, 1968 (No. 4).

Contents of Mullard Technical Communications **10**, No. 93, 1968:

Contents of Valvo Berichte **14**, No. 3, 1968:

Volume 29, 1968, No. 11 pages 325-352 Published 24th January 1969

PHILIPS TECHNICAL REVIEW

VOLUME 29, 1968, No. 12

The iodide discharge lamp

L. B. Beijer, C. A. J. Jacobs and T. Tol

Gas discharge lamps have the highest luminous efficiency of all radiation sources: the low-pressure sodium lamp excels in this respect with efficiencies of 130 lm/W, and even up to 175 lm/W when special heat insulation is used. Until recently the only elements that could be used for the excitation of light in gas discharge tubes were sodium, mercury and the inert gases. The other elements emitting suitable radiation are either not volatile enough or else they attack the types of glass which can be used for the discharge tube. The article below describes a method which gets around these limitations, making it possible to produce radiation sources whose spectral distribution can be freely chosen within certain limits and which give an efficiency only surpassed by that of the sodium lamp.

Principle of the iodide lamp and its application in the HPI lamp

The first high-pressure mercury-vapour lamp (the HP lamp[*]) for street lighting appeared in about 1933. Right from the start a great deal of effort was spent in trying to improve the colour rendering of this lamp. In the early attempts at improvement the metals zinc and cadmium, which emit blue and red radiation, were added to the mercury discharge in order to fill up the "holes" in the mercury spectrum. This was not successful mainly because these elements attacked the quartz glass at the temperature required to obtain sufficient vapour pressure. All the work done appeared to indicate that only mercury and the inert gases were suitable for the excitation of light in high-pressure discharge lamps (although sodium could be used in low-pressure discharge lamps).

In about 1937 a notable success was achieved by an entirely different approach, in the development of the HPL lamp. In this lamp the ultra-violet radiation emitted by the mercury is converted into additional visible radiation by means of a fluorescent powder on the outer bulb [1]. This lamp quickly superseded the HP lamp. Nevertheless, its colour rendering and also its luminous efficiency still left something to be desired.

Ir. L. B. Beijer, C. A. J. Jacobs and Dr. T. Tol are with the Philips Lighting Division, Eindhoven.

A few years ago it was found that the earlier conclusions needed to be revised. It was established — or rather rediscovered, since the possibility had been indicated in a patent as long ago as 1900 — that many elements can be evaporated to give a sufficiently high vapour pressure if they are introduced in the form of special volatile *compounds*. When this is done it is possible to make the vapour pressure of the element higher *at the centre* of the discharge than it would be if the element itself were added. The vapour pressure of the element is not increased in the rest of the tube, but this is not necessary, since the most active part of the gas discharge is found at the centre of the discharge, where most of the radiation is produced. This special method permits the use of elements which are not themselves sufficiently volatile, and it also removes the danger of chemical attack. It was found that *iodides* are generally very suitable for this purpose.

High-pressure mercury-vapour lamps were in fact made in which the desired additional radiation was produced by the addition of an iodide. The development did not rest long at this stage, however, for it

[*] In the United Kingdom the lamp types HP, HPL and HPI are designated as MB, MBF and MBI respectively.

[1] See J. L. Ouweltjes, W. Elenbaas and K. R. Labberté, Philips tech. Rev. **13**, 109, 1951/52, and also W. Elenbaas, Fifty years of the high-pressure mercury-vapour lamp, Philips tech. Rev. **18**, 167-172, 1956/57, in particular p. 171.

soon became apparent that much greater advantages were to be gained by using the iodides not simply to supplement the mercury spectrum, but as the actual source of the radiation. This results in a considerable gain in efficiency, since it enables elements to be used which, unlike mercury, emit *resonance radiation* in the visible part of the spectrum.

Resonance radiation is radiation which is generated when an electron returns from an excited state to the *ground level*. Generally speaking, electron transitions of this nature yield the strongest emission lines. To make use of resonance lines it is necessary to ensure that the absorption of these lines by the gas itself (self-absorption) is kept to a minimum. In low-pressure discharge lamps, like the low-pressure sodium lamp, the fluorescent lamp, etc. the vapour pressure of the emitting element is kept low for this reason [2].

The iodide discharge lamp to be described here is a high-pressure discharge lamp which has elements with a lower excitation level than mercury added to it in the form of iodides to give resonance radiation. The iodides are added in small enough quantities to ensure low self-absorption. The excitation level of the added elements must be much lower than that of mercury to make sure that the mercury spectrum is *not* emitted even though the mercury vapour pressure is much higher. In these lamps the mercury has an entirely different function: it takes little or no part in the excitation and ionization processes but helps to maintain the discharge under high pressure and high temperature; it serves as a "buffer gas". A high temperature is needed for sufficient evaporation of the iodides.

We first used the principle outlined here in the development of a lamp for road lighting, the HPI lamp, which offers both a better luminous efficiency and a better colour rendering than the HPL lamp. Three elements were chosen as the radiation emitters for this lamp: Na, Tl and In, giving reddish-yellow, green and blue light respectively. As in the HP and HPL lamps, argon was adopted as the ignition gas. With careful choice of the quantities of the three iodide additives [3], efficiencies from 80 to 110 lm/W are obtained, compared with 55 to 60 lm/W from the HP and HPL lamps. The colour-rendering index is about 70 compared with 25 for the HP lamp and 45 for the HPL lamp.

HPI lamps of 400, 2000 and 5000 W (*fig. 1*) are al-

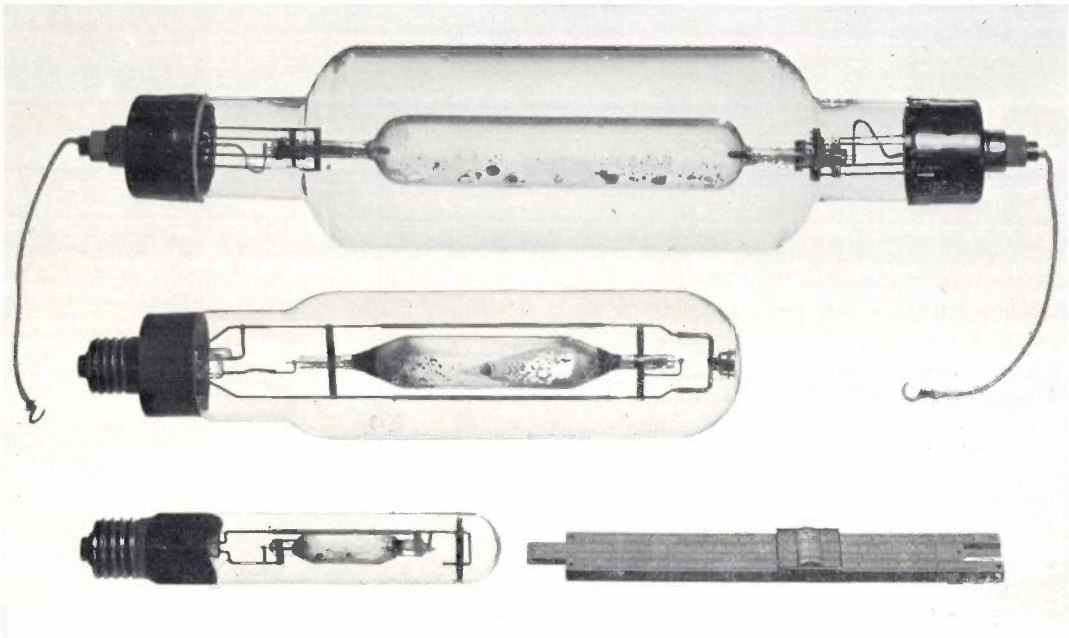


Fig. 1. The 400, 2000 and 5000 W HPI lamps in ascending order. The luminous efficiencies are 80, 95 and 110 lm/W respectively, the operating voltages are 125, 245 and 550 V, and the currents are 3.5, 9 and 10 A. The lamps have a cylindrical outer bulb of hard glass; the inner bulb is of quartz glass. The outer bulb of the 400 W lamp is evacuated and contains a getter; the outer bulb of the 2000 and 5000 W lamps is filled with nitrogen to improve the heat conduction.

[2] A more detailed discussion of low-pressure and high-pressure lamps is given in G. Heller, Comparison between discharge phenomena in sodium and mercury vapour lamps, I and II, Philips tech. Rev. **1**, 2-5 and 70-75, 1936.

[3] For optimization of the quantities to be added (the "dose"), see T. Holmes and J. B. de Boer, Conf. Associated Public Lighting Engineers, Edinburgh 1964, paper No. 2.

[4] L. B. Beijer, Proc. 7th Int. Conf. on phenomena in ionized gases, Belgrade 1965, Vol. III, p. 182.

ready on the market. They are not only used for road lighting but, because of their good colour rendering, for interior lighting as well.

An account of investigations which have been carried out on the HPI lamp will be given in the last section of this article. However, one or two of the investigations do help to give a clearer idea of the principle underlying the iodide discharge lamp, and are perhaps best dealt with here.

perimental HPI lamp was found to be about 1000 °K, and about the same value was measured for the HP lamp. At the axis of the discharge, we measured a temperature of about 4800 °K for the HPI lamp, and about 5700 °K for the HP lamp [4]. The temperature of the arc is thus substantially lower in the HPI lamp. Since both lamps were designed to have about the same concentrations of Hg atoms, and since we do not want any mercury radiation from the HPI lamp, a lower arc

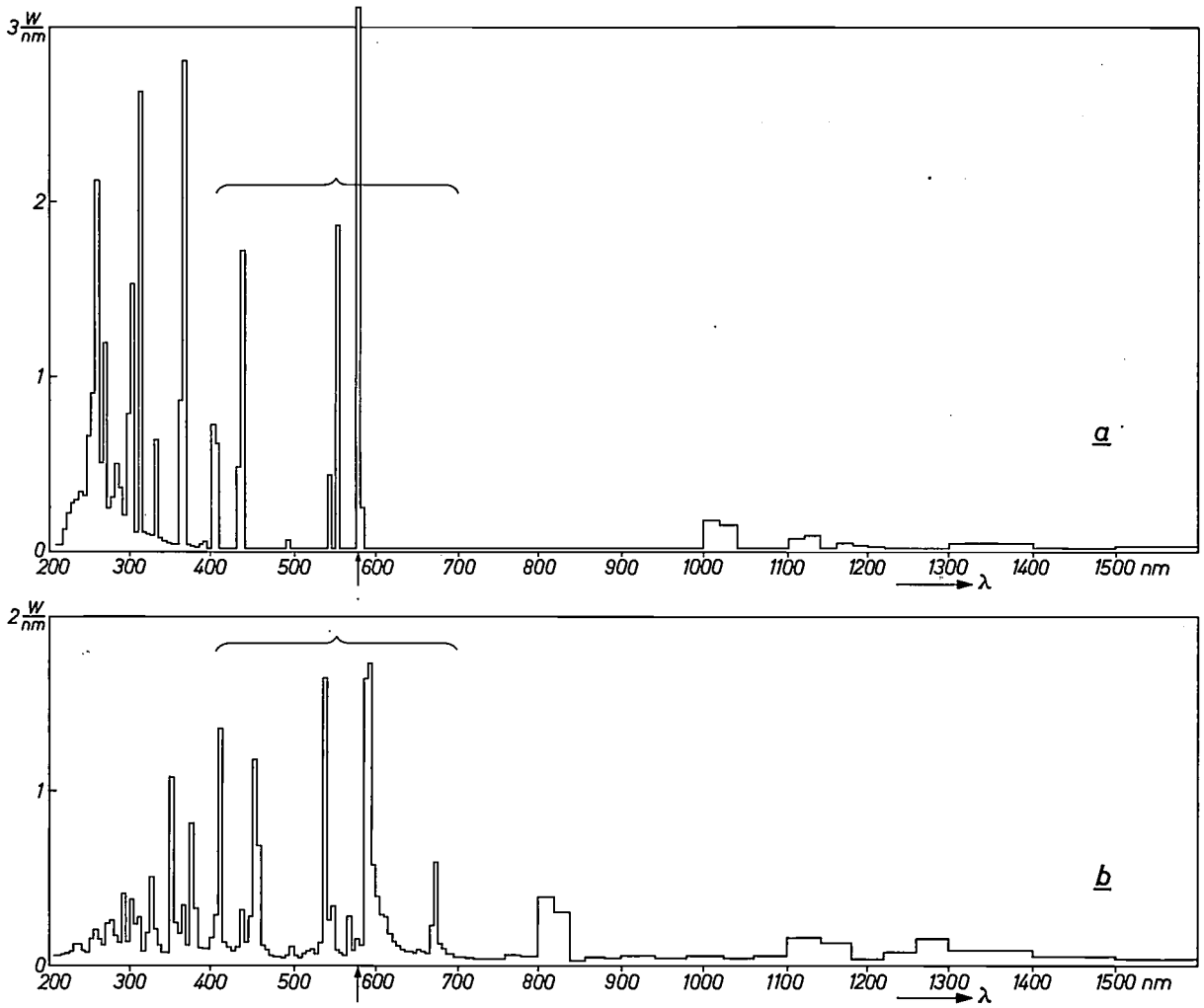


Fig. 2. Spectral energy distributions of two geometrically identical 500 W discharge lamps whose bulbs are made entirely of quartz glass: *a*) HP lamp, *b*) HPI lamp. The visible part of the spectrum is indicated by the bracket. It can be seen that the HPI lamp emits less radiation in the ultra-violet than the HP lamp but rather more radiation in the infra-red. It is also noticeable that the emission in the visible part of the spectrum is more uniform in the HPI lamp than in the HP lamp.

For comparative experiments we used special HP and HPI lamps, both with an arc length of 7 cm and run at 500 W. The production versions have the outer bulb of hard glass and the inner bulb of quartz glass, but in the special lamps *both* bulbs were made of quartz glass to permit measurements over the widest possible range of wavelengths. The wall temperature of this ex-

temperature in the HPI lamp is a necessary condition.

The spectral energy distribution of the experimental HP and HPI lamps are shown in *figs. 2a* and *b* respectively. It can be seen that the spectrum of the HPI lamp consists mainly of radiation contributions from Na, Tl and In, and not from Hg. This is particularly clear from the low emission in the ultra-violet — the region in

which most of the radiation of the HP lamp is produced. The three yellow mercury lines at 577.0, 579.0 and 579.1 nm (not separated in the figure, see arrow) are also very strong in the HP lamp. The only indication of the presence of these emission lines in the spectrum of the HPI lamp is a very small peak, which is only 0.037 times the height of the corresponding peak in the HP spectrum.

This experimentally determined ratio of radiant intensities was verified by a calculation in the following way. Both the upper and lower energy levels of the lines of interest are well above the ground level, and this means that their occupation densities will decrease very rapidly with temperature. We may therefore assume that the radiation is produced mainly in the hot core of the arc and that this radiation is only absorbed to a slight extent as it leaves the arc. As a first approximation therefore, the only temperature we need to know is that at the core of the arc. By inserting in Boltzmann's formula the measured temperatures of 4800 °K for the HPI lamp and 5700 °K for the HP lamp, together with the known values of the appropriate excitation levels, we found that the ratio of the radiant intensities was 0.038, which is in good agreement with the experimental value.

Further discussion of the principles; other applications

Until recently, only a few elements were considered for the excitation of radiation in gas discharge lamps. As we have seen, the reason for this limited choice was that the other elements were either not volatile enough or too chemically active, or both. The first difficulty can be avoided by introducing the element in the form of an iodide — provided the iodide is

a) reasonably stable at the temperature of the wall, but
b) more volatile than the element itself, and
c) largely dissociated at the temperature encountered at the axis of the tube.

The condition that the iodide should be more volatile arises from the fact that the vapour pressure of the element in equilibrium can never be higher than the saturation vapour pressure of the iodide at the wall temperature. The significance of the stability of the iodide near to the wall becomes clearer when we realize that, if the iodide is too unstable, the vapour pressure of the element may become higher than the local saturation pressure, resulting in the partial condensation or deposition of the element. In these circumstances the required pressure cannot be reached at the axis.

The second difficulty limiting the choice of elements — the high chemical activity — is also overcome if the iodide is very highly stable at the wall temperature.

For introduction in the form of iodides we chose elements with very strong emission lines in the wavelength regions of practical importance. The elements that seem most suitable, partly because of the properties of their iodides, are Li, Na, Tl, In, Ga and Pb (*Table I* and *fig. 3*). The iodides of Tl, In, Ga and Pb are more volatile and the iodides of Li and Na are less active than the respective elements.

The wider choice of elements, with accurate control of the quantity of the iodide, allows us to increase the vapour pressure of the element at the centre of the tube to any value we require, provided it remains below the saturation pressure of the iodide at the wall temperature. This means that we can control the spectral energy distribution of the iodide lamp within certain limits and match the lamp to the needs of a particular application. In the development of the HPI lamp the choice of the iodides and the quantity used was largely determined by the need for the best possible luminous efficiency with reasonably good colour rendering. In the development of a *projector lamp* (CSI), where the standard of colour rendering has to be higher, the same three iodides were no longer sufficient and we had to use in addition the red-emitting lithium, as the iodide. The resultant improvement in the colour rendering of the CSI lamp, however, was gained at the expense of its luminous efficiency.

We have also made lamps that produce coloured light with a high efficiency (*floodlight lamps*): a reddish

Table I. Some data relating to the elements with the strongest emission lines in the wavelength regions of practical interest, and their iodides. T is the temperature at which the saturation pressure is 1 millibar, and λ is the wavelength of the strongest emission lines.

	T (°K)	λ (nm)
Li	980	323.3 - 610.4 - 670.8
LiI	985	
Na	705	589.0/589.6
NaI	1026	
Tl	1082	276.8 - 351.9 - 377.6 - 535.0
TlI	702	
In	1510	303.9 - 325.6 - 410.2 - 451.1
InI	624	
InI_2	637	
InI_3	510	
Ga	1690	287.4 - 294.4 - 403.3 - 417.2
GaI_3	430	
Pb	1226	283.3 - 363.9 - 368.3 - 405.8
PbI_2	743	
I		178.3 - 183.0 - 206.2
I_2	308	
Hg	394	185.0 - 253.7 - 365.0/365.5/366.3 - 404.7 - 435.8 - 546.1 - 577.0/579.0/579.1
HgI_2	425	

yellow lamp with NaI, a green lamp with TlI, and a blue lamp with InI. Other important applications are to be found in the field of *photochemical reactions*. Lamps we have made for these applications include the types that produce Ga and Pb radiation, mainly in the wavelength region between 340 and 450 nm. Used for the activation of diazo reactions (in photographic reproduction) and for the polymerization of photosensitive lacquer (in photo-etching processes), these lamps give an efficiency 50% higher than that of the mercury lamps which had to be used before. Another interesting application of iodide discharge lamps is for *pumping solid-state lasers*: a super-high-pressure lamp [5] containing TlI has been used for pumping ruby lasers.

One special application is that of the *spectral lamps* used for calibrating the wavelength adjustment of spectroscopes, spectrophotometers, etc. It is important in this application to minimize the broadening and displacement of the emitted lines. This is done, for example, by keeping the concentrations of the emitting elements low. The efficiency of such lamps is clearly not very important. Even so, in the past the makers have been limited to a few elements for these lamps, while the users on the other hand would like to have a sufficiently wide range of lamps to permit measurements and calibrations to be made at all points in the spectrum. This application has also benefited from the wider choice of elements that can be used with the iodide technique. Since high pressures are undesirable owing to the line-broadening they cause — and since the efficiency is not important — these lamps are produced as low-pressure types (with no buffer gas) [6]. In practice an adequate coverage is obtained by adding lamps with Tl, In, Ga and Bi to the existing range.

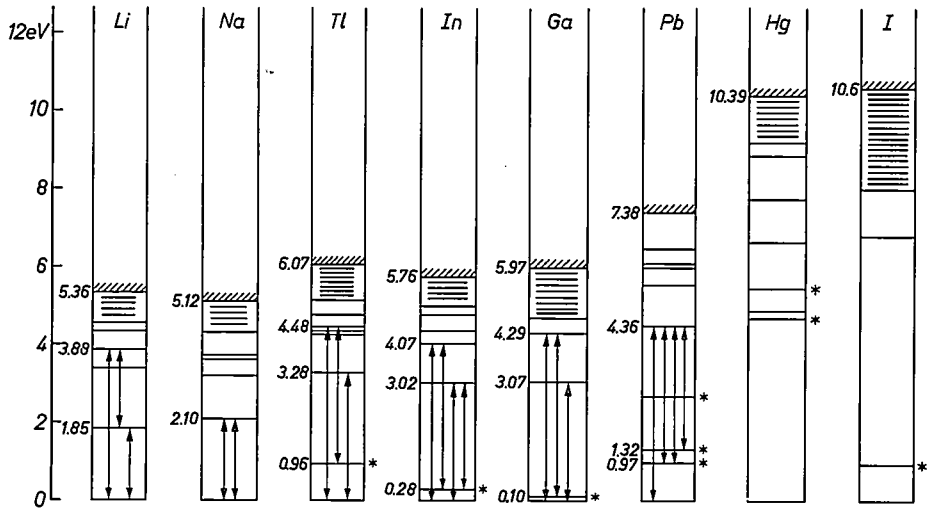


Fig. 3. The principal energy levels of the elements to be used for producing the light. The metastable levels, which cannot be "emptied" by emission, are indicated with an asterisk. It can be seen that Hg and I (apart from a metastable level at 0.94 eV) are more difficult to excite and to ionize than the other elements.

Some of the lamps mentioned above can be seen in *fig. 4*.

Investigations on the HPI lamp

Of the iodide discharge lamps we have developed, the HPI lamp is the one which has been studied in the greatest detail. These studies included comparisons with the HP lamp. The temperature at the core of the two discharges and the spectral distribution have already been mentioned. We shall now discuss in turn the composition of the discharge medium, the division of the input power between radiation and heat (power balance), and the ignition.

Composition of the discharge medium

It is not at all easy to give an exact description of the chemical equilibrium situations in discharge lamps with

[5] The principle of the super-high-pressure lamp has been described by E. G. Dorgelo in Philips tech. Rev. 2, 165, 1937.

[6] In this case the role of the iodides appears to be rather different from that in the high-pressure lamps — the details are not yet completely explained.

such a mixed filling as in the HPI lamp, including as it does Hg, HgI_2, I_2, I, NaI, Na, TlI, Tl, InI and In. A good insight can nevertheless be obtained by means of a special thermodynamic approach to the problem. The chief results of these thermodynamic calculations are given below; the actual thermodynamic treatment is explained in the Appendix.

The following picture is obtained:

In the core of the discharge, that is to say in the temperature zone from about 3500 to about 4800 °K, the three iodides NaI, TlI and InI are completely dissociated; at these temperatures the iodine is liberated in the atomic state.

At the wall temperature of about 1000 °K these three iodides are only very slightly dissociated. Any iodine liberated by the dissociation of these iodides combines to form HgI_2 because of the excess of Hg.

The HgI_2 is not very stable and the result is that free iodine is formed in the lamp at temperatures above 1300 °K. Between 1300 °K and 1500 °K the iodine occurs in the molecular form and above 1500 °K it occurs in the atomic form. The fact that there is only a narrow temperature zone in which molecular I_2 occurs (and then only in a very low concentration) is important since molecular I_2 absorbs in the visible region of the spectrum.

Effective dissociation of NaI, TlI and InI begins in the temperature range above about 2000 °K.

The approximate partial pressures are:

Hg:	4 bars,
Na + NaI:	1 millibar,
Tl + TlI:	200 millibars,
In + InI:	50 millibars.

Since the sum of the partial pressures of the element plus the corresponding iodide is constant over the whole temperature range, this value is also the partial pressure of the element on *complete dissociation* of the iodide in the central region of the discharge. For Tl and In this pressure is substantially higher than the maximum pressure that can be reached with the addition of the element itself. This is not the case with Na; for this element the iodide technique is desirable only because it is so chemically active.

With Tl and In the partial vapour pressure in the core is so high that one might think that the self-absorption of the Tl and In lines would be too high. In the surrounding *colder* part of the lamp, however, where the self-absorption mainly occurs, the partial pressure of the absorbing element is *lower* than in the core because the equilibrium has shifted more towards the iodide. This explains why good results are obtained with Tl and In in spite of the self-absorption.

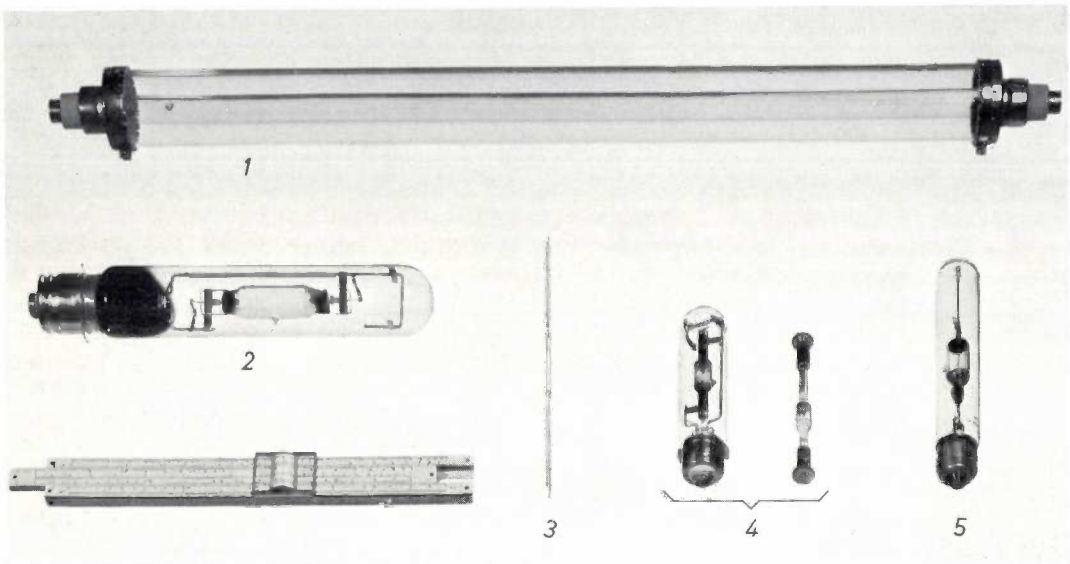


Fig. 4. Some iodide lamps for various applications. *1* printing lamp, 2000 W, with Ga and Pb as emitters (principally radiation between 340 and 450 nm). *2* 400 W lamp with the same emitters as *1*, used for the polymerization of photo-sensitive lacquers. *3* super-high-pressure lamp, 1000 W, for optical pumping of ruby lasers; Tl is the emitter. *4* 250 W projection lamps (CSI) containing LiI, NaI, TlI and InI; the right-hand lamp is made without outer bulb for optical reasons. *5* spectral lamp emitting In lines. This lamp differs from the others in that a low pressure discharge takes place and no mercury is used.

Power balance

As fig. 2 shows, the HPI lamp emits less radiation in the ultra-violet than the HP lamp. In the infra-red region of the spectrum the opposite situation holds. However, partly because the HPI lamp converts a greater part of the input power into radiation than the HP lamp does, the overall result is that the HPI lamp radiates considerably more in the visible region of the spectrum. With the two experimental 500 W lamps, the radiated power in the visible region is 120 W for the HPI lamp and 73 W for the HP lamp. In both cases the effective spectral sensitivity of the eye (relative spectral energy distribution of the radiated light) is about the same, so that the improvement in luminous efficiency is about 60%. As we shall explain in the following, this gain in luminous efficiency is even greater at higher powers.

From radiation measurements on vertically suspended high-pressure mercury lamps, Elenbaas [7] has calculated that the quantity of radiation Σ produced per unit length of the discharge column depends in a very simple way on the power P dissipated per unit length. Taking W/cm as the unit, the relation becomes:

$$\Sigma = 0.72 \times (P - 10). \quad \ldots \quad (1)$$

We may interpret this by saying that, whatever the dissipated power may be, there is always a loss of about 10 W/cm due to heat conduction from the core of the arc to the wall of the discharge tube (the effect of *convection* is negligible in vertically suspended lamps). Of the remainder of the input power $(P - 10)$ converted into radiation in the core of the arc, about 28% is lost due to absorption by the mercury vapour, the quartz glass of the bulb wall and the surrounding air.

We have performed similar measurements on the HPI lamp. Reproducible results are more difficult to obtain here than with the HP lamp, since the iodides show a tendency to distribute themselves non-uniformly over the length of the tube. However, the results of the measurements do indicate that in this case as well, for power inputs between 30 and 75 W/cm there is a simple relation between the radiation Σ produced per cm and the power P dissipated per cm:

$$\Sigma = 1 \times (P - 20). \quad \ldots \quad (2)$$

Comparison with equation (1) shows that the HPI lamp loses twice as much heat per cm (20 W/cm) by conduction as the HP lamp. On the other hand, the radiation produced in the core is not absorbed at all.

Further examination of the two equations reveals that when P is increased, the efficiency of the HPI lamp increases more than that of the HP lamp. This indicates that the power supplied to the lamp should be dissipated at the smallest possible electrode spacing. However, the electrodes should not be mounted so close to one another that a lot of the radiation is trapped and absorbed at the ends of the discharge tube. Because of this limitation a relatively low value of P has to be accepted with low-power lamps.

Table II shows the power balance of the 500 W lamps used for the comparative experiments.

Table II. Distribution of the input power for the experimental 500 W HPI and HP lamps made entirely of quartz glass.

	HPI	HP
Input power	500 W	500 W
Distribution of input power		
a) not radiated from the discharge	185 W	110 W
comprising: estimated electrode loss	45 W	40 W
b) radiated from discharge	315 W	390 W
comprising:		
ultra-violet radiation	44 W	112 W
visible radiation	120 W	73 W
infra-red radiation	151 W	95 W
absorbed	—	110 W
Effective sensitivity of the eye	0.5	0.5
Luminous efficiency	87 lm/W	52 lm/W

Ignition of the HPI lamp

There are special problems in the ignition of iodide discharge lamps: they are more difficult to ignite than HP lamps. The difficulty stems from two main causes: the presence of iodine vapour, and the presence of hydrogen.

The reason why iodine vapour pushes up the ignition voltage is that iodine is an electronegative element, i.e. an iodine atom readily associates with an electron to form a negative ion. Consequently, many of the electrons liberated in an iodide discharge lamp are withdrawn from the ignition process, resulting in a drastic reduction of the electrical conductivity of the gas. Another serious and related problem is that the *alkali* atoms used in some iodide lamps (Na in the HPI lamp and Na and Li in the CSI lamp) so very easily disappear: some of them react with traces of oxygen to form alkaline oxides which are difficult to evaporate, and others diffuse through the quartz wall. In itself, the loss of the alkalis is not so serious, since it can be compensated by providing a large initial surplus, but the loss in alkali corresponds to an increase in the concentration of iodine and therefore to a further increase in the ignition voltage.

It was found that the production of hydrogen in HPI lamps is much greater than it usually is in high-pressure mercury-vapour lamps. Iodides are hygroscopic, and during the operation of the lamp they give off

[7] W. Elenbaas, Physica **4**, 413, 1937. A refinement of the equation discussed here is given in W. Elenbaas, Philips Res. Repts. **20**, 213, 1965.

water which dissociates. The oxygen then combines with one or more of the metals present in the discharge tube, leaving hydrogen gas behind. At room temperature the gas atmosphere now consists of a mixture of mercury vapour, argon and hydrogen, which is more difficult to ionize than a mixture of mercury vapour and argon. For argon (without mercury) we know that the presence of 0.01% of hydrogen is sufficient to cause a perceptible increase in the ignition voltage [8]. Thus, assuming the usual argon pressure of 30 millibars in the lamp, a hydrogen pressure of only 3×10^{-3} millibars is enough to cause an increase in the ignition voltage. In the HPI lamp the increase occurs at even lower hydrogen pressures, because the mercury-argon mixture appears to be more sensitive to hydrogen than argon alone. The iodides must therefore be very carefully dried and kept dry before being added to the lamp.

Another source of hydrogen is the quartz glass of which the discharge tube is made. Why so much hydrogen can enter the iodide lamp from the quartz glass has not yet been completely explained. It has been found that the evolution of hydrogen can be considerably reduced by prolonged outgassing of the quartz glass at the beginning of the production process.

The higher ignition voltage in iodide discharge lamps makes it necessary to use a somewhat more elaborate circuit than that of the HP lamp. With reliability particularly in mind, we have chosen the following two methods of ignition from the several methods available.

a) For the 400 W and 2000 W HPI lamps: an electronic ignition unit which applies voltage peaks of 3000 V at each half-cycle until the lamp is ignited, used in conjunction with conventional HP chokes. When existing HP installations have been fitted with this unit they can also be used with HPI lamps.

   A familiar difficulty with high-pressure mercury-vapour lamps is that it takes a fairly long time to reignite a lamp which has just been switched off and is still hot. This "brute force" ignition method just mentioned is not only very reliable for the ignition of a cold lamp, it also considerably reduces the time needed to reignite a hot lamp.

b) For the 5000 W HPI lamp: since there are no HP installations for lamps of this rating, a simpler solution is possible in this case, consisting of a transformer with a high open-circuit voltage (860 V r.m.s.). This voltage is high enough to ensure stable operation of the lamp and is suitable for igniting the lamp in conjunction with a glow-discharge starter.

Appendix: the gas composition in the HPI lamp

The thermodynamic calculations presented here, relating to the HPI lamp, are based on the assumption that the gas mixture in the high-pressure discharge lamp with iodides is to a good approximation in thermal equilibrium everywhere [9], and that the presence of the radiation field and gas-kinetic effects may therefore be neglected. The initial consideration is that there is a hot central region around the axis of the gas discharge tube and a colder outer zone, and that the two regions can be treated separately. It is known that the iodides of mercury are highly unstable compounds, and we may therefore expect that, in spite of the higher mercury concentration, there will be no iodide of mercury present in the hotter regions of the discharge tube. Similarly, there will be virtually no molecular iodine present there. In the colder regions, however, the mercury may be expected to react with iodine and molecular iodine will also be found.

We shall now consider first the hot central region of the gas discharge.

To illustrate the thermodynamic approach for this region, we take the equilibrium $\mathrm{InI} \rightleftarrows \mathrm{In} + \mathrm{I}$ as an example. The constant K_p of this equilibrium is:

$$K_p = \frac{p_{\mathrm{In}} p_{\mathrm{I}}}{p_{\mathrm{InI}}},$$

where p_{In} is the partial vapour pressure of In, etc. The partial vapour pressure of the iodine p_{I} is due *not only* to partial dissociation of InI, but also to partial dissociation of the other iodides. If α is the degree of dissociation of InI, we can then also write K_p as

$$K_p = \frac{\alpha}{1-\alpha} p_{\mathrm{I}}.$$

The vapour pressure of the In iodide no longer occurs in the right-hand side of this expression, and we therefore only need to know p_{I} to be able to calculate the degree of dissociation α as a function of the temperature T. The value of p_{I} is for the time being unknown, but we can set the existing partial vapour pressure of iodine between two estimated limit values. We take the upper limit as the value which the partial vapour pressure of iodine would have on complete dissociation of TlI, InI and NaI, rounded off to a power of 10 and expressed in bars. The temperature, which is 1000 °K at the wall and 4800 °K at the axis, is assumed to have a parabolic distribution. The lower limit is taken to be a factor of 10 lower. We thus arrive at the limit values for p_{I} of 100 and 10 millibars. Using each of these values we have calculated α as a function of T and the result is shown in *fig. 5*. The degrees of dissociation of NaI, TlI and $\mathrm{I_2}$ were calculated in the same way as a function of T. It can be seen from this figure that at temperatures above 2000 °K the $\mathrm{I_2}$ is already completely dissociated whereas the metallic iodides are hardly dissociated at all. We now take the "hot" region of the gas discharge to refer to those places where the temperature is greater than 2000 °K. Our tacit assumption thus far that all free iodine is atomic is therefore justified in this temperature range.

We can get to know something about the composition of the gas atmosphere in the colder region of the lamp, below 2000 °K, by taking the following approach. Much of the total amount of iodine present in the colder region of the discharge tube is combined with the elements Na, Tl and In. The iodine so combined (corresponding to 250 millibars) will not be taken into account since in this temperature range the chemical bonds are so strong that no perceptible changes in the degree of dissociation occur. The rest of the iodine consists of an amount β combined in $\mathrm{HgI_2}$, an amount γ in atomic form, and an amount $(1-\beta-\gamma)$ in the molecular form. The possibility of mercurous iodide being pres-

[8] M. A. Uman, Phys. Rev. 133, A 1266, 1964.
[9] See C. F. Gallo, Appl. Optics 5, 1285, 1966.

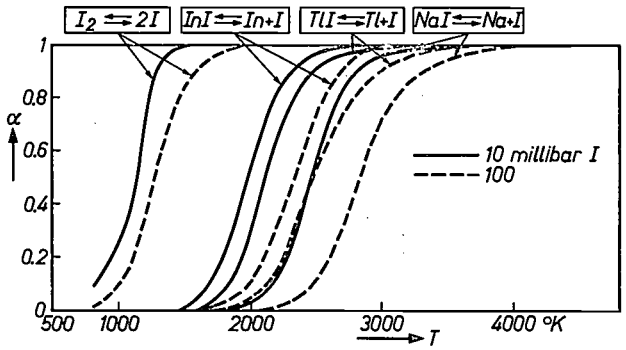


Fig. 5. The degree of dissociation α of I_2, InI, TlI and NaI as a function of temperature T, calculated for values of the partial vapour pressure of iodine p_I of 10 millibars (solid curves) and 100 millibars (dashed curves).

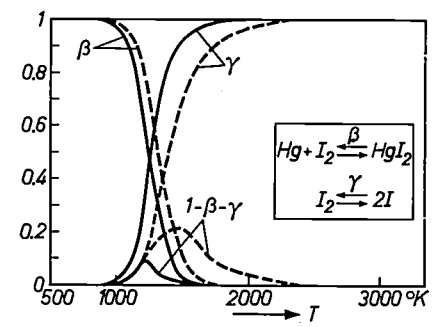


Fig. 6. Distribution of the iodine not combined with Na, Tl and In in the temperature range from 1000 to 2000 °K, similarly calculated for $p_I = 10$ millibars (solid lines) and 100 millibars (dashed lines). β is the fraction present in the form of HgI_2, γ is the fraction present in atomic form, and $1 - \beta - \gamma$ is the remaining fraction of molecular I_2.

ent is not taken into account since this iodide is more unstable than mercuric iodide. The values of β and γ can now be calculated with the aid of two equations, in which the quantity of mercury and the quantity of iodine, whether combined with mercury or not, and both expressed in terms of partial pressures, are used together as parameters. The mercury pressure in the HPI lamp is calculated from the known added quantity and the temperature distribution to be about 4 bars. We have again taken limiting values of 10 and 100 millibars for the iodine vapour pressure. The result is given in *fig. 6*. Below 2000 °K the formation of HgI_2 and of I_2 must be taken into account. It has in fact been found that the iodine at 1000 °K (i.e. close to the wall) has reacted virtually completely with the mercury. This ties up with the underlying idea of dividing the thermodynamic treatment into two parts. Fig. 6 also shows that we are justified in not considering HgI_2 when dealing with the temperature region above 2000 °K.

We now have all the information we need: the temperature distribution, the quantities added and the degree of dissociation of the iodides. In principle we can therefore calculate the partial vapour pressures of the iodides at every point in the HPI lamp. To simplify the work of calculation, it is useful to introduce the further assumption that, for every metal, the sum of its vapour pressure and the vapour pressure of its iodide is constant throughout the tube, since every metallic iodide molecule that goes to a hotter location and dissociates there will yield one metal atom, and *vice versa*. Since it is a fairly good approximation to take the total pressure as being constant throughout the discharge tube, this assumption implies that the sum of the partial vapour pressure of the mercury, the HgI_2 and the iodine is also constant throughout the tube.

We have thus arrived at the results summarized on page 358: the partial vapour pressure of mercury ($\approx$ 4 bars), Na + NaI ($\approx$ 1 millibar), Tl + TlI ($\approx$ 200 millibars) and In + InI ($\approx$ 50 millibars). In addition we have calculated the vapour pressures of the metals separately, and these are shown in *fig. 7* as a function of the distance to the axis of the discharge, with the partial iodine vapour pressure as a parameter.

Since the curves for $p_I = 100$ millibars and $p_I = 10$ millibars lie close together, it is evident that the calculated partial vapour pressures of the metals are not closely dependent on the partial

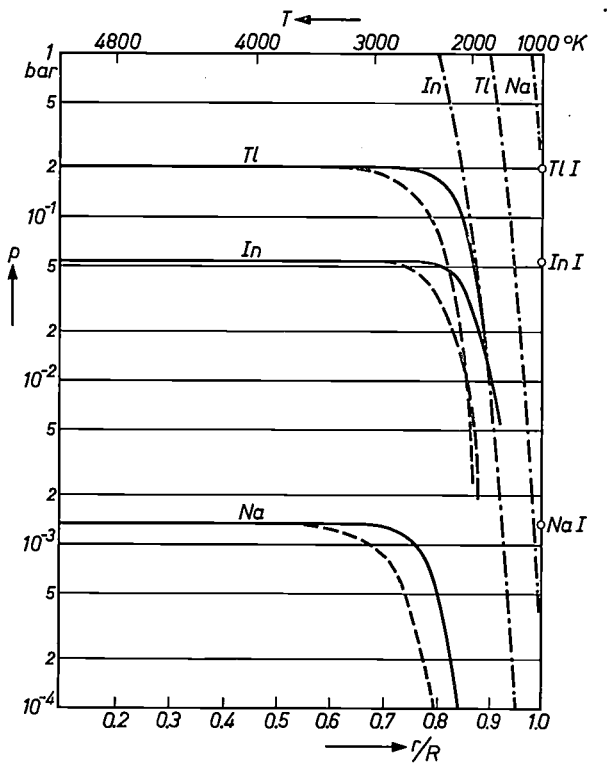


Fig. 7. Partial pressures of Na, Tl and In in the HPI lamp as a function of temperature (indicated at the top), and as a function of r/R, i.e. the ratio of the distance to the axis and the radius of the tube (indicated at the bottom). The partial pressures were calculated for iodine vapour pressures of 10 millibars (solid lines) and 100 millibars (dashed lines). The graph also gives curves of the saturation vapour pressures of the free elements (chain-dotted lines). If the elements themselves are introduced, i.e. not as iodides, the partial pressure cannot become higher than the saturation vapour pressure corresponding to the lowest temperature of the lamp. Taking this to be the wall temperature of 1000 °K, we see from the figure that Tl and In (*not* Na) have acquired a considerable higher pressure at the centre after being introduced in the form of iodides. Another important point is to find out whether the curve for the saturation pressure cuts the curve for the pressure of the element introduced as an iodide. Intersection of the curves indicates that the partial pressure of the element introduced as the iodide exceeds the saturation pressure at the temperature corresponding to the point of intersection. In this case a liquid or solid phase will occur and the desired object cannot be achieved. The shape of the curves seems to indicate that this undesired effect will occur with In. In practice, however, this complication has not arisen. In any case, this kind of situation can be prevented by appropriate measures based on these calculations.

vapour pressures of the iodine. We thus have a good picture of the composition of the vapour mixture in the HPI lamp.

Fig. 7 also shows the curves for the saturation vapour pressures of the free metals, determined by the temperature at each location in the tube. It can be seen that nearly all the calculated vapour pressures of the metals remain below the saturation values and are therefore practicable. Difficulties are only to be anticipated with In since the curve for the partial In vapour pressure corresponding to $p_I = 10$ millibars cuts, or perhaps grazes, the curve for the saturation vapour pressure of free In, and this makes the situation rather uncertain. If the curves really did intersect, liquid In would condense in the region to the right of the point of intersection. This would limit the value of the partial In pressure that could be built up in the core. However, practical experience with In has been very reassuring on this point. It does appear, however, that if a higher In vapour pressure should be required, the dissociation equilibrium would have to be shifted more towards the iodide; this can be done by using a higher iodine vapour pressure.

For completeness it should be mentioned that the iodine also reacts with the electrodes. The problems connected with these reactions are complicated and have not yet been resolved sufficiently to justify further discussion. We shall just note here that in most types of iodide lamp tungsten electrodes have been found to be resistant to iodine vapour and do not therefore extract much iodine.

Summary. A gas discharge lamp is described in which high-efficiency emission is obtained from elements such as Li, Na, Tl, In, Ga and Pb which have been introduced in the form of iodides. The lamp is designed as a high-pressure type in which mercury acts as a buffer gas, i.e. the mercury helps to maintain the discharge under high pressure and high temperature, but is hardly at all excited or ionized. The elements introduced as iodides could not be used in the elementary form because they are either not volatile enough or too liable to attack the materials of the lamp. By varying the concentrations of the iodides, the spectral distribution of the emitted radiation can be controlled within certain limits, so as to obtain optimum efficiency for a variety of applications. The lamps made include types for road lighting (HPI lamps), with Na, Tl and In as radiation emitters, projector lamps in which the red emission of Li is added to improve the colour rendering (CSI lamps), coloured lamps for floodlighting (reddish yellow with Na, green with Tl and blue with In), radiation sources for photochemical processes (using Ga and Pb) and for the activation of lasers (using Tl), and spectral lamps. Finally some investigations on the HPI lamp are discussed.

Fast logic circuits with low energy consumption

A. Slob

Logic circuits are widely used in computer and telecommunications systems, in measuring and control equipment, and in other applications. Their manufacture has shown a natural trend towards standardization and rationalization [*]. A recent study of basic circuits has shown that a significant improvement is to be obtained from a modification of a well-known trigger circuit — a development of considerable promise in the realization of large-scale integration.

The dominant trends in the development of logic circuits, which are so vital to applications like computers, are the efforts to achieve *speed* and *economy*. Clearly, any attempt to increase the speed of operation will imply shorter transit times and hence smaller dimensions. And the effort to reduce the costs of manufacture encourages the replacement of conventional circuits by integrated circuits, which require fewer assembly operations for each circuit. This tendency has become more noticeable in the last few years as control of manufacturing processes has improved and the reject rate for integrated circuits has declined. The integrated circuit, particularly the solid or "monolithic" form, at the same time represents the ultimate in miniaturization.

The next step in this development is "large-scale integration", in which a number of elementary circuits are accommodated on a single crystal-chip. Again, this is something which has become a reality because of the improved yield of the manufacturing processes. If the digital system to be built is properly divided up into modules, each to be made on a single chip, then the number of internal connections on one chip can be optimized with respect to the number of input and output leads. The chip then has the minimum number of soldered joints and thus an increased reliability.

Large-scale integration is today taken to mean something like a hundred or more basic logic circuits per chip. However, the great difficulty encountered when trying to achieve this is the heat which is generated. The basic logic circuits known and used at the present time have a dissipation of about 60 mW. If the chips are not going to be specially cooled, the maximum permissible dissipation per chip is $\frac{1}{4}$ W. This means that no more than four conventional basic circuits could be allowed per chip. This number might be increased by good cooling, but it would still fall short of what is needed.

Circuits therefore have to be found which are at least comparable in speed with the existing ones but generate less heat. This means that the circuit should contain only a few elements and the operating swing of the circuit, i.e. the voltage difference between the two logic levels, should be smaller than has been used up to now. When the swing is reduced, the supply voltage can be reduced proportionately, the heat due to dissipation diminishing as the square of the swing.

The circuits described in this article go a long way towards meeting these requirements. The starting point for the design is the Schmitt trigger, a circuit already widely used in other applications. The use of this circuit, with suitable component values, permits a drastic reduction of the swing. It will be shown here that the swing need be no higher than 200 mV, giving a dissipation of only 3 to 4 mW, yet permitting switching speeds of 2 to 5 ns to be reached.

This is a substantial improvement compared with existing circuits.

The basic circuit

The basic circuit, which can be extended to form various types of logic circuit, is the difference amplifier of *fig. 1*. As a further simplification the resistance R_1 with the supply voltage in series will be represented by a current source I. In this well-known circuit, for fixed

A. Slob is with Philips Research Laboratories, Eindhoven.

[*] See E. J. van Barneveld, Digital circuit blocks, and C. Slofstra, The use of digital circuit blocks in industrial equipment, Philips tech. Rev. 28, 44-56, 1967 (No. 2), and 29, 19-33, 1968 (No. 1).

V_r transistor *1* conducts when $V_0 \gg V_r$, while transistor *2* conducts when $V_0 \ll V_r$. In the first case the V_1 output is at a low level ($= -IR$) and the V_2 output is at a high level ($= 0$), while in the second case the situation is reversed. The voltage variation which thus appears at the output, of value IR, will be referred to as the "theoretical swing". Now it has been found that reliable operation with a *small swing* can be obtained if positive feedback is applied with the base voltage V_r of the second transistor made equal to V_1. (The resultant circuit corresponds in fact to a Schmitt trigger [1].) In this case, making a few simplifying assumptions, we can derive the expression given below for the relation between input and output voltages. This equation takes its simplest form if we calculate all voltages with respect to a level $-\frac{1}{2}IR$, that is to say half the swing. The output voltages V_1 and V_2 are then equal and opposite, i.e. $V_1 = -V_2$. It is also convenient to express all the voltages in terms of the quantity kT/q (kT/q has the dimension of a voltage; k is Boltzmann's constant, T the absolute temperature, q the charge of the electron), which is equal to about 25 mV at room temperature. Using the abbreviation z for $\frac{1}{2}IR$ (also expressed in terms of kT/q) we can then write:

$$\frac{V_0}{z} = \frac{1}{z} \ln\left(\frac{1 + V_2/z}{1 - V_2/z}\right) - \frac{V_2}{z}.$$

The relation between V_0 and V_2 is shown in *fig. 2* for various values of z (i.e. the half-swing). It can be seen that for $z > 2$ part of the curve has a negative slope. The circuit here is unstable. If the input voltage decreases, starting from Q (see *fig. 3*), then when point P_a is passed the output voltage suddenly changes from $V_2(P_a)$ to $V_2(P_b)$. An unstable region of the curve like this helps to give a good discrimination between "low" and "high" level at the output, and it is the presence of this unstable region which permits a very small swing to be used. We shall now work out just how large the swing should be.

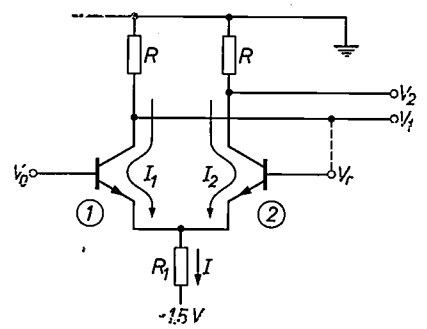


Fig. 1. Circuit of two transistors (difference amplifier, known as a "long-tailed pair") used as the basic circuit for the applications discussed here. The vital feature is the use of positive feedback, achieved by making V_r equal to V_1.

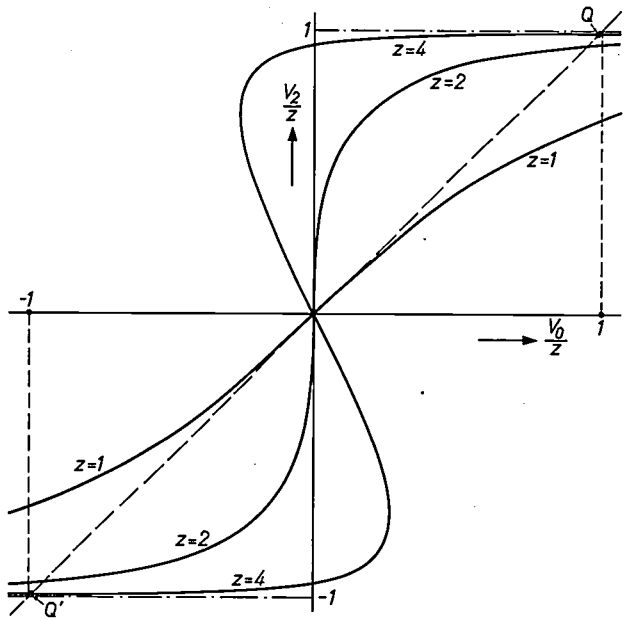


Fig. 2. Idealized characteristics of the basic circuit of fig. 1. Depending on the choice of the parameter z (half of the "theoretical swing") different types of curve are obtained.

Noise margin

The best choice of operating curve from the set shown in fig. 2 is determined by considering the available "margin" for noise or other unwanted signals at the input.

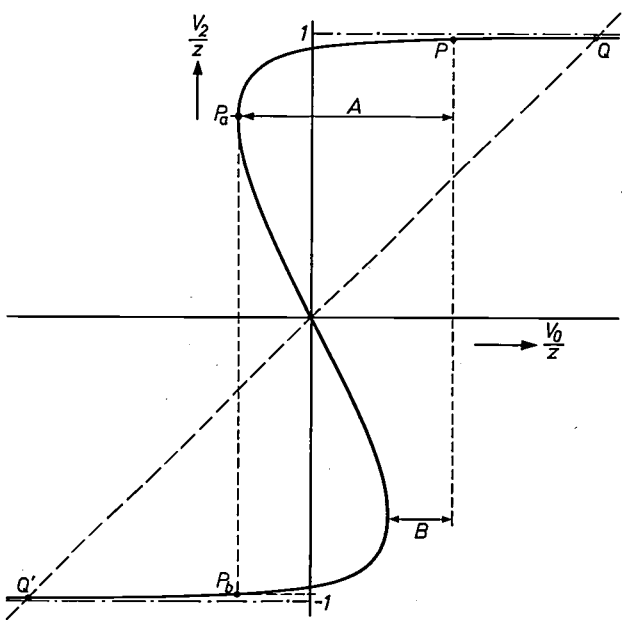


Fig. 3. There are two critical values, A and B, for the noise that can occur at an arbitrary operating point P.

Referring to the arbitrary operating point P in fig. 3, we can distinguish between two critical noise levels:

1) Noise leading to an undesired switchover from the point P; here A is the value that must not be exceeded.
2) Noise preventing point P from being reached from the low level; this occurs if the value of B is exceeded.

Let us now consider a long chain of identical circuits in which the same level has to be passed on from one to the next. Each circuit is assumed to have an identical noise source of voltage $-m$. If the input voltage for the first circuit is $V_0(Q_1)$ — see *fig. 4* — then the effect of

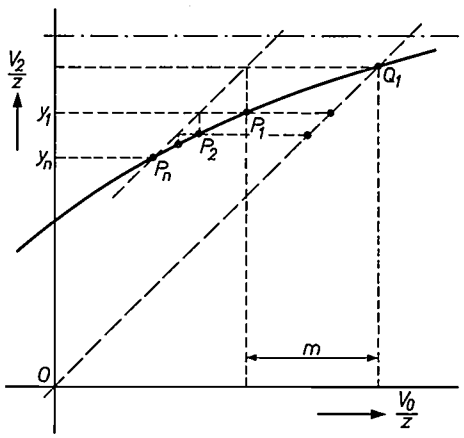


Fig. 4. Effect of identical noise voltages (of magnitude m) in a long chain of logic circuits in which the output voltage V_2 of one circuit acts as the input voltage of the next.

the noise voltage is that the output voltage is not $V_2(Q_1)$, but y_1, corresponding to an operating point P_1. Taking this value as the input voltage (using the line OQ_1, since V_2 is the input value V_0 for the next circuit) and again applying the same noise contribution $-m$, we find the operating point P_2 for the second circuit. It follows, therefore, that the final operating point P_n for the chain of circuits is found as the point where the curve being considered cuts the line parallel to OQ_1 and separated from it by a horizontal distance m. Although a long chain of identical circuits like this would not perhaps be found very often in practice, this kind of situation does occur when the output is fed back to the input, as when the circuit is used as a storage element, with the noise source m located in the feedback path.

It follows from this that the maximum noise signal that just fails to make the circuit switch over is found with the aid of a straight line which makes an angle of 45° to the axes and is tangent to the appropriate curve. (see *fig. 5*). This gives $m_0 = P_nQ_n$ as the value of the

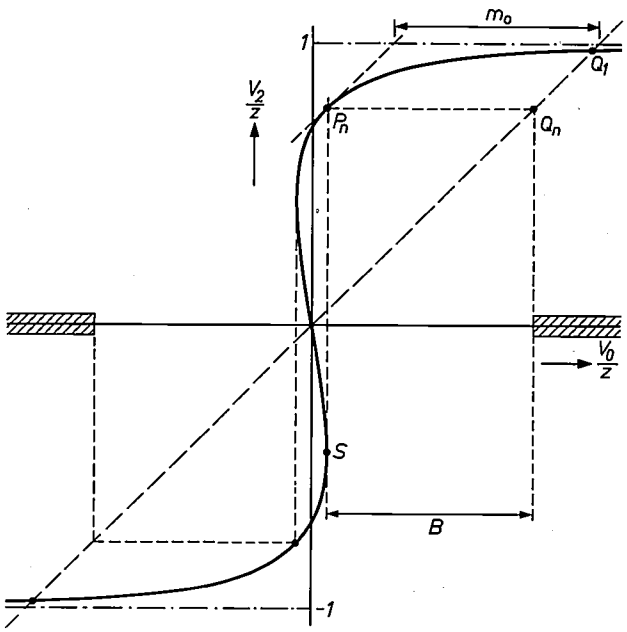


Fig. 5. Optimum shape of the characteristic of the basic circuit. The turning-point S must lie vertically beneath the point P_n where the slope is unity. The maximum noise margin is $m_0 = P_nQ_n$.

maximum noise, and the hatched area of the V_0 axis as the permissible region of input values. On the other hand, if Q_n is to be reached from the low level, the distance B in figs. 3 and 5 must not be smaller than P_nQ_n. We thus arrive at the requirement: the turning-point S on the curve and the point P_n where the 45° tangent meets it must lie on the same vertical line if the two kinds of noise margin are to be identical and at the same time as large as possible. We have thus established the value of the parameter z, the half-swing: it turns out to be $z = 2.47$. This then gives $V_0/z = 0.06, 0.985$ and 0.77 for the points S, Q_1 and Q_n, so that $P_nQ_n = 0.71$ (all expressed in terms of kT/q). From this we should therefore find the "theoretical swing" RI to be equal to $2 \times 2.47 \times kT/q$ or about 125 mV, with the noise allowed to reach about 35% of the theoretical swing.

Although certain simplifying assumptions were made in the above calculations, the results obtained give a fairly realistic idea of the situation encountered in practice. Thus, the optimum value of z as measured in a practical circuit is $z = 4$ (see the characteristics in *fig. 6* for the V_1 and V_2 outputs), i.e. a swing of about 200 mV. Here, at the point Q_n we have $V_0/z \approx 0.90$ for the "V_2" output and about 0.80 for the "V_1" output, while at the point S we have $V_0/z \approx 0.10$ and $m \approx 0.70$, so that the noise margin can again be about 35% of

[1] See G. Klein and J. J. Zaalberg van Zelst, Precision electronics, Philips Technical Library, Eindhoven 1967, page 363.

the total swing. A noise margin of this magnitude seems somewhat excessive, but it also has to serve for taking up tolerances in supply voltages, resistors and transistors. Moreover, the figures apply to an unloaded circuit, and loading also makes things less favourable.

Taking into account various design criteria which we shall not deal with here, the appropriate supply voltage is found to be 1.5 V.

The total *dissipation* of the circuit is now found after establishing the most suitable current value for the rise time, in this case 2 to 3 mA. The product of the current and the supply voltage is then about 3 to 4 mW. Switching times from 2 to 5 ns were measured at these values.

The lower parts of the curves in fig. 6 have a slightly different shape because the current source which was assumed in the calculations has in practice to be replaced by the common resistor R_1 (fig. 1). When transistor *1* starts to conduct, this resistor transmits some of the input voltage variation to the output V_1.

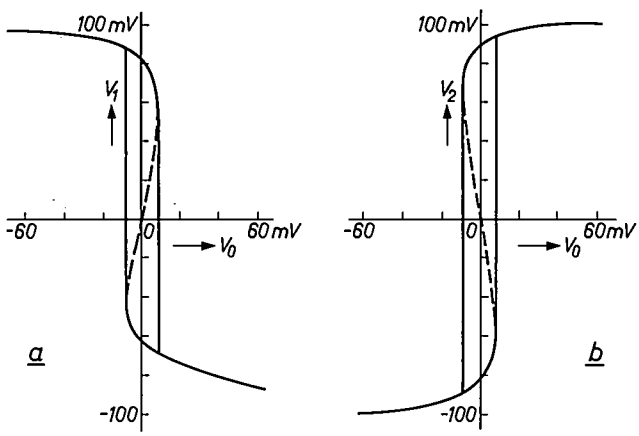


Fig. 6. Measured characteristics *a*) for V_1 (NOR output) and *b*) for V_2 (OR output).

Applications of the basic circuit

The basic circuit which has been described, and modifications of it, can be applied in various ways.

a) Multi-input circuit.

Fig. 7 gives an example of the same circuit with several inputs. The presence of a high voltage at one of the points *A*, *B* or *C* is sufficient to give V_1 a low value and V_2 a high value. The circuit is thus a NOR/OR circuit for high voltages, with V_1 as the NOR output and V_2 as the OR output.

b) Threshold logic.

In threshold logic the presence of a high voltage at the output depends on at least k out of n inputs being high, as suggested by the block symbol of *fig. 8*.

If $k = 1$ the circuit corresponds to an OR circuit; with $k = 2$ and $n = 3$ the circuit is useful for forming the carry digit in a binary addition, a carry occurring

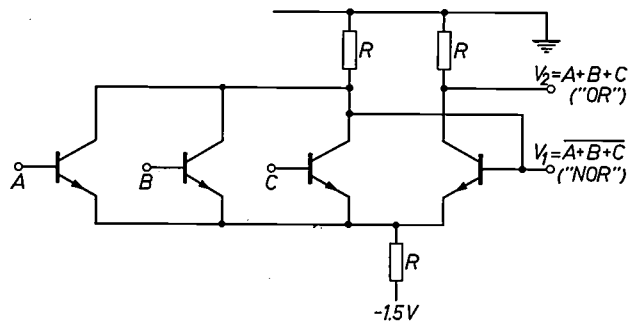


Fig. 7. The basic circuit used as NOR/OR circuit for high voltages (i.e. in "positive logic").

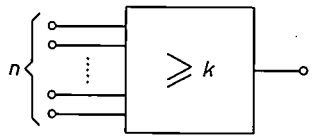


Fig. 8. Block symbol for a threshold-logic circuit. The circuit gives a "1" when k or more of the n inputs supply a "1".

where two or more "ones" occur in the digits *a* and *b* to be added and the incoming carry digit *c*. The circuit shown in *fig. 9* includes four of the basic circuits of the type shown in fig. 1. The first three have the 75 Ω resistor as common resistance, so that at point *C* about $\frac{1}{3}\times$ the sum of the three output voltages appears ($75/220 \approx \frac{1}{3}$), which serves as the input for the fourth circuit. The "threshold" action is given by the presence of this resistance, since currents of $2\times I$ or $3\times I$ flowing in it will keep point *C* below the switchover point, while a current $1\times I$ or $0\times I$ will bring *C* above it.

This circuit can readily be manufactured with present-day techniques.

c) "Clocked" storage element (*fig. 10*).

We connect the output V_2 of the basic circuit to the input V_0 (the transistors have been drawn the other way round to make the figure neater). An additional pair of transistors D_1, D_2 is provided to give a means of external control. Since the output voltage V_2 is now equal to the input voltage V_0, the operating point of the circuit goes to one of the points Q or Q' of the characteristic shown in fig. 3. Consequently the circuit has only two stable stages: V_2 high and V_1 low or *vice versa*. The configuration thus acts as a storage element. The input transistors D_1 and D_2 can be used to bring the circuit into one of the two stable states.

If one of the transistors D_1 or D_2 is made to conduct, current is drawn from the corresponding branch in the basic circuit. This works exactly like the "noise" source introduced in the discussion of the noise margin. It

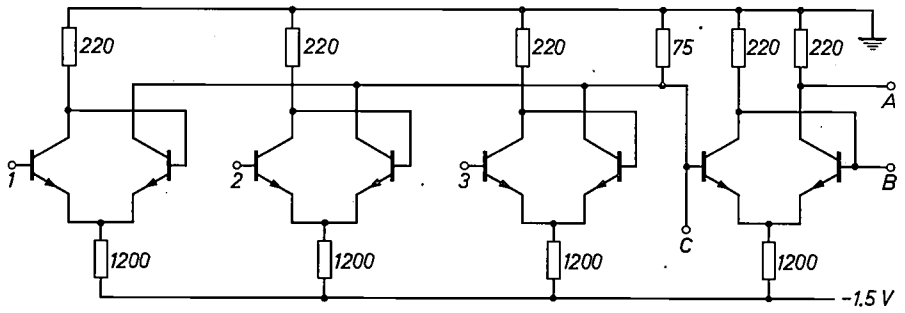


Fig. 9. The basic circuit used for threshold logic. The circuit gives a high voltage at A as soon as two or more of the inputs *1*, *2* and *3* have a high voltage. This can be used in forming a carry digit in binary addition.

follows, therefore, that the current I_D, which has to be supplied by an input transistor to cause the circuit to change over, is given by I_D = (noise margin)/R. The situation now is that the current I_D can flow only during the time a positive "clock pulse" is present at the base of transistor K. In fact the transistors D_1 and D_2 have a high emitter voltage when K is not conducting. Because of this the circuit is only sensitive to the inputs D_1 and D_2 when K is conducting, that is to say when the base voltage of K is high. We thus have a "clocked" storage element.

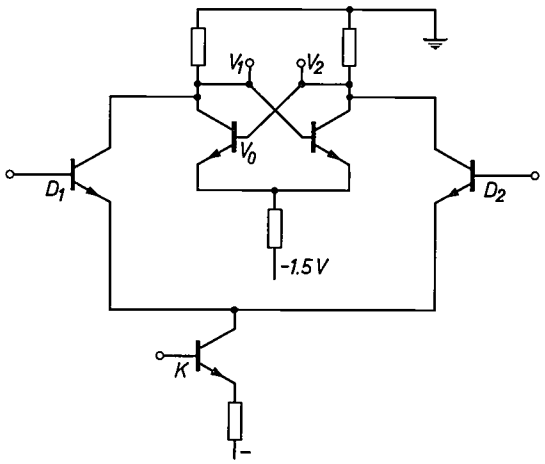


Fig. 10. The basic circuit as a "clocked" storage element. The circuit has two stable states. Which one it assumes can be controlled by making one of the input transistors D_1 or D_2 conduct. The configuration is such that the transistors D_1, D_2 can only operate when transistor K is conducting, during the period of a positive "clock pulse".

In this case the voltages for K are at a different level from those for D_1 and D_2, but this is of no consequence here. This storage element is exceptionally suitable for applications in shift registers and storage matrices of the fully-integrated type.

Plans based on the new developments outlined in this article are already under way for their application in various large-scale integrated devices.

Summary. A circuit which is closely related to the Schmitt trigger can be used as a basis for logic circuits, with the advantage that good discrimination is obtained using only a very small "swing". A result of this is that the amount of heat dissipated is quite small (3 or 4 mW or so), so that numerous logic functions can be accommodated on a single chip in an integrated circuit, without sacrificing switching speed: switching times may still be as low as from 2 to 5 ns. Since in addition the circuits contain relatively few components, this new circuit will be very useful in making large scale integration a practical reality.

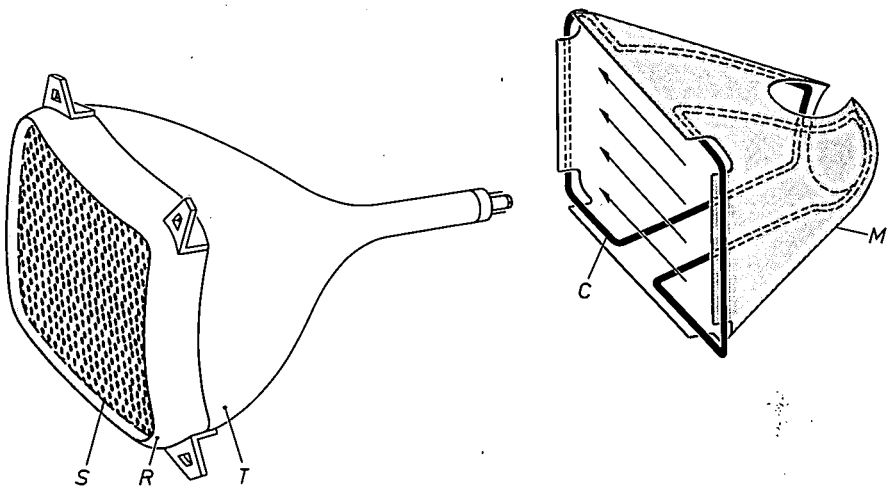
Degaussing a colour-television picture tube

In colour-television tubes of the type now in use (shadow-mask tubes) the picture is traced out on the screen by three electron beams. The red, green and blue fluorescent phosphor dots are arranged on the screen in a repetitive pattern designed so that the transmission by the shadow mask of the three beams, arriving from different angles, allows each beam to strike only the dots for one colour.

Even quite small deviations in beam angle can give colour errors in the picture. Such deviations can be caused by interfering magnetic fields. External fields, and in particular the Earth's magnetic field, are shut out

by means of the magnetic shield M around the tube and the reinforcement rim R around the screen. The shield and rim, like the shadow mask (S), are made of magnetically soft material. To give this material the highest possible effective permeability, and to eliminate any residual magnetization which would itself cause colour errors, the material must be completely degaussed.

When the tube has been made, the degaussing must be carried out before the picture quality can be checked.



This is done by means of a large field coil, which generates an alternating field near the tube and then after a short time is slowly moved away from the tube. While he is doing this the operator sees the striking display of colour patterns illustrated here, which are produced on the screen by the distorted electron beams. Once the field has been removed and degaussing is complete, the whole of the screen is the same colour all over.

The exposure time for this colour photograph presented no problem: since the frame frequency and the field frequency are the same (50 Hz), the colour pattern is stationary when the coil is held still.

This single degaussing operation does not complete the story. Every displacement of the receiver with respect to the Earth's magnetic field or exposure to transient interfering fields can induce a remanent magnetization, which must again be eliminated. This is very likely to happen once the receiver has been installed in the home, where interfering fields can be caused by all kinds of domestic appliances. This problem has been solved by using a degaussing coil (C in the figure) which is fitted around the tube of every receiver as shown in the diagram. This coil is energized by a high current, derived from the mains, every time the set is switched on. In the few seconds that elapse before the picture appears the current is reduced gradually to a level at which it will not affect the picture. This can be achieved simply and effectively by connecting a PTC resistor [1] in the supply lead of the degaussing coil: at first the PTC resistor has a very low resistance, but as it warms up the resistance becomes very high.

[1] PTC resistors are described in: E. Andrich and K. H. Härdtl, Philips tech. Rev. 26, 119, 1965.

The Philips electron microscope EM 300

C. J. Rakels, J. C. Tiemeijer and K. W. Witteveen

Two parallel lines of attack can be distinguished in the development of the Philips electron microscopes. Improvement of the resolving power goes hand in hand with improvement of the arrangement and design of the instrument, giving maximum efficiency of operation and ease of use. The latest Philips electron microscope, the EM 300, offers some noteworthy advances in both of these aspects.

Introduction

The range of electron microscopes put on the market by Philips in the last 20 years, the EM 100, the EM 75 and the EM 200 [1], has recently been extended by the addition of a new instrument, the EM 300. The main difference between this microscope and the others is its even better resolving power (less than 5 Å). The EM 300 is also simpler to operate than its predecessors, and in its equipment and design even more attention than before has been paid to the convenience of the user. A large number of matching accessories have been developed, so that the microscope can be used for a very wide variety of investigations without requiring major changes of the instrument. The picture can be displayed on a television monitor [2] as well as on the usual fluorescent screen. The magnification can be varied in 31 calibrated steps from 220 times to 500 000 times; the range of possible magnifications thus partly overlaps that of the optical microscope.

As can be seen from *fig. 1*, the microscope tube of the new instrument is vertical, as it is in the EM 200 and the EM 75. This arrangement has proved to be the best one for stability, ready accessibility of the operating controls, easy attachment of accessories and simple maintenance. The cabinet on the left-hand side contains the sensitive electronic circuits, such as the stabilizing circuit for the high voltage and the lens currents, and various auxiliary circuits. The vacuum system is accommodated in the central section, behind the space for the operator's feet. This central section, which carries the microscope tube, is very sturdily built to ensure mechanical stability, and its weight is taken directly by the floor, independently of the two side cabinets.

In the projection chamber, fitted with three viewing windows, the fluorescent screen is located at the height of the table-top. This screen can be swung aside to give the electrons access to a plate camera or to a television camera tube ("Plumbicon" [*] tube). When the camera tube is used, an image intensifier can be interposed. Facilities are also available for taking photographs with two types of roll film camera in addition to the plate camera.

The electronic circuits are all transistor circuits. Some of them are water-cooled. All circuits are contained in standard exchangeable modules. The operating controls not regularly needed are on sliding panels immediately below the table-top (*fig. 2*). There are no controls on the table-top itself.

One of the distinguishing features of the vacuum system is that it works automatically, and another is that the roughing pump can be switched off for hours once a sufficiently high vacuum has been reached. This eliminates a major source of vibration and noise. The vacuum is better and is reached more quickly than in the predecessors of the EM 300. A number of vacuum locks are provided, so that the filament, photographic plates or films and the specimen can be changed without having to admit air into the entire microscope tube.

Photographic exposures are also very easy to make. The film and plate transport is motor-driven, and the EM 300 is fitted with an exposure meter which is arranged so that photographs may be taken in much the same way as with a semi-automatic camera: for correct

Ir. C. J. Rakels, J. C. Tiemeijer and Ir. K. W. Witteveen are with the Philips Industrial Equipment Division (PIT), Eindhoven.

[*] Registered Trade Mark for television camera tubes.

[1] A description of the EM 100 electron microscope is given in: A. C. van Dorsten, H. Nieuwdorp and A. Verhoeff, Philips tech. Rev. **12**, 33, 1950/51, and of the EM 75 in A. C. van Dorsten and J. B. Le Poole, Philips tech. Rev. **17**, 47, 1955/56.

[2] See P. H. Broerse, A. C. van Dorsten and H. F. Premsela, Philips tech. Rev. **29**, 294, 1968 (No. 10).

Fig. 1. The Philips EM 300 electron microscope.

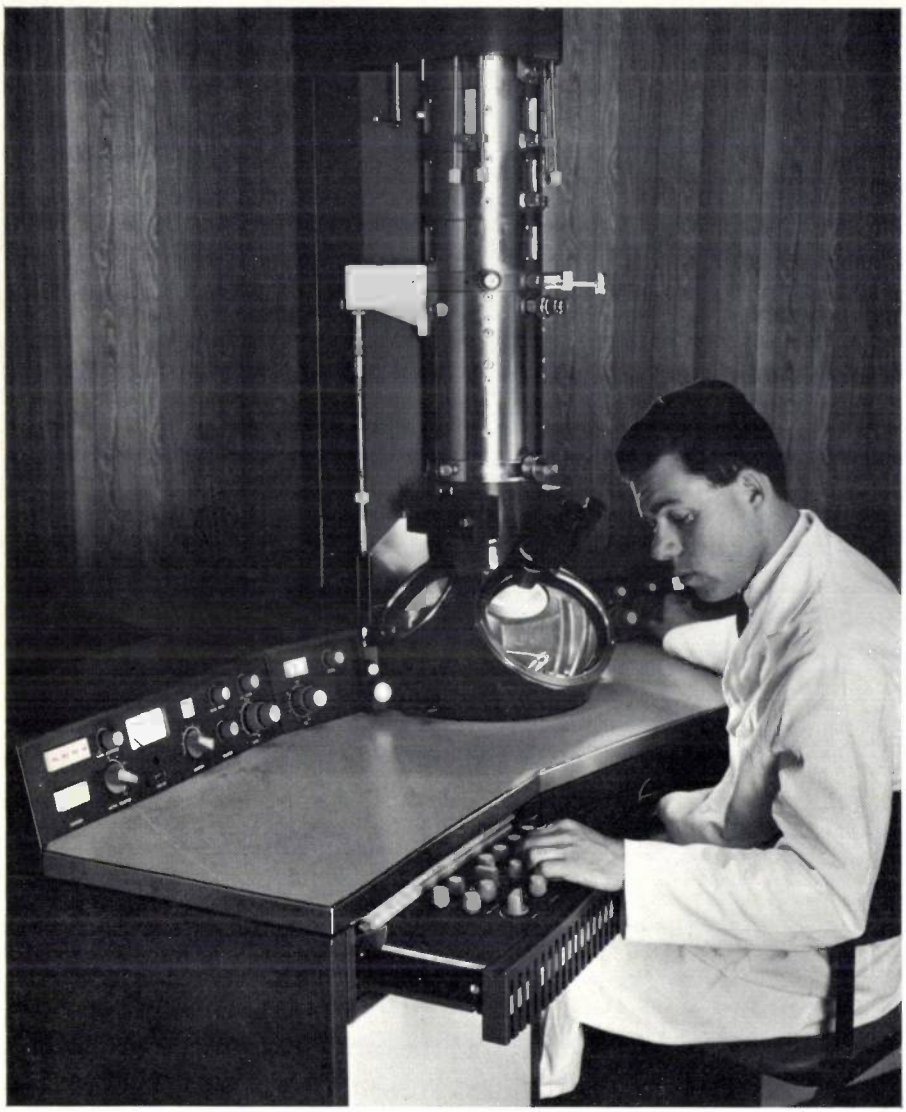


Fig. 2. The operating controls that are most frequently needed are within easy reach above the table-top; the others are on sliding panels immediately below it.

exposure the operator adjusts the shutter speed or current density on the screen until the pointer of a meter on the control panel is at the centre of the scale. To make an ordinary exposure and a diffraction exposure of the same object one after the other, it is simply necessary to turn the knob of the "function switch".

An important feature of the new microscope is that all parts of the electron-optical system can be separately aligned. This minimizes displacement of the image in the image plane, due say to variations in the strength of the lenses.

The very high resolving power of the new microscope is illustrated by *fig. 3*. It is so good partly because of the mechanical stability noted earlier and partly because of the excellent stabilization of lens currents and high voltage, the low sensitivity to external fields and, in particular, the extremely short focal length of the objective lens (1.6 mm). The importance of making the focal length short is explained in the next section, where the image formation and the electron-optical factors that determine the resolving power are dealt with. The other sections will be concerned with technical features of the microscope and accessories.

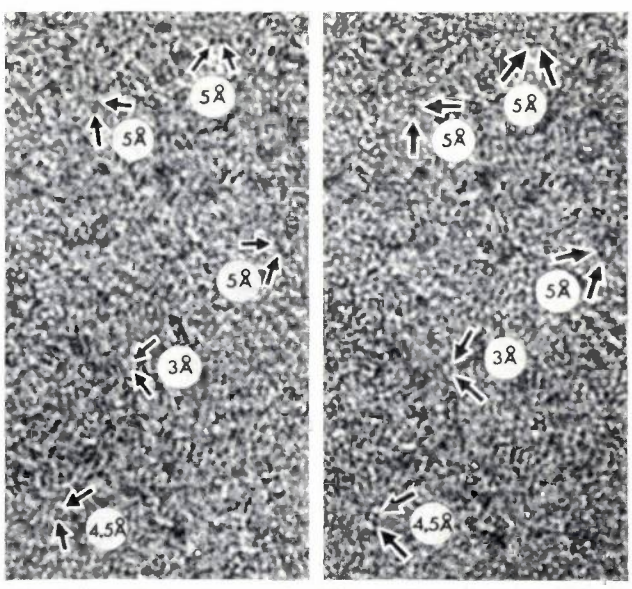


Fig. 3. Two photomicrographs of the same specimen in which details situated 5 Å apart can be distinguished. The fact that the same features are seen on both exposures proves that they are not grains from the photographic emulsion.

Electron optics and resolving power

Fig. 4a shows schematically the principal electron-optical components inside the microscope tube. Apart from various diaphragms, these components (from

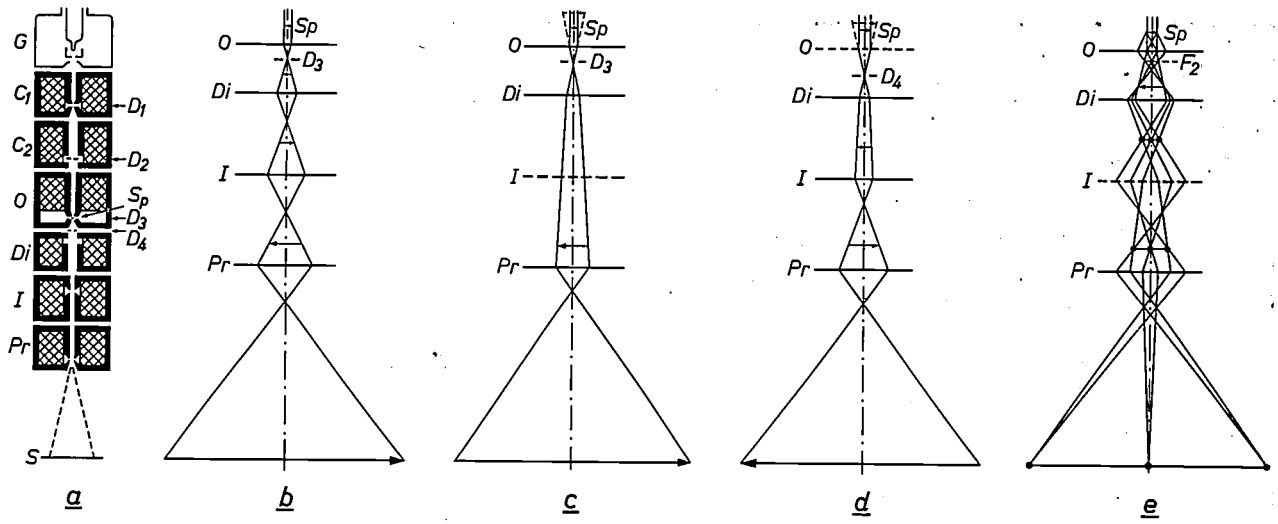


Fig. 4. *a*) Schematic arrangement of the microscope tube. *G* emission chamber with electron gun. C_1 first condenser lens. C_2 second condenser lens. *O* objective lens. *Sp* specimen. *Di* diffraction lens. *I* intermediate lens. *Pr* projector lens. *S* fluorescent screen. D_1 and D_2 condenser diaphragms. D_3 objective diaphragm. D_4 selector diaphragm.
b) Path of rays in normal use (high magnifications). The intermediate lens and the diffraction lens are both in operation.
c) Path of rays for medium magnifications. The objective lens forms a virtual image of the object. The intermediate lens is switched off.
d) The diffraction lens used as objective (very small magnifications, for initial investigation of a specimen). In the case illustrated the objective lens is weakly excited.
e) Electron diffraction. At the object plane of *Pr* the diffraction lens now no longer forms an image of the object, but of the focal plane F_2 of the objective. If necessary the intermediate lens can be switched off.

top to bottom) are the electron gun, two condenser lenses, the objective lens, the diffraction lens, the intermediate lens and the projector lens. Altogether, therefore, there are six lenses, all of them electromagnetic. The specimen is located approximately at the upper focal plane of the objective lens. There is a choice of accelerating voltages of 20, 40, 60, 80 or 100 kV.

The gun and the two condenser lenses constitute the "illumination system". The first condenser lens C_1 is a powerful lens whose excitation can be varied. The minimum focal length is 1.5 mm. This lens forms a reduced image of the electron source, with a diameter of about 0.3 μm at maximum excitation (for most applications the maximum excitation will not be used, but one giving a spot diameter between 5 and 1 μm). The second condenser lens C_2 is a relatively weak one, whose excitation can also be varied. This lens is adjusted in such a way that the image given by C_1 appears with unit magnification at or near the specimen — usually slightly above the specimen since this setting gives a small illuminating aperture, which is desirable in certain cases (see below).

The size of the spot projected on the specimen, and its intensity (current density) can be varied over a wide range by varying the strength of the two condenser lenses (*fig. 5*). This has considerable advantages which are not present in an illumination system using only one condenser lens [3]. In the first place, the electron

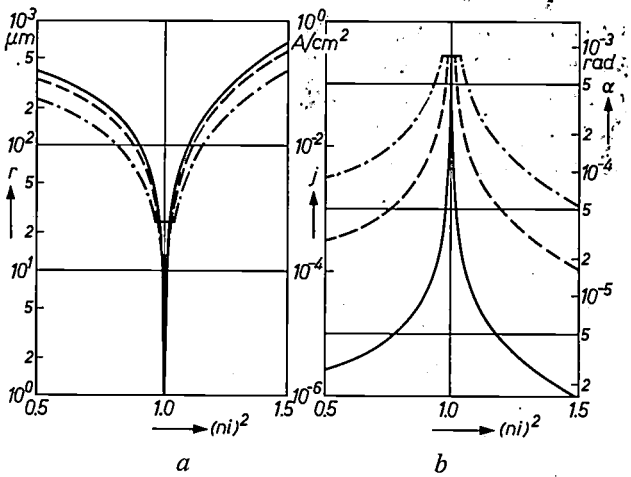


Fig. 5. *a*) Variation of the radius *r* of the electron spot on the specimen as a function of the excitation of C_2 with constant excitation of C_1 (*ni* is the number of ampere-turns in arbitrary units) and constant diameter of the condenser diaphragms.
b) Similar curves for the aperture α and the current density *j* in the specimen plane; these curves coincide if the ordinate scale is properly chosen.
In (*a*) and (*b*), the solid-line curves apply for C_1 strongly energized, the dashed curves for weak energization. The chain-dotted lines are found when C_1 is switched off.
When the strength of C_1 is varied, while C_2 keeps the spot in focus on the specimen, the size of the illuminated part of the specimen changes, but not the aperture or the current density. These can be reduced only by defocusing C_2.

[3] For a discussion of the advantages of a condenser with two lenses, see S. Leisegang, Proc. 3rd Int. Conf. on Electron Microscopy, London 1954.

spot can be made almost as small as the part of the specimen to be investigated, without illuminating it more strongly than is necessary. This permits the thermal loading of the specimen to be kept to a minimum. Secondly, when the light spot is small, only a few electrons are scattered in parts of the specimen that are outside the field of view, and this improves the quality of the image. Thirdly, when the current passing through the specimen is small, there is less trouble from impurities liberated from parts of the microscope or from charging effects.

The specimen can be illuminated in the normal way, or with *oblique* illumination (dark-field illumination). The maximum value that can be chosen for the angle between the axis of the beam and the optical axis of the microscope is as much as 5°.

The optical characteristics of the next component, the *objective lens*, which are of great importance to the quality of the image formation and the resolving power will be dealt with in more detail later. It is sufficient to note here that this lens can be varied in strength, that the coil is situated *above* the specimen, and that the specimen is located *inside* the magnetic field of this lens (immersion objective).

The *diffraction lens*, which is relatively weak and has a wide bore, and the *intermediate lens* can also both be varied in strength. If these two lenses are used — either together or separately — as an intermediate lens, the whole range of magnifications can be covered without having to alter the strength of the projector lens [4]. For high and very high magnifications (greater than 20 000 times) the combination shown in fig. 4*b* is used, for moderate magnifications (4000 to 20 000 times) the combination of fig. 4*c*, and for small magnifications that of fig. 4*d*. In the last case the diffraction lens in fact functions as a (non-immersion) objective lens; the diffraction lens could hardly be used in this way if the specimen were *above* the coil of the objective lens, because the distance from the specimen to the lens would then be too great.

When both the diffraction and the intermediate lenses are used, as in fig. 4*b*, the magnification is varied by varying the strength of the intermediate lens. (The strength of the objective lens is also varied slightly in order to keep the image in focus.) The diffraction lens is always set to maximum strength. With the arrangement of fig. 4*c*, the intermediate lens is switched off and the magnification is varied by means of the objective and the diffraction lenses. As can be seen, the objective here forms a virtual image of the object. This image, which moves upwards with increasing magnification, is projected by the diffraction lens on to the object plane of the projection lens. With the mode of operation shown in fig. 4*d*, the objective lens is usually weakly energized so as to cause the electron paths to cut the optical axis at the location of the selector diaphragm D_4. This can then act as an objective diaphragm, so that good contrast is still obtained when the instrument is used in this way. If it is desired not to subject the specimen to a magnetic field, or to subject it only to a field whose strength is accurately known [5], the objective lens is not excited. As a result of the arrangements and modes of operation illustrated in fig. 4*a-d*, the distortion of the magnified image obtained with the EM 300 is low even at the weakest magnifications. Finally, fig. 4*e* shows the path of the rays when the microscope is used as an instrument for making electron diffraction photographs.

Resolving power

The resolving power of the EM 300 is primarily determined by the objective lens. The less fundamental factors affecting the resolving power, such as the mechanical and electrical stability, are of lesser importance in the new microscope. We shall therefore briefly review the way in which the resolving power is affected by the principal lens errors: spherical aberration, chromatic aberration and astigmatism.

Chromatic aberration cannot be neglected because an electron beam in an electron microscope can never be completely monoenergetic: there is energy dispersion immediately outside the gun since the electrons are liberated by thermionic emission, velocities are affected by space charge in regions where the paths cross one another (Boersch effect) and finally not all the electrons are subject to the same amount of energy loss in the specimen. These effects make it necessary to take account of chromatic aberration even in an electron microscope that is electrically perfectly stabilized [6].

Every magnetic lens with circular symmetry suffers from positive spherical aberration, that is to say the focal length is smaller for the peripheral rays than for the paraxial rays. Unfortunately, no practical method has yet been found for correcting this lens error. The spherical aberration can, however, be minimized by choosing appropriate values for the spacing and bore of the pole-pieces, and in particular for the excitation of the lens. This minimum of spherical aberration decreases as the focal length of the lens is made smaller. This is the primary reason for making the focal length of the objective lens as small as possible in order to obtain a good resolving power.

The chromatic aberration of an electron lens is also less when the focal length is smaller. The chromatic aberration constant C_c of the EM 300 (see below) is therefore only 1.3 mm. This means that the spread of the energy of the electrons will prevent the theoretical resolving power from being achieved only when the electrons pass through a relatively thick specimen. Provided the specimen is sufficiently thin (e.g. a carbon film

of 20 nm), only a small fraction of the electrons looses energy in the specimen.

If spherical aberration only were present, and if the situation in an electron microscope were entirely analogous to that in an optical microscope — in particular with respect to the illumination of the object — the theoretical resolving power found from the classical treatment due to Rayleigh would be 2.3 Å (0.23 nm) at an accelerating voltage of 100 kV.

In ordinary optics the classical treatment of resolving power has been supplemented in the last 20 years by the theory of contrast transmission, which uses contrast-transfer functions and curves. For some years now this type of treatment has also been applied to electron lenses [7]. It will appear that the value of resolving power quoted above can be approached very closely, but also that the interpretation of the image of an object whose details are of the order of magnitude of ångströms is a complicated matter.

We shall now consider how the transfer of contrast in the object is affected by spherical aberration, defocusing of the microscope and, through chromatic aberration, by fluctuations in the high voltage and the lens currents.

Let us consider a beam of rays emitted from the first focal point. If there were no defects in the lens, the spherical wave-front of the beam would emerge from the lens as a planar wave-front. The lens defects do however cause some deviation, known as the wave-front aberration Δ, which depends on the radius vector r_b in the plane under consideration and on the position of that plane. (For object points lying in the first focal plane but away from the optical axis, different values of Δ are found at a given r_b, but this complication may be neglected in an electron microscope since the objects are relatively small.) Taking only the spherical aberration into account, then the wave-front aberration in the second focal plane is given by $\Delta(r_b) = -Ar_b^4$. If the object point is on the axis but not at the first focal point, an extra term Br_b^2 enters into the equation, where B depends on the distance Δz_F of both points, the "defocusing". The total wave-front aberration is then:

$$\Delta(r_b) = -Ar_b^4 + Br_b^2. \quad . \quad . \quad . \quad . \quad (1)$$

Fig. 6 gives a curve of this function for the objective lens of the EM 300, for the case $B = 0$ and also for four cases where there is a certain amount of defocusing. The wave-front aberration Δ is expressed in terms of the wavelength of electrons at 100 keV (0.037 Å). A second horizontal scale beneath the graph gives r_b in terms of the line spacing (grating constant) d of a sine grating that diffracts an electron beam of 100 keV through an angle α such that the diffracted beam is focused into

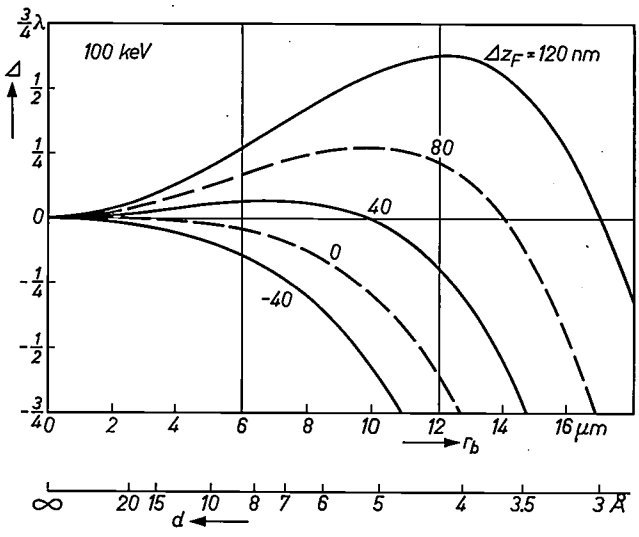


Fig. 6. The wave-front aberration Δ in the second focal plane of the objective, caused by spherical aberration and defocusing, as a function of the radius vector r_b, for an electron beam of 100 keV energy, emitted from the primary focus F_1 or from a point a distance of Δz_F away from it on the optical axis. The scale beneath the figure gives the grating constant d corresponding to r_b.

the second focal plane at a distance r_b from the optical axis.

From the curves of fig. 6 we can immediately find the effect of the spherical aberration and the defocusing on the diffracted image obtained from a given grating in the second focal plane, and thus at the same time obtain some idea of the imperfections that the magnified image on the screen will have. This comes about since an object can be regarded as a superposition of a variety of phase and amplitude gratings with different constants. (An object is a phase object when the emergent wave has the same amplitude and thus the same intensity at all points, but differs locally in phase.) For example, at zero defocusing the beam diffracted by an amplitude grating with $d = 5.3$ Å has a wave-front aberration of $-\frac{1}{4}\lambda$. This means that the image of this grating looks like that of a phase grating, i.e. the grating structure is completely imperceptible on the screen. On the other hand, if the object is a phase grating, its image will have maximum contrast. Because of the spherical aberration the structure of the phase grating, which in the ideal case would be imperceptible, can now be observed.

If we determine the contrast with which amplitude and phase gratings with a different grating constant are reproduced, and we plot the result in a graph, we obtain what are known as contrast-transfer curves. The

[4] The advantage of using a third magnifying lens, in addition to the objective and the projector lens, was first pointed out by J. B. Le Poole in Philips tech. Rev. **9**, 33, 1947/48.

[5] See for example H. B. Haanstra and H. Zijlstra, Philips tech. Rev. 29, 218, 1968 (No. 7).

[6] See the second article quoted in reference [1].

[7] See for example K.-J. Hanszen, Naturwiss. **54**, 125, 1967 (No. 6).

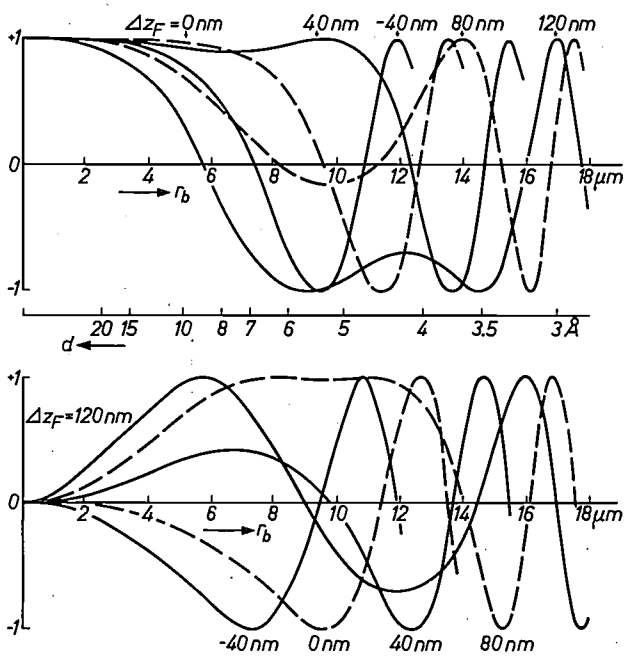


Fig. 7. Contrast transfer curves for the objective lens of the EM 300, derived from the curves of fig. 6; upper curves for amplitude gratings and lower curves for phase gratings.

curves for the EM 300 are shown in *fig. 7*. It can be seen from this figure even more clearly than from the previous one that there is no setting at which the finest grating structures ($d < 10$ Å) can all be seen at the same time. For instance, with a defocusing of 120 nm amplitude gratings with $d = 3$ Å and $3.5 < d < 6.5$ Å are reproduced with good contrast, but those where d is in the region of 3.25 and 9 Å are not. The best setting in this respect is the one with a defocusing of 40 nm. The corresponding defocusing for phase structures is 80 nm. We should add that there are many specimens (particularly biological ones) that yield only very weak amplitude contrasts. In the investigation of such specimens the presence of phase contrasts can be a help instead of an interfering effect.

From the foregoing we can find the stability requirements which the accelerating voltage V_0 and the excitation current I of the objective lens should meet in order to minimize the effect of fluctuations in I and V_0 on the resolving power. The displacement δz of the focal point caused by variations δV_0 and δI is given by:

$$\delta z_F \doteq C_c \{(\delta V_0/V_0) - 2(\delta I/I)\}. \quad . . \quad (2)$$

The quantity C_c is the chromatic aberration constant mentioned previously. We can now calculate that the contrast decreases to 90% if, in the period of observation (one minute at the most), the wave-front aberration Δ varies between $+\frac{1}{4}\lambda$ and $-\frac{1}{4}\lambda$, and all values are equally likely. Taking this as the criterion, we find from the wave-front aberration curves what variation in defocusing is permissible, and then from equation (2) what relative variation of V_0 and I this corresponds to at a given value of d (*Table I*). For the EM 300 the relative drift of V_0 is less than 5×10^{-6} per minute, and that of I less than 2.5×10^{-6} per minute. It is evident from this that the drift of I and V_0 in the EM 300 does not prevent the theoretical resolving power from being reached.

Table I. Maximum deviation in the accelerating voltage V_0 and the lens current I at which a phase object with a periodic structure (phase grating) of line spacing d, still gives an image with good contrast.

d (Å)	$\delta V_0/V_0 - 2\,\delta I/I$
10	110×10^{-6}
5	26×10^{-6}
4	17×10^{-6}
3	9.7×10^{-6}
2.3	5.7×10^{-6}

If the same is to apply for the effect of the astigmatism of the objective lens, then there should be virtually no change in the contrast-transfer curves when the plane in which the diffracted beam lies changes position. If this plane is rotated one full turn around the optical axis, the defocusing range must not be greater than about 10 nm. This means that the requirements for the circular symmetry of the pole-pieces are very exacting. These requirements cannot be met using normal machining methods and tools. The pole-pieces of the EM 300 are therefore made on a specially designed ultra-high precision machine equipped with hydrostatic bearings [8]. Before the pole-pieces are mounted in the microscope they are checked with a precision instrument, one of whose special features is that its own mechanical deviations do not affect the measurement [9].

Even with these special devices and provisions the astigmatism still cannot be reduced to a sufficiently small value. The residual astigmatism is compensated by means of a *stigmator*, which in principle is simply a combination of two cylindrical magnetic lenses whose strength can be varied. In the EM 300 the stigmator is a set of eight small coils mounted directly under the objective lens.

Apart from the above-mentioned characteristics of the objective lens and the more trivial factors referred to in the introduction, like the mechanical stability and sensitivity to stray magnetic fields, other factors affecting the resolving power are the quality of the specimen stage and the brightness of the electron source.

[8] See J. G. C. de Gast and H. J. J. Kraakman, shortly to be published in Philips tech. Rev.

[9] This instrument was built by the Central National Organization for Applied Scientific Research in the Netherlands (TNO), Technical University of Delft, and was described by J. Kramer, J. B. Le Poole and A. B. Bok in TNO-Nieuws **20**, 297, 1965 (in Dutch, with English summary).

The specimen stage must be capable of holding the specimen in position to within a few ångströms during the period of an exposure. The specimen stage in the EM 300 is the same as the one used in the EM 200, which meets these requirements.

The brightness of the electron source comes into the picture in the following way. We have just seen that in order to make the very finest details visible, it is often necessary to make use of phase contrasts. This calls for highly coherent illumination, and this condition is most easily met with a small illuminating aperture. (The condenser system of the EM 300 does in fact permit operation with a small illuminating aperture, as we noted earlier.) In order to obtain a sufficient current density at the fluorescent screen when using a small illuminating aperture and the maximum magnification of the instrument, it is necessary to have an extremely bright electron source.

The microscope tube

In this section we shall consider the principal design features of the components contained in the microscope tube. A simplified cross-sectional drawing of the tube is shown in *fig. 8*. In addition to the components in fig. 4, the drawing shows

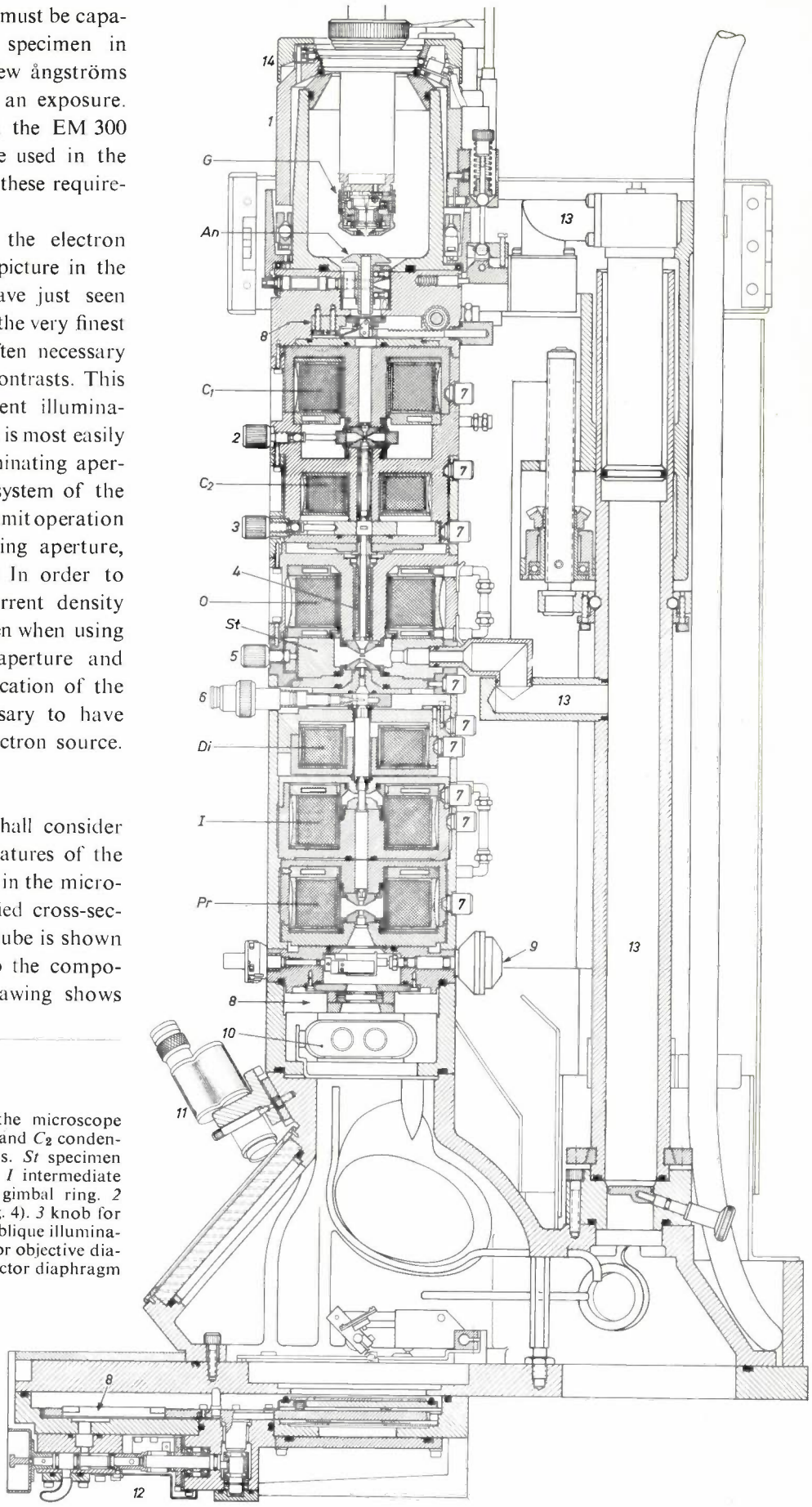


Fig. 8. Cross-section of the microscope tube. *G* gun. *An* anode. C_1 and C_2 condenser lenses. *O* objective lens. *St* specimen stage. *Di* diffraction lens. *I* intermediate lens. *Pr* projector lens. *1* gimbal ring. *2* knob for diaphragm D_1 (fig. 4). *3* knob for diaphragm D_2. *4* coils for oblique illumination and focusing. *5* knob for objective diaphragm D_3. *6* knob for selector diaphragm D_4. *7* alignment controls for lenses and pole pieces. *8* vacuum valves. *9* shutter. *10* camera (35 mm). *11* binocular microscope. *12* plate camera. *13* vacuum line. *14* rapid coupling device. The camera *10* can be replaced by an electron-diffraction device (i.e. a specimen holder for large specimens).

the specimen stage, the stigmator, the coils for oblique illumination, the photographic shutter and a number of vacuum valves and locks, as well as a number of controls for operating the various diaphragms and for the alignment of lenses, pole-pieces, etc.

Electron-optical section

The electron gun is of the conventional type, comprising a filament, a Wehnelt cylinder and an anode. It can be aligned by displacing the cathode (here taken to be the combination of filament and Wehnelt cylinder) with respect to the anode and rotating it about two axes perpendicular to the centre-line of the microscope tube and passing through the emitting tip of the filament. The mounting of the cathode in a gimbal bearing ensures that it moves smoothly and can be accurately aligned. The height of the anode can be adjusted while the gun is in operation. This is important in order to achieve maximum brightness from the electron gun at voltages lower than 100 kV. Facilities such as the rapid coupling arrangement and the vacuum lock situated immediately below the electron gun enable a filament to be replaced in less than a minute.

Each of the condenser lenses is fitted with a diaphragm holder containing three interchangeable diaphragms. Fitted against the lower pole-piece of C_2 is a disc which contains the stigmator. This compensates for departures from circularity of the illuminating beam caused by astigmatism in the condenser system. (If the beam is not of circular cross-section, the specimen may receive more thermal energy from the beam than necessary, and there may be unwanted charging effects at the edges of diaphragms, etc.)

In addition to the above-mentioned facilities for displacing and tilting the gun, the illumination system has a variety of other correction and alignment facilities which cannot all be mentioned here.

The design of the objective lens is illustrated in *fig. 9*. This lens, together with the specimen stage (shown shaded), forms a single assembly, which however can easily be removed in order to change from one type of specimen stage (see below) to another.

The upper pole-piece of the objective lens cannot be displaced and constitutes one of the fixed points used for alignment (the other is the centre of the fluorescent screen). The lower pole-piece can be aligned. When the current in the coil is at its maximum value (5 A), the magnetic field between the pole-pieces is about 1.5 Wb/m^2. Situated immediately under the lower pole-piece is the stigmator of the objective lens.

Fitted in the bore of the magnet coil, above the lens field, are the deflection coils for oblique illumination. These operate on the same principle as Le Poole's focusing device [10], and they can also be used for this purpose. Here, however, there are two pairs of coils instead of one. These pairs are arranged at right angles to one another, so that the azimuth of the plane in which the deflection takes place can be chosen as required.

The deflection coils of the focusing device are wound in such a way that the magnetic field they produce is highly uniform. There are also a few correction coils to compensate for any difference in orientation that may exist between the upper and the lower coils. With these facilities the angle between the beam and the optical axis can be increased to the high value of 5° mentioned earlier, with no noticeable change in the location at which the beam meets the specimen. When a set of these coils is used for focusing, very sharp images can be obtained even at quite high magnifications.

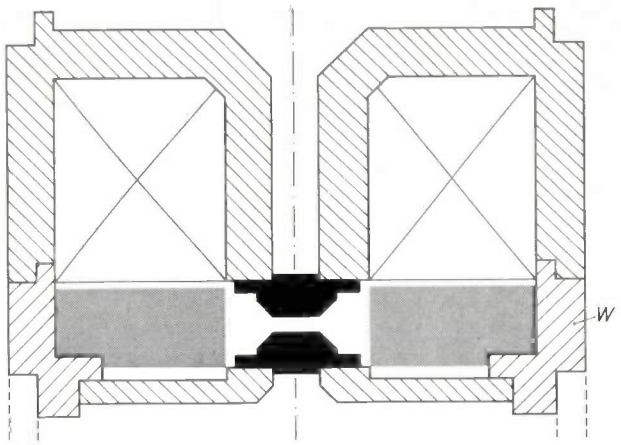


Fig. 9. Schematic cross-section of the objective lens and specimen stage (shaded). The pole-pieces are shown black. When the upper part (including the coil) is detached from the ring *W*, the specimen stage can easily be removed. The lower pole-piece can be aligned.

Fig. 10a shows how the specimen stage is mounted in the microscope tube. Fig. 10*b* shows the central section, which contains a cross-slide system. This system is driven by a system of pushrods and two levers, each driven in turn by micrometer screws, operated from the control panel. Exceptionally smooth and accurate movement of the specimen is obtained by proper balancing of frictional forces, spring forces and masses and by the appropriate choice of material. As in the previous Philips microscopes, the specimen is introduced through a vacuum lock by means of an injector.

The diffraction lens, the intermediate lens and the projector lens are relatively simple in design. The diffraction and intermediate lenses can be fully aligned, i.e. the lens can be displaced as a whole and in addition one of the two pole-pieces can be displaced with respect to the other. The projector lens can only be dis-

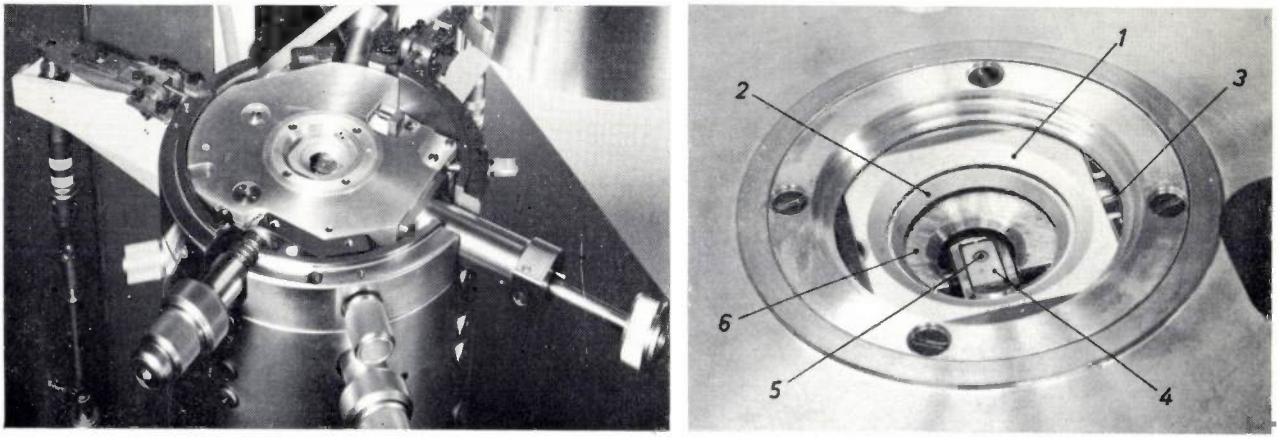
placed as a whole, which is sufficient because this lens is not varied in strength and no further magnification takes place after it.

Projection chamber and aids for viewing and recording

The aperture of the beam that converges at one point of the image formed in the projection chamber of an electron microscope is very much smaller than the aperture of the projector lens itself, and therefore the depth of focus is very great. This gives the designer considerable freedom in the positioning of the various viewing aids. A 35 mm camera is accommodated directly under the projector lens, and a plate camera just below the screen. Below this there is room for a camera for 70 mm film or for a television camera tube, which may if necessary be used in conjunction with an image intensifier. When the image intensifier is used, exposures can be made at a very low current density, permitting the investigation of specimens that would be damaged at the normal current density. In addition to the ordinary fluorescent screen there is a small screen with a very fine grain; the image formed on this can be viewed with a binocular microscope at a magnification of 10 times.

The arrangement of the plate camera is illustrated schematically in *fig. 11*. The main component is a rotating disc situated eccentrically with respect to the axis of the microscope tube. At every revolution this disc picks up a plate from a magazine, carries it into position under the microscope tube and, after the exposure is made, lets it drop into a box. The disc also carries the vacuum valve which shuts off the projection chamber from underneath. This is done by turning the disc until the vacuum valve is immediately under the tube opening and then raising it until it closes the opening.

Between the projector lens and the 35 mm camera there is a rotating shutter, which can be used for all photographic exposures. A rotating shutter has constructional advantages in view of the requirement that opening and closing must cause the least possible vibration.



a *b*

Fig. 10. *a*) The standard specimen stage in the microscope. The upper section of the tube shown here is the ring *W* (fig. 9) of the objective lens. *b*) A detailed view of the central section. *1* upper cross-table slide. *2* lower slide. *3* push-rod for the movement of the upper slide. *4* specimen holder. *5* specimen. *6* cooling element (—180 °C); this accessory does not belong to the stage but is introduced separately from the side.

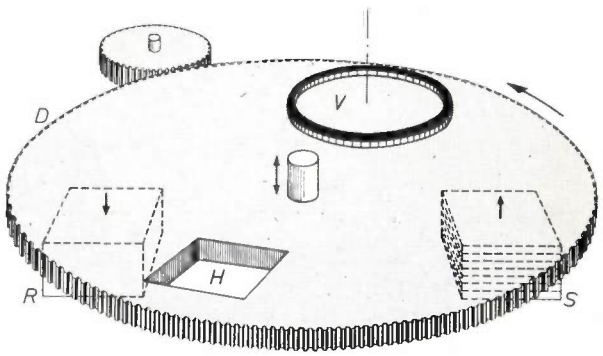


Fig. 11. Schematic representation of plate camera. The toothed disc *D* meshing in a smaller gear is motor-driven. *S* plate magazine. *R* box for receiving exposed plates. *V* vacuum valve. When the opening *H* arrives above *S*, a plate is pushed in. This travels around with the disc into position under the microscope tube, where it is exposed and then dropped into *R*.

[10] See the first article quoted in reference [1].

Vacuum system and electrical equipment

The vacuum system of the EM 300 (*fig. 12*) is basically conventional — comprising a roughing pump, vacuum reservoir, mercury ejector pump and oil diffusion pump — but differs in some significant details from the system used in the earlier Philips microscopes.

In the first place facilities are provided which make the entire pumping process completely automatic. Besides making the instrument very simple to operate, automation has the advantage that the pumping process is optimized and human errors in operation are excluded, thus reducing the risk of breakdown. In the first phase only the roughing pump is in operation and all the valves, apart from V_1, are closed. When the pressure at the location of the Pirani gauge M_1 has dropped to about 0.1 torr and the two diffusion pumps are hot enough, V_1 closes and V_2 opens. If the pressure at M_1 again has dropped to 0.1 torr, V_2 closes and V_1 and V_3 open. The buffer vessel, the mercury ejector pump and the oil diffusion pump are then in series between the roughing pump and the tube. Once the pressure above the oil diffusion pump at the location of the Penning gauge M_2 has dropped to 10^{-3} torr, valve V_1 closes, the roughing pump stops and V_4 opens. Since after the second phase the roughing pump is isolated from the microscope tube by the closing of V_2, contamination of the vacuum system by oil from the roughing pump is minimal; there is no back diffusion of oil vapour above about 0.1 torr. Another favourable feature here is that the roughing pump normally operates for so short a time that it barely gets warm.

There are of course various safety devices in the automatic system which protect the microscope from damage if the water-cooling should fail, if there should be a serious vacuum leak, if a heating element should break down, and so on. The high-vacuum valve V_3, for example, is arranged so that the valve closes immediately if there is a mains failure.

Other differences between the equipment of this and the earlier Philips microscopes are the incorporation of a pump unit with a very much higher pumping speed and the provision of a number of vacuum valves in the microscope tube and in the viewing chamber. As we noted earlier, there is no need to let air into the entire microscope tube of the EM 300 when changing a filament or loading the plate camera.

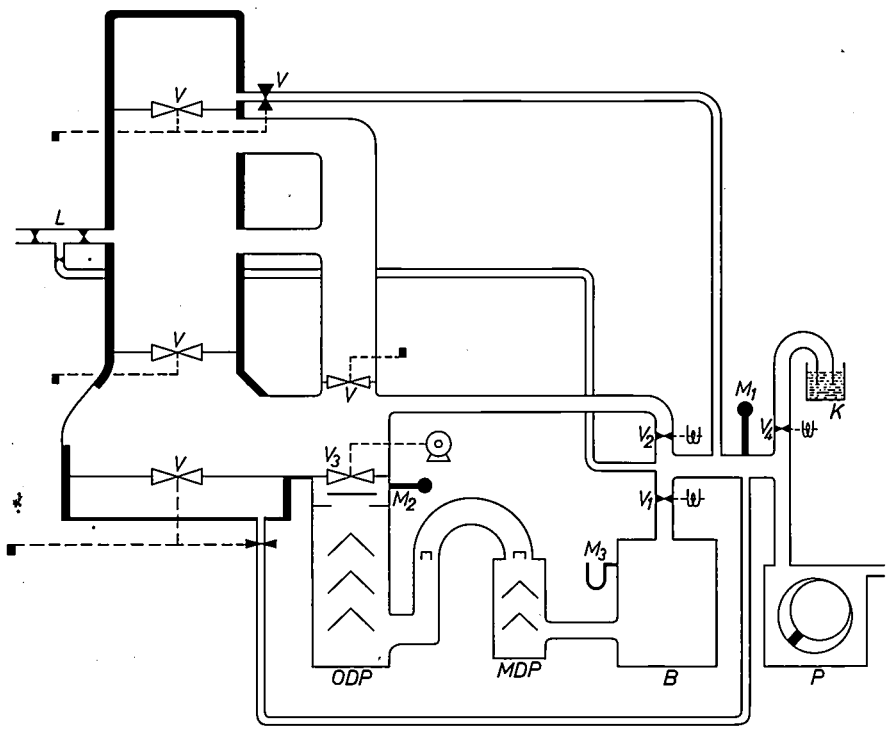


Fig. 12. The vacuum system. The microscope tube, shown on the left, is connected by three broad tubes with the oil diffusion pump *ODP*. *MDP* mercury ejector pump. *B* fore-vacuum vessel. *P* roughing pump. M_1 Pirani gauge. M_2 Penning gauge. M_3 mercury pressure gauge. *V* vacuum valves. *L* vacuum lock, which can be separately evacuated. *K* vessel with drying agent. Valves V_1, V_2 and V_4 are electromagnetic, V_3 is motor-driven.

The *electrical equipment* of the EM 300 is rather different from that of the earlier Philips microscopes. The most conspicuous difference is the use of transistors instead of valves; in the whole of the electrical equipment there are only two valves, in the high-voltage stabilizing circuit. The advantage of transistor circuits is that they are more reliable and develop less heat. It is also much easier to use printed wiring with transistor circuits than with valve circuits. The changeover to transistors made it necessary to develop new circuits for most sections of the electrical equipment.

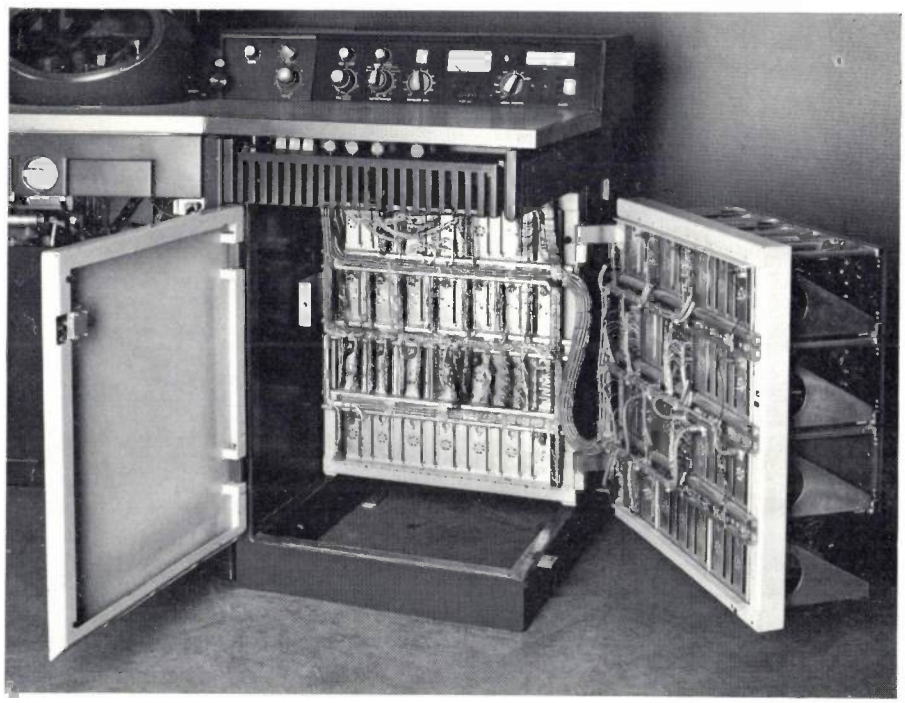


Fig. 13. The electrical circuits are contained in modules that fit into hinged racks.

The equipment has also been extended. In addition to the vital components, like the circuits that supply the high voltage and the lens currents, there is a monitoring circuit that can be connected to no fewer than 18 points of these circuits, enabling the user to measure (or record) the principal currents and voltages of the microscope or to check their stability.

To minimize magnetic interference with the beam, the transformers used in the circuits are all C-core types, which have an extremely weak stray field.

All the electrical circuits are contained in uniform modular units. This makes the most efficient use of the available space, and all the components are readily accessible (*fig. 13*). The power supply for the complete equipment comes from a separate transformer, arranged in such a way that the microscope can be fed from any 50 or 60 Hz single-phase mains supply whose voltage is between 110 and 440 V.

In the following parts of this section we shall deal at somewhat greater length with some important circuits.

The universal control amplifier

In order to be able to stabilize the lens currents and high voltage to such a degree that their fluctuations do not prevent the theoretical resolving power from being achieved, fluctuations of about 1 in 10^6 have to be detected. Since low voltages are used in transistor circuits, the control voltage derived from a lens current is also low, only a few volts. Variations of a few microvolts in the control voltage must therefore still be readily detectable. This calls for a highly sensitive amplifier with a carefully designed wiring layout: there must be no risk of interference from magnetic or electric fields or from coupling with other circuits. To offset these difficulties, there is the advantage that in transistor circuits the resistances across which the control voltages appear are small, so that the amplifier does not need to have a high input impedance.

The preamplifier which we have developed for the control circuits required in the EM 300, and which meets the above-mentioned requirements, is a special

type of chopper amplifier. Its principle of operation is illustrated in *fig. 14*, and the main features are summarized in the caption. The d.c. input voltage is converted by the transistor circuit on the left (the chopper) into an alternating voltage with a frequency of 1200 Hz and then passed by a transformer to an a.c. amplifier. The output signal from this amplifier is converted into a d.c. voltage again by the circuit on the right.

The distinguishing feature of the universal control amplifier is that the conversion from d.c. into a.c. voltage is not effected mechanically, but by means of a transistor circuit. This arrangement was chosen for maximum reliability and long life. As can be seen, the input circuit is separated by a transformer from the a.c. amplifier; this gives considerable freedom from interference, and permits the unit containing the sampling resistor and the input circuit to be earthed separately. The output signal from the a.c. amplifier is not rectified by an ordinary rectifier but by a circuit which is more or less the counterpart of the chopper circuit, connected to the same oscillator (synchronous detection).

The chopper circuit in fig. 14, like any other chopper, causes interference signals, mainly during switching. Various measures have been taken to minimize these, but they can still reach a peak value of about 200 μV, which is far in excess of the variations of the control signal. Nevertheless, the new amplifier will give faithful amplification of control signals of 1 μV. This is due in the first place to the synchronous detection, and also to the fact that the amplification has not been made unduly high, so that input signals may be as high as 200 μV without blocking occurring in the later stages of the amplifier.

Finally, some remarks on the chopper circuit. As can be seen, it consists of a symmetrical arrangement of four transistors. These are germanium transistors, selected in matched pairs for identical collector-emitter voltage at zero collector current. Germanium transistors are preferable here to silicon transistors, because in the transition from "completely" off to "completely" on (and *vice versa*) they require a much smaller signal on the base (70 mV r.m.s.). The use of germanium transistors therefore means that there is much less interference from crosstalk between the switching signal and the control signal — the principal type of interference — than there would be if silicon transistors were used. Germanium transistors, however, have a greater leakage current; as a result, a fairly large drift can occur if the voltage source delivering the signal to be amplified has a high internal impedance.

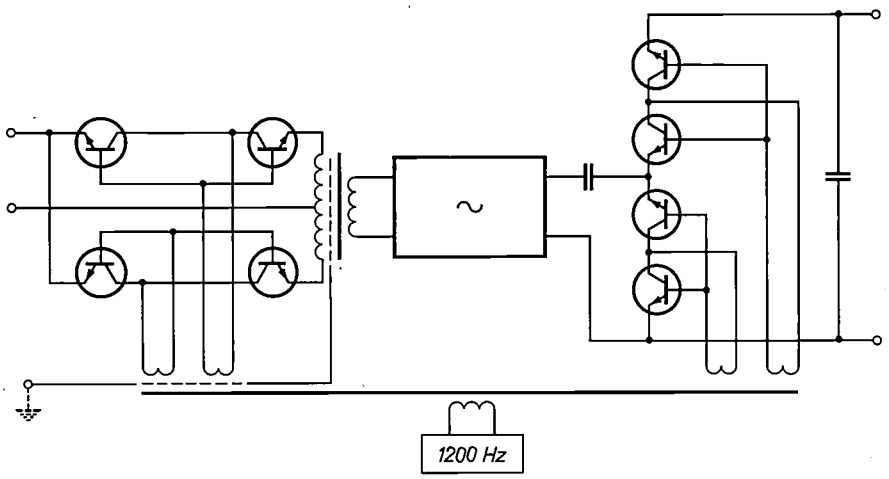


Fig. 14. Universal control amplifier. The circuit consists of a transistor chopper which converts the d.c. input voltage into an a.c. voltage, a transformer, an a.c. voltage amplifier and a synchronous detector. The chopper and the synchronous detector operate at a frequency of 1200 Hz and are controlled by a common oscillator via a transformer. Data: Floating input. Input impedance >20 000 Ω. Amplification 1000 times. Frequency response 0-100 Hz (—3 dB). Output voltage swing ±10 V. Between 0 and 50 °C the input drift at a maximum source impedance of 1000 Ω is less than 1 μV/°C and less than 1 μV/8 hours; the noise (peak-to-peak; 0-10 Hz) is less than 1 μV. At a maximum source impedance of 20 000 Ω these values are three times as high.

Although it might appear that only two transistors could be used for the chopper circuit, the arrangement with the two *pairs* of transistors connected in opposition has been chosen since a transistor is not such an ideal switch as a mechanical switch. In the present circuit the zero errors of the members of each pair generally compensate each other.

The chopper amplifier is included with a d.c. difference amplifier of conventional design (gain 2000) in a

rack-mounted module of the kind already mentioned (*fig. 15*). The electrical equipment of the EM 300 contains altogether nine such modular units. They are used for the lens-current and high-voltage stabilizers, for the monitoring circuit and also for the circuit that supplies the reference voltage with which the control signal is compared in each stabilizing circuit.

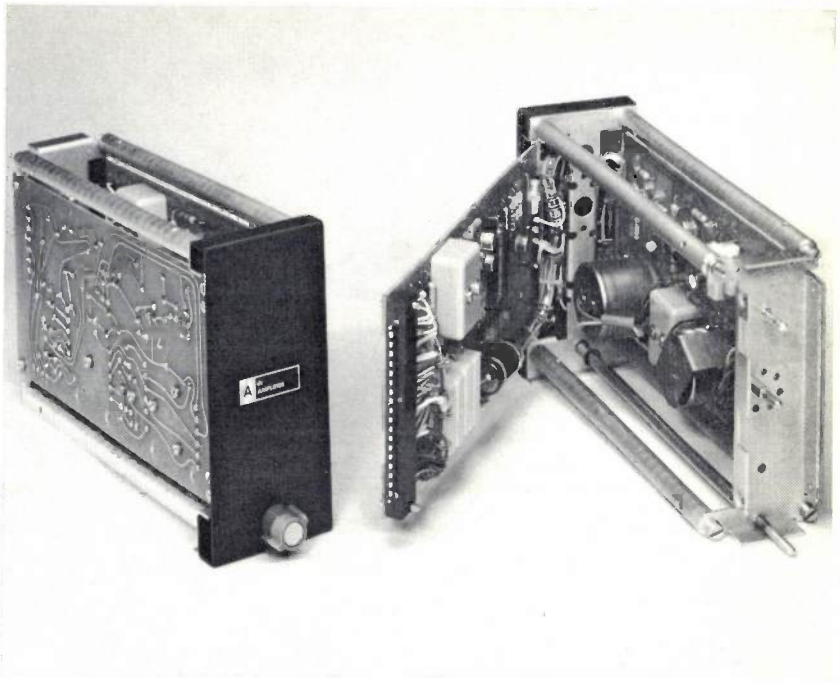


Fig. 15. Standard control amplifier, containing a circuit like that of fig. 14 and a conventional d.c. difference amplifier. The two amplifiers are mounted on printed wiring boards. Note the mu-metal screening around the transformers.

The central reference voltage

For stabilizing the high voltage and the lens current a reference voltage is needed that meets at least the same requirements of stability. A simple and very effective solution is to derive the reference voltage from a battery, but in that case the battery must not deliver any current. In the EM 300 this is not practicable for the following reason. As the power for the lenses comes from transistor circuits, the currents used are relatively high. The sampling resistors must therefore be small (of the order of 1 Ω). This means that the current cannot be varied by varying the sampling resistance, because in a resistance of about 1 Ω the variations in the contact resistance would be too large compared with the total value. The lens current can therefore be varied only by varying the reference voltage, that is to say by connecting a potentiometer circuit across the source supplying this voltage. This rules out the use of a battery. For the EM 300 a reference-voltage source has therefore been developed that can deliver a current and meets the appropriate stability requirements. Its output voltage is about 10 V and varies less than 1 in 10^6 per minute and less than 3 in 10^5 in 24 hours. The circuit can deliver 50 mA and is short-circuit proof. The temperature coefficient averages 10^{-6} per °C and is never more than 3×10^{-6} per °C. All stabilizing circuits derive their reference voltage from this single voltage source.

Fig. 16 shows a block diagram of the reference voltage source. The main units are the amplifiers discussed above (fig. 15) and a bridge circuit consisting of three resistors (R_1, R_2, R_3) and a Zener diode selected for low noise. The bridge circuit is housed in a small thermostat. The principle of operation is very simple. As soon as the part of the output voltage determined by R_2 and R_3 differs from the operating voltage of the Zener diode, the chopper amplifier receives an input signal and the output voltage is returned to the correct value. Variations too fast to pass the chopper amplifier are fed via capacitor C straight to the input of the difference amplifier.

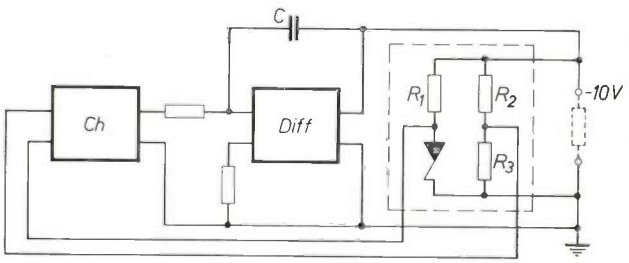


Fig. 16. Generator for the central reference voltage. *Ch* chopper amplifier. *Diff* difference amplifier (see fig. 15). The dashed line encloses a bridge circuit consisting of three resistors and a Zener diode, housed in a thermostat. *C* feedback capacitor for the suppression of fast fluctuations.

Generation and stabilization of lens current and high tension

Fig. 17 gives a block diagram of the power supply system for the lenses and shows how the lens currents are stabilized. All the lenses are fed from a single rectifier which can supply a current of 20 A. The output voltage is 42 V unstabilized and 33 V stabilized. From the output of the voltage stabilizer the current is distributed to the six lens circuits, each of which is a series configuration of a current regulator T_D, a sampling resistor R_m and the lens coil with a polarity reversal unit *RU*. A power transistor, or a number of power transistors in parallel, act as the current regulator which, to-

gether with the sampling resistor, forms part of the current stabilizing circuit; these transistors are mounted on a water-cooled plate. All stabilizing circuits are short-circuit proof — if there is a short-circuit or overload, the current is reduced to a safe value. In addition, the stabilizer for the objective-lens current contains a cut-out which switches off the lens current if one of the parallel transistors becomes overloaded or fails. (With a parallel arrangement of five transistors this is not immediately noticed, but it is with one or two.) The sampling resistors are resistance-alloy coils, housed in a water-cooled oil bath to ensure stability.

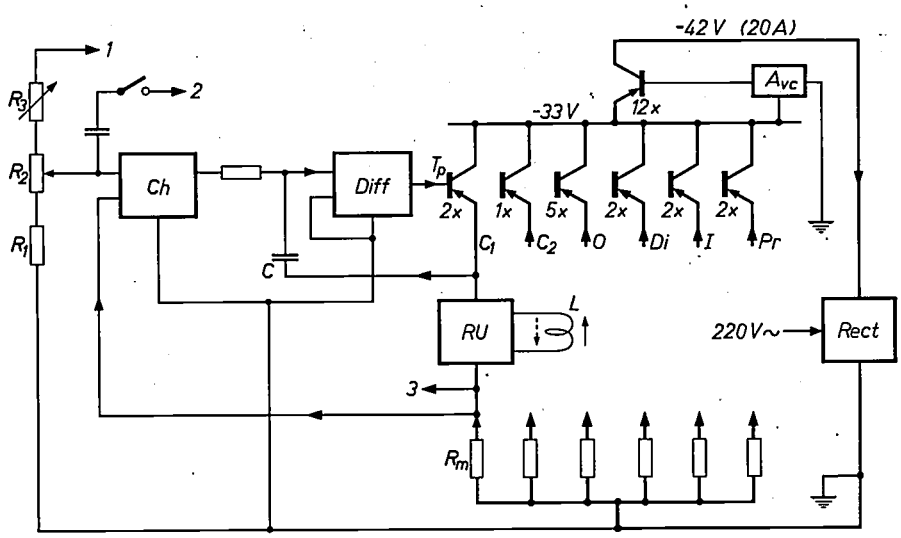


Fig. 17. Circuit for generating and stabilizing the lens currents. *Rect* rectifier. A_{vc} voltage-stabilizing amplifier. T_p power transistor(s). *RU* current reversal unit. *L* lens coil. R_m sampling resistor. *Ch* chopper amplifier. *Diff* difference amplifier. *1* connection with central reference voltage. *2* connection with 1.5 Hz generator. *3* connection with monitoring circuit. The resistor R_3 is connected to the high-voltage selector switch. The lens current can be varied by means of the potentiometer R_2.

The current is reversed by means of mercury-wetted reed relays. The entire operation is completely automatic and its sequence is as follows. First the lens coil is short-circuited and almost simultaneously cut off from the current source. After about half a second, which is sufficient for the dissipation of the field energy stored in the coil, the coil connections are reversed, the short-circuit is removed and the current again switched on. These relays are extremely reliable, and they are also bounce-free and have a very low and stable contact resistance.

Slow variations in the lens current are suppressed by applying the difference between the voltage across the sampling resistor and the reference voltage as an input signal to the control amplifier described above (fig. 15). The output signal from this amplifier drives the power transistor T_p. This control does not respond to fast fluctuations, as the chopper amplifier cannot handle them. The output of T_p is therefore directly connected to the input of the difference amplifier through a capacitor C (cf. fig. 16).

As explained above, changing over to another lens current entails altering the reference voltage. This is done by means of a potentiometer circuit (on far left in fig. 17) consisting of a fixed resistor R_1, a resistor with tap R_2 and a variable resistor R_3. The form of R_2 differs for each lens, varying from a very simple to a highly complicated arrangement. The objective-lens current can be varied by means of a divider in very small steps over a large range (the smallest steps are 4 in 10^6).

More than one potentiometer is provided for three lenses. This makes it possible for the user to switch from normal magnification to very low magnification, or to electron diffraction, by merely turning the function switch mentioned previously.

There is a similar facility, as in previous Philips microscopes, for changing over to another high voltage. The resistor R_3, which controls the magnitude of the current in the potentiometer circuits, is coupled to the high-voltage switch and controls the current in such a way that when the high voltage is changed to another value the illumination, magnification and focusing remain practically unchanged.

A low-frequency alternating current (1.5 Hz) can be applied to the input of the chopper amplifier. In this case the lens current is modulated approximately 7%, which is necessary for alignment. (Changing the strength of a lens causes a rotation of the image, so that the point where the optical axis cuts the image plane can be found.)

The recording shown in *fig. 18* gives a good idea of the stability achieved. There is little variation in current even over periods much longer than the one minute mentioned earlier.

The high voltage is stabilized by means of two control circuits, which have some sections in common. The first is completely analogous with the circuit of fig. 17; the correction signal produced in this circuit is added to the output voltage of the high-voltage generator. This circuit has a limited control range, however (about 1200 V on either side), which would rapidly be exceeded if no further provision were made. This difficulty is avoided by arranging that the output voltage also controls the stabilizing circuit that holds the primary voltage for the high-voltage transformer constant. In this way only slow variations can be smoothed out,

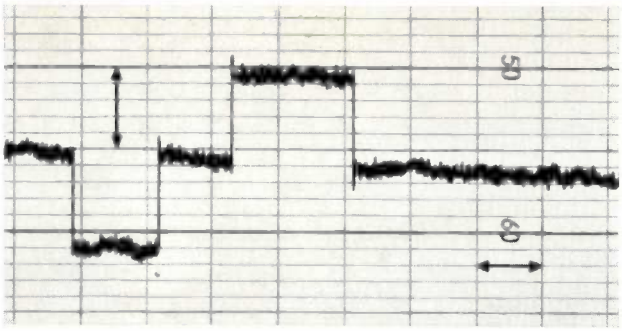


Fig. 18. Recording of lens current. The length of the horizontal double arrow corresponds to a time interval of two minutes, the length of the vertical one to a relative current variation of 5×10^{-6}. The fast current fluctuations and also the drift during one minute are no greater than about 1 in 10^6. During the recording the current was varied twice in steps of 5×10^{-6} (corresponding defocusing 12.5 nm).

but the control range is very wide (0 to 100 kV). Together, the two circuits are capable of keeping the high voltage constant to the extent required. For interfering signals at a frequency lower than about 10 Hz the control is principally achieved through the supply voltage. As a result of the control method described, the loop gain is very high at low frequencies (more than 10^5 below 1 Hz).

The accessories

The accessories for an electron microscope are almost as important as the microscope itself, since they often determine whether or not an investigation can be carried out. Besides the accessories we have already mentioned, like the cameras, the television camera tube and the image intensifier, the EM 300 has many others: we shall just mention a few of them here. One important accessory is a cooling element (below 120 °K), which can be brought close to the specimen in the tube to reduce contamination of the specimen by carbon formed by the breakdown of hydrocarbon molecules (see fig. 10*b*). There are also a large number of specimen holders enabling the specimen to be put under tension, cooled (to below 120 °K), heated (to 1250 °K) and so on. A special type of specimen stage, called a goniometer stage, enables the user to put the specimen into any required orientation (*fig. 19*). This is arranged so that the specimen can be tilted about an axis that

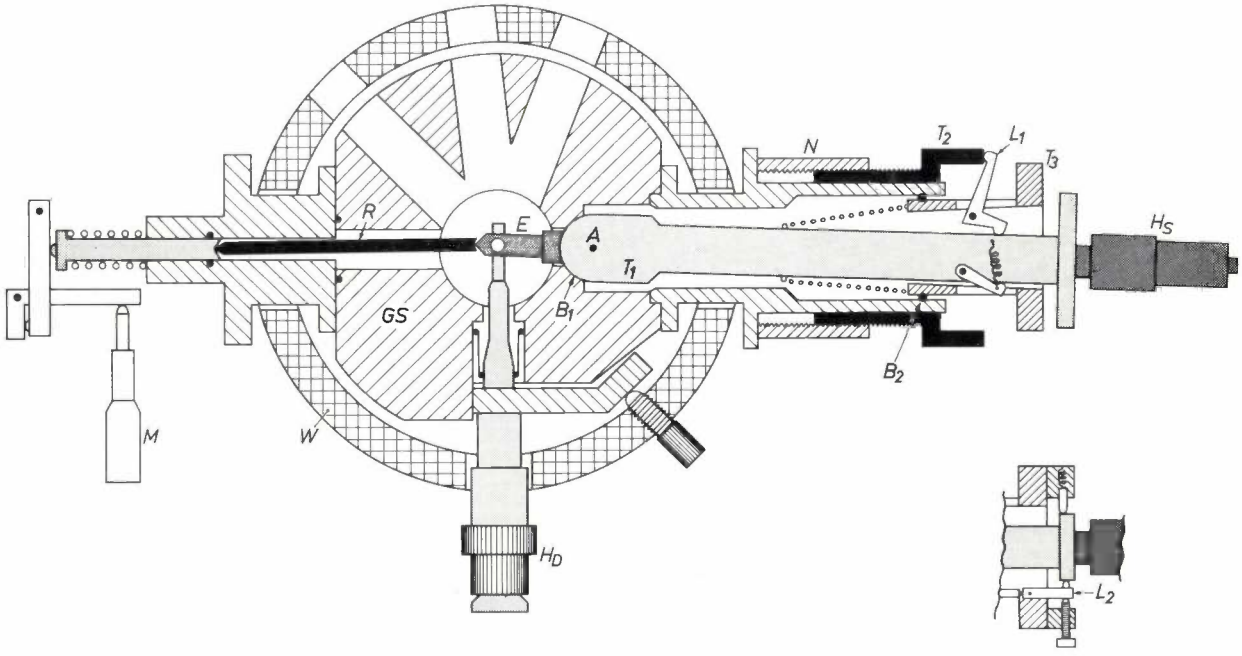


Fig. 19. Horizontal cross-section of the microscope tube with goniometer specimen stage. *W* microscope tube (cf. fig. 9). *GS* specimen stage which has a system of tubes attached to its right-hand side. Inside these tubes there is a much narrower tube T_1 (lightly shaded), whose left-hand end is hemispherical and bears against the spherical surface B_1. The position of the other end of T_1 is determined by the lever L_1, which is moved by the tube T_2 when the nut *N* is turned. In the plane perpendicular to the plane of the drawing there is a second lever L_2 (see detail at bottom right). The end *E* of the specimen holder H_S can be introduced into the microscope tube through the tube T_1; the depth of insertion is determined by the push-rod *R*, which is adjusted by the micrometer *M*.

When the tube T_3 resting in ball-bearing B_2 is rotated about its axis, the tube T_1 and the specimen holder follow it, so that the angle between the plane of the specimen and the horizontal plane also varies. In principle this angle can have any value, but in practice it is limited to about 60°, since at larger angles the electrons will no longer be able to pass through the opening in the specimen holder. At every position of the specimen plane the specimen can be moved in two directions in this plane: in the axial direction of the specimen holder by means of the micrometer *M*, and at right angles to this by rotating the nut *N*. Rotating the nut causes the tube T_1 to rotate around point *A*, so that in the drawing, for example, *E* rises when L_1 pushes down the right-hand end of T_1. H_D is the objective-diaphragm holder.

cuts the microscope axis, and so that any given detail of the specimen under investigation can be made to coincide with this point of intersection. A change in the position of the specimen thus causes no change in focus or magnification, nor any change in effective camera length for diffraction exposures.

Among the many specimen holders that can be fitted in the goniometer stage there is one that permits the specimen to be rotated 360° in its own plane about an axis through its centre. This facility allows a free choice of the azimuth of the tilt we have just mentioned.

When the goniometer stage is used, the pole-pieces of the objective lens illustrated in fig. 9 have to be replaced by others with a wider gap since otherwise there will not be enough room for the tilting of the specimen. This has the result of giving the objective a greater focal length (4.1 mm). In order to concentrate the field as much as possible in the space below the specimen, the pole-pieces are unequal. When the goniometer stage is used with the objective lens modified in this way a resolving power better than 15 Å can be achieved.

The very extensive range of accessories and the good resolving power make the Philips EM 300 electron microscope particularly suitable for many widely divergent types of investigation.

Summary. The Philips electron microscope type EM 300 differs from its predecessors in that it has an even better resolving power. In normal use this is always better than 5 Å and in favourable circumstances the theoretical resolving power can be closely approached. Operation has also been further simplified by including various vacuum locks, a semi-automatic exposure meter and a "function switch", etc., and also by the design of the instrument as a whole. The image can be displayed on a television screen as well as on the usual fluorescent screen. The magnification can be varied from 220 times to 500 000 times. Nearly all of the electron-optical sections can be aligned. The electrical circuits are all transistorized. The vacuum system is fully automatic in operation and has a high pumping speed. The excellent resolving power is attributable to the very short focal length of the objective lens (1.6 mm), to good mechanical stability and to the careful stabilization of lens current and high voltage (the relative variations are less than 2.5×10^{-6} and 5×10^{-6} per minute respectively). There is a wide range of accessories, including a goniometer specimen stage, a cryogenic anti-contamination device and a variety of specimen holders.

Recent scientific publications

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

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

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

G. A. Acket: The influence of doping fluctuations on limited space-charge accumulation in *n*-type gallium arsenide.
Physics Letters **27A**, 293-294, 1968 (No. 5). *E*

F. R. Badcock & D. R. Lamb (Associated Semiconductor Manufacturers Ltd., Wembley, England): Stability and surface charge in the MOS system.
Int. J. Electronics **24**, 1-9, 1968 (No. 1).

J. R. A. Beale: Modern developments in semiconductor devices.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **1**, 1229-1240, 1968 (No. 10). *M*

P. Billard & C. Hily: Les systèmes optiques à champ étendu utilisables dans l'infrarouge.
Acta electronica **11**, 237-325, 1968 (No. 3). *L*

G. Blasse: Fluorescence of uranium-activated compounds with rocksalt lattice.
J. Electrochem. Soc. **115**, 738-742, 1968 (No. 7). *E*

G. Blasse: Crystal structure and fluorescence of compounds $Ln_2Me^{4+}Me^{6+}O_8$.
J. inorg. nucl. Chem. **30**, 2091-2099, 1968 (No. 8). *E*

G. Blasse: Fluorescence of compounds with fresnoite ($Ba_2TiSi_2O_8$) structure.
J. inorg. nucl. Chem. **30**, 2283-2284, 1968 (No. 8). *E*

P. F. Bongers, C. F. van Bruggen (Rijksuniversiteit Groningen), **J. Koopstra** (R.U. Gr.), **W. P. F. A. M. Omloo** (R.U. Gr.), **G. A. Wiegers** (R.U. Gr.) **& F. Jellinek** (R.U. Gr.): Structures and magnetic properties of some metal (I) chromium (III) sulfides and selenides.
J. Phys. Chem. Solids **29**, 977-984, 1968 (No. 6). *E*

G. Bouwhuis: Interferometrie met gaslasers.
Ned. T. Natuurk. **34**, 225-232, 1968 (No. 8). *E*

A. Bril, W. L. Wanmaker (Philips Lighting Division, Eindhoven) **& J. W. ter Vrugt** (Philips L.D.): $Sr_3(PO_4)_2$-Tb, a bluish-white emitting cathode-ray phosphor with a long decay time.
J. Electrochem. Soc. **115**, 776-777, 1968 (No. 7). *E*

K. Bulthuis: Anomalous penetration of Ga and In implanted in silicon.
Physics Letters **27A**, 193-194, 1968 (No. 4). *E*

K. H. J. Buschow: Crystal structures, magnetic properties and phase relations of erbium-nickel intermetallic compounds.
J. less-common Met. **16**, 45-53, 1968 (No. 1). *E*

B. L. Cardozo & R. J. Ritsma (Institute for Perception Research, Eindhoven): On the perception of imperfect periodicity.
IEEE Trans. **AU-16**, 159-164, 1968 (No. 2). *E*

H. B. G. Casimir: On Bose-Einstein condensation. Fundamental problems in statistical mechanics II, editor E.G.D. Cohen, North-Holland Publ. Co., Amsterdam 1968, p. 188-196. *E*

A. Cohen (Institute for Perception Research, Eindhoven): Errors of speech and their implication for understanding the strategy of language users.
Z. Phonetik, Sprachwiss. & Komm.f. **21**, 177-181, 1968 (No. 1/2).

W. F. Druyvesteyn, C. A. A. J. Greebe, A. J. Smets & R. Ruyg: Faraday rotation of Alfvén waves in liquid sodium.
Physics Letters **27A**, 301-302, 1968 (No. 5). *E*

Y. Genin: A note on linear minimum variance estimation problems.
IEEE Trans. **AC-13**, 103, 1968 (No. 1). *B*

A. A. van der Giessen, J. G. Rensen & J. S. van Wieringen: A study of the constitution and freezing behaviour of iron oxide-hydrate gels by means of the Mössbauer effect.
J. inorg. nucl. Chem. **30**, 1739-1744, 1968 (No. 7). *E*

L. Heijne: Ionic conductivity in oxides: experimental problems, survey of existing data.
Mass transport in oxides, Proc. Symp. Gaithersburg, Maryland, 1967 (NBS Special Publ. No. 296, 1968), p. 149-164. *E*

H. Jonker, C. J. G. F. Janssen & Th. P. G. W. Thijssens: PD-Photoplating. A new photographic additive printed circuit process.
Photographie und Film in Industrie und Technik, Ber. I. int. Kongress, Köln 1966, p. 249-253; 1968. *E*

A. van Katwijk (Institute for Perception Research, Eindhoven): A functional grammar of Dutch number names.
Grammars for number names, editor H. Brandt Corstius, D. Reidel Publ. Co., Dordrecht 1968, p. 1-8.

D. R. Lamb & F. R. Badcock (Associated Semiconductor Manufacturers Ltd., Wembley, England): The effect of ambient, temperature and cooling rate, on the surface charge at the silicon/silicon dioxide interface.
Int. J. Electronics **24**, 11-16, 1968 (No. 1).

R. Memming & G. Schwandt: Electrochemical properties of gallium phosphide in aqueous solutions.
Electrochim. Acta **13**, 1299-1310, 1968 (No. 6). *H*

J. Neirynck: Predistortion of high-pass filters for the compensation of parasitic capacitances.
Network and switching theory, editor G. Biorci, Academic Press, New York 1968, p. 273-282. *B*

A. van Oostrom: Field emission and surface structure.
A survey of phenomena in ionized gases, Int. Atomic Energy Agency, Vienna 1968, p. 291-308. *E*

G. den Ouden: Internal friction of weld metal.
Brit. Welding J. **15**, 436-447, 1968 (No. 9). *E*

P. J. L. Reijnen: Sintering behaviour and microstructures of aluminates and ferrites with spinel structure with regard to deviation from stoicheiometry.
Science of Ceramics **4**, 169-188, 1968. *E*

C. J. M. Rooymans: The effect of very high pressures on ceramic systems.
Science of Ceramics **4**, 133-149, 1968. *E*

J. M. Stevels: Permitted and forbidden structural units in vitreous systems.
Mat. Res. Bull. **3**, 599-609, 1968 (No. 7). *E*

G. W. Tichelaar & J. G. Verhagen: Woher kommen die Poren beim Schutzgasschweißen unter Mischgas?
Der Praktiker **20**, 132, 1968 (No. 6). *E*

J. Tillack: Gravimetrische Analyse von Halogeniden und Oxidhalogeniden nach der „H-Rohr-Methode".
Z. anal. Chemie **239**, 81-87, 1968 (No. 2). *A*

J. A. Weaver: Machines that read.
Science Journal **4**, No. 10, 66-71, 1968. *M*

K. Weiss: Die Beweglichkeit von Elektronen in AgBr.
Z. phys. Chemie Neue Folge **59**, 242-250, 1968 (No. 5/6). *E*

K. Weiss: Zur Umwandlung der „Tieftemperaturphasen" des AgJ in die α-Modifikation.
Z. phys. Chemie Neue Folge **59**, 318-320, 1968 (No. 5/6). *E*

R. M. G. Wijnhoven: A simple high-speed magnetic access switch matrix.
IEEE Trans. C-**17**, 542-550, 1968 (No. 6). *E*

F. J. Witts & P. W. Whipps: An experimental slow-rate crystal puller for seeded growth from a flux.
J. sci. Instr. (J. Physics E), ser. 2, **1**, 970-971, 1968 (No. 9). *M*

Contents of Electronic Applications **28**, No. 3, 1968:

Contents of Mullard Technical Communications **10**, No. 94, 1968:

Volume 29, 1968, No. 12 pages 353-388 Published 28th February