PHILIPS TECHNICAL REVIEW

VOLUME 32, 1971, No. 1

Integrated linear basic circuits

Th. J. van Kessel and R. J. van de Plassche

Monolithic integrated circuits, which have developed in a matter of ten years from a laboratory experiment into a mass-produced product, can be divided into two main classes, digital and linear. The linear devices can fulfil many functions, but they always contain the same basic circuits. The authors present some elegant solutions for these linear basic circuits, making use of the special capabilities of integrated-circuit technology, in what amounts to a new departure in electronics.

Introduction

The unsuspecting layman viewing an integrated circuit for the first time under a microscope might well believe that he is looking at a piece of modern art. The abstract play of patches and lines, often beautifully coloured, does not suggest a deliberate pattern of shapes designed to replace a whole board packed with resistors and transistors.

It is not surprising that many people believe integration technology to be capable of making the impossible possible. Admiration for an imperfectly understood technology may lead people to overestimate its capabilities. It should be remembered, however, that integration technology is only one of the methods of manufacturing electronic circuits. Like any other method, it has its advantages and disadvantages, and a finished product of high quality will not be produced unless the electronic engineer's design is able to exploit the one and avoid the other.

In this article we shall describe some examples of linear basic circuits which take advantage of the particular capabilities offered by integrated-circuit technology. To appreciate the beauty of these circuits we must first, however, take a closer look at the features of the new technology.

Integration techniques

In recent decades constant efforts have been made to produce electronic circuits more efficiently than can be done by making wire connections between individual components [1].

The first step was the introduction of printed circuits, the printed wiring pattern being applied to an insulating base, usually a resin-bonded paper board, by a photo-etching process. A second step was the advent of the "hybrid" circuit [2], in which not only the wiring but also the resistors and smaller capacitors are evaporated on to a glass or ceramic substrate. The active devices and other components are soldered on later. The circuits are called "hybrid" because of the combination of individual components with vacuum-evaporated ones; there is integration to the extent that some of the components are fabricated as a whole.

Both of these techniques obviously allow complete freedom in the choice of the active devices. There is no reason why *NPN*, *PNP*, field-effect and MOS transistors should not be included side by side in the same circuit. The introduction of these technologies therefore did not radically alter the work of the circuit designer.

A complete change was brought about, however, by the monolithic or solid circuit [3], which is usually what is intended by the term "integrated circuit". Here all the circuit elements, both active and passive, are

Ir. Th. J. van Kessel and Ir. R. J. van de Plassche are with Philips Research Laboratories, Eindhoven.

[1] P. W. Haaijman, Integration of electronic circuits, Philips tech. Rev. 27, 180-181, 1966.

[2] E. C. Munk and A. Rademakers, Integrated circuits with evaporated thin films, Philips tech. Rev. 27, 182-191, 1966.

[3] A. Schmitz, Solid circuits, Philips tech. Rev. 27, 192-199, 1966.

formed at the same time in a thin layer of a silicon wafer by the "planar" technique, in a succession of oxidizing, photo-etching and diffusion processes. Finally the elements are interconnected by means of an evaporated pattern of conductor. The fact that all the elements are formed in the same steps of the process makes them interdependent. For example, the starting material and the individual steps could be chosen so as to produce optimum *NPN* transistors. But generally speaking this choice would then not be optimum for the other elements. The introduction of the monolithic integration technique therefore demanded a different approach from the circuit designer; the planar technique has its own special advantages, but it also has its limitations.

Capabilities and limitations of the planar technique; a new electronics

Advantages of the planar technique

In every circuit the external contacts are possible sources of undesirable effects, and the reliability of a circuit generally decreases with an increase in the number of contacts. Complicated circuits are more reliable in integrated form than when they are built up from separate transistors, since the total number of contacts is then much smaller than the total number of contacts of all the individual transistors. This greater reliability is the important factor that has led to growing interest in the building of systems from standard integrated-circuit units.

Provided the masks for the photo-etching operations in the planar technique are accurately drawn and the various manufacturing steps are carried out with scrupulous care, it is possible to produce almost identical transistors in an integrated circuit. The variation of base-emitter voltage V_{BE} with collector current I_{C} can be reproduced fairly easily ($\Delta V_{\mathrm{BE}}/V_{\mathrm{BE}} < 1\%$), but less success is achieved with the current-gain factor h_{FE} ($\Delta h_{\mathrm{FE}}/h_{\mathrm{FE}} < 10\%$).

The reason for this is that, as can be seen from the relation

$$I_{\mathrm{C}} = I_{\mathrm{C0}}\,(\mathrm{e}^{eV_{\mathrm{BE}}/kT} - 1) \qquad (1)$$

(where e is the electronic charge, k Boltzmann's constant and T the absolute temperature), the collector current I_{C} at a given base-emitter voltage V_{BE} is proportional to the leakage current I_{C0}, and I_{C0} is in turn proportional to the surface area of the emitter. This area is critically determined by the masks used. The current gain factor h_{FE}, on the other hand, is dependent on the thickness d of the base layer, i.e. on the depth of the base diffusion less the depth of the emitter diffusion (*fig. 1*), and of course the difference between these two diffusion depths is much more difficult to make identical than the emitter areas.

The d.c. operating point of a transistor is affected by the temperature (see equation 1). Now in integrated circuits the distances between the elements are so small, and the thermal conductivity of the silicon is so high, that two closely adjacent transistors vary with temperature in practically the same way, provided that dissipating elements are kept sufficiently far apart.

This can be established by measuring the small difference in drift of the base-emitter voltage V_{BE} between two transistors in the same circuit that carry identical currents I_{C}. A value of 1 µV/°C is quite feasible. Now it can be demonstrated that at a V_{BE} of 0.6 V — approximately the value of V_{BE} at 100 µA — a temperature difference of 1 °C would cause a difference of 2 mV in V_{BE}. This indicates that the temperature difference between the transistors varies by only about 0.0005 °C for the same temperature change of 1 °C. By using balanced circuits, such as differential amplifiers, the temperature effects of the individual transistors can be made to compensate each other almost completely. Any slight temperature drift remaining is not so much due to temperature differences as to slight physical differences between the devices.

The price of monolithic circuits is determined by the initial costs, such as the costs of design and drawing the masks, and by the production costs, which in turn depend on chip size and on the number of contacts per circuit. Where large quantities are produced the initial costs are usually negligible; if the designer can succeed in keeping the chip size down and minimizing the number of contacts required, then the planar technique is very suitable for the inexpensive manufacture of reliable electronic circuits of high quality.

Limitations of the planar technique

It is not possible, as we have said, to choose the starting material and processes of the planar technique in such a way that optimum *NPN* and *PNP* transistors can be produced at the same time, perhaps with MOS or other field-effect transistors as well. *NPN* transistors

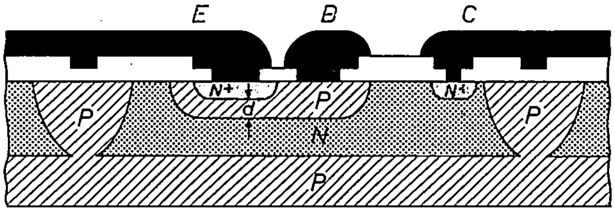


Fig. 1. Cross-section of an *NPN* transistor in an integrated circuit. The integrated circuit is made in a layer of *N*-type silicon applied epitaxially to a *P*-type silicon substrate. Part of the epitaxial *N*-type layer, separated from the rest by a *P*-type diffusion, serves as the collector. A *P*-type diffusion in the collector island forms the base of the transistor, an N^+ diffusion in the base forms the emitter. A second N^+ diffusion forms the contact with the *N*-type collector. *E*, *B* and *C* are metal conductors for the connections to emitter, base and collector. *d* thickness of base layer.

and resistors are usually regarded as the main product and *PNP* transistors as secondary products. Consequently the starting material is an epitaxially grown layer of *N*-type silicon, which serves as the collector material, into which a *P*-type region is subsequently diffused as the base, followed by the diffusion into this *P*-type zone of an *N*-type zone as the emitter (fig. 1). *PNP* transistors are made by diffusing two closely spaced *P*-type areas into the *N*-type layer. The characteristics of these lateral *PNP* transistors are not so good as those of the *NPN* transistors. The quality of the *PNP* transistors can be increased by means of a number of extra operations, but these of course make the circuit more expensive.

The resistors in a monolithic circuit are usually formed by channels of *P*-type material which are produced at the same time as the bases of the transistors in the epitaxial *N*-type layer. As might be expected, the accuracy of these resistors is not very high ($\Delta R/R \approx 10\%$), since the value depends not only on the surface area, which is determined by the masks, but also on the concentration of the *P*-type doping. The relative values of two resistors are maintained much more accurately (about 3%, and as good as 1% for resistors of the same value).

The same applies to the temperature coefficient. This is fairly high and, depending on the sheet resistance of the *P*-channel, varies between 0.1 and 0.3% per °C (the sheet resistance or surface resistivity is measured at the surface between two opposite sides of a square; the value is independent of the size of the square). Because the temperature difference is so small, the temperature coefficient of the ratio of two resistors may be much smaller.

Integration technique imposes a certain limitation on the size of the resistors. The area of one 10 kΩ resistor, for example, is equal to that of six transistors. In the circuits to be described transistors have deliberately been used instead of resistors wherever possible.

In cases where high resistances are indispensable, "buried resistors" sometimes provide the answer. These are resistors of *P*-type material covered by an *N*-type layer that is applied at the same time as the emitter diffusion. This has the effect of increasing the resistance value about ten times. Such a resistor has a field-effect-transistor configuration, and consequently the resistance depends on the voltage and there is some spread in the value.

In an integrated circuit a reverse-biased *PN* junction can be used as a capacitor. Its capacitance depends strongly on the reverse voltage. Another possibility is to apply an aluminium layer above an N^+ layer with the protective layer of silicon dioxide in the circuit acting as the dielectric.

The capacitance of both types of capacitor is proportional to the area that they occupy on the chip. A 200 pF capacitor occupies an area of the order of 0.1 mm². Only capacitors of very small values are therefore eligible for integration in a monolithic circuit.

Inductors cannot be made by the monolithic technique.

An integrated circuit always contains parasitic elements; the resistors and transistors, for example, always have parasitic capacitance to the *P*-type substrate (fig. 1). The parasitic *PNP* transistor formed in every *NPN* transistor by the base diffusion (*P*), the epitaxial layer (*N*) and the substrate (*P*) can often be particularly disturbing. This starts to conduct as soon as there is a forward voltage across the collector junction of the *NPN* transistor, which happens when it is driven into saturation. It may also happen, however, when the *NPN* transistor is operated as a diode by connecting the collector and base together (*fig. 2*), owing to the effect of the parasitic collector series resistance $r_{cc'}$ formed by the relatively poorly conducting epitaxial *N*-type layer. If the current through the transistor in a diode configuration becomes high enough for the voltage across the collector resistance $r_{cc'}$ to make the parasitic *PNP* transistor conduct, the current *I* will not flow entirely through the diode but partly through the *PNP* transistor to the substrate, which is at a negative potential to maintain the reverse-biased junction between substrate and epitaxial *N*-type layer. The collector series resistance can be reduced by means of a buried N^+ layer under the *N*-type silicon of the collector [3]. The collector contact diffusion is sometimes made so deep that it joins up with the buried layer, forming a "collector wall".

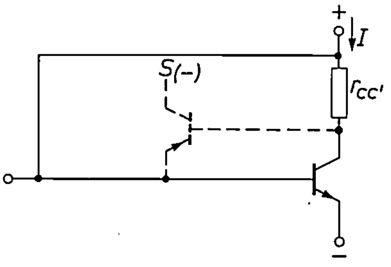


Fig. 2. Every integrated *NPN* transistor incorporates a parasitic *PNP* transistor, whose collector is formed by the substrate. If the *NPN* transistor is connected as a diode, the voltage across the internal collector resistance $r_{cc'}$ can make the parasitic *PNP* transistor conduct, so that part of the current flows to the substrate *S*.

A new electronics

The capabilities and limitations of integration technology make it necessary to rewrite or add new material to our textbooks on electronics.

In the chapter on *Basic Circuits*, for example, the ordinary amplifier stage (*fig. 3a*) is not suitable for integration because of its many resistors and the large decoupling capacitor across the emitter resistor. The differential amplifier (fig. 3*b*) can fulfil the same function [4] [5], and is very suitable for integration, particularly since the load resistors can be replaced by a controlled current source — a new basic circuit that consists entirely of transistors. The resistors needed for limiting thermal drift in the circuit shown in fig. 3*a* are superfluous here since the drift in two identical transistors operating at the same temperature is exactly the same and does not give rise to any voltage between the output terminals, and therefore produces no output signal. Nor are these resistors needed for the d.c. bias, since this is also supplied by a current source. Under the heading of *Basic Circuits* a considerable amount of space will therefore have to be devoted to current sources.

The use of differential amplifiers and current sources also offers a wide variety of possible ways of coupling amplifier stages; these would have to be included in the chapter on *Amplifier Circuits*.

In the following we shall discuss in turn a number of circuits (current sources, input amplifiers, output amplifiers) which have been designed on the principles of this new integration electronics. Combination of these component circuits on a single chip of silicon gives complete integrated circuits, such as the operational amplifiers that are used in instrument electronics.

Current sources

A current source has to supply a current that does not vary with the voltage across it; the ideal current source therefore has an infinitely high output impedance. In many cases it is desirable to be able to control the magnitude of the output current; in the circuits described here this is done by means of a reference current.

Controlled current source using two transistors

The simplest controlled current source consists of two identical transistors, one of which is connected as a diode (*fig. 4*). The two transistors have the same emitter area and therefore the same leakage current I_{C0}. Since they have the same base-emitter voltage V_{BE}, their collector currents I_{C1} and I_0 are also equal:

$$I_{C1} = I_0 = I_{C0}\,(e^{eV_{BE}/kT} - 1).$$

The base currents are thus $I_{C1}/h_{FE} = I_0/h_{FE}$ and are supplied by the reference-current source I_{ref}, so that $I_0 = I_{ref} - 2\,I_0/h_{FE}$ or

$$I_0 = I_{ref}\left(1 - \frac{2}{h_{FE}+2}\right), \qquad (2)$$

where the difference term expresses the two base currents. Since h_{FE} is of the order of magnitude of 100, I_0 is approximately equal to I_{ref}. The circuit gives gain because I_0 is delivered across a high output impedance, the collector impedance of transistor *2*, while the reference-current source (see page 7) has a conducting diode as its load and therefore does not need a high output impedance. Because of the symmetric structure of the circuit it is relatively insensitive to variations in temperature (eV_{BE}/kT has the same value for both transistors) and to voltage fluctuations.

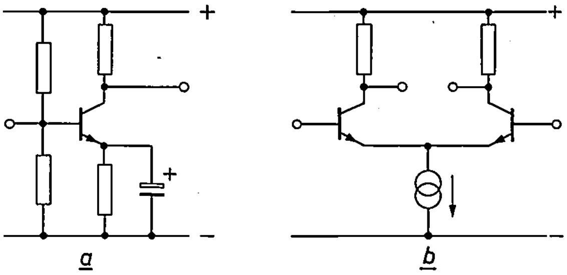


Fig. 3. *a*) Conventional transistor amplifier stage, not suitable for integration. *b*) Differential amplifier stage, very suitable for integration, particularly when the load resistors are replaced by transistor circuits.

With transistors of unequal emitter areas the ratio of the reference and output currents will be the same as the ratio of the emitter areas. Since these areas are fixed when the openings in the masks are drawn, their ratio can be fairly well controlled. To ensure accuracy in this ratio each of the two emitters is sometimes built up from a number of diffusions of equal magnitude.

As equation (2) shows, the emitter areas only determine the currents if the current gain h_{FE} of the transistors is sufficiently high.

An investigation of the behaviour of the circuit as a function of frequency involves the quantity h_{fe}, the current gain for the a.c. component in the base current of a transistor. At relatively low frequencies, h_{fe} is independent of frequency but at high frequencies h_{fe} decreases with rising frequency. The behaviour of h_{fe} as

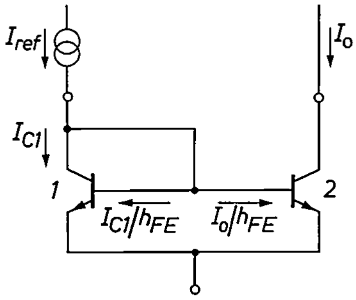


Fig. 4. Controlled current source with two transistors. Transistor *1* is connected as a diode. The output current I_0 is independent of the voltage across the output and is approximately equal to the reference current I_{ref}.

a function of frequency f is approximated by the expression

$$h_{\mathrm{fe}} = \frac{h_{\mathrm{fe0}}}{1 + \mathrm{j} h_{\mathrm{fe0}} f/f_{\mathrm{T}}}, \tag{3}$$

where h_{fe0} is the value of $|h_{\mathrm{fe}}|$ at low frequencies, and f_{T} is the frequency at which $|h_{\mathrm{fe}}|$ has decreased to 1. Substitution of this expression in (2) gives the following equation for the a.c. components in output and reference current:

$$I_{\mathrm{o}} = \frac{I_{\mathrm{ref}}}{1 + 2/h_{\mathrm{fe0}} + 2\mathrm{j}\, f/f_{\mathrm{T}}} \approx \frac{I_{\mathrm{ref}}}{1 + 2\mathrm{j}\, f/f_{\mathrm{T}}}. \tag{4}$$

We see from this that the equality of I_{o} and I_{ref} is no longer adequate when $f > \frac{1}{2} f_{\mathrm{T}}$; the frequency characteristic of this current source is given by curve *a* in *fig. 5*.

The collector-emitter breakdown voltage $V_{(\mathrm{BR})\mathrm{CE0}}$ of transistor *2* in fig. 4 is two or three times greater in this circuit than that of the transistor itself.

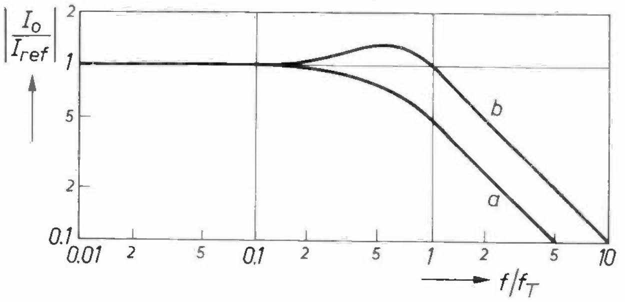


Fig. 5. Frequency characteristic of the controlled current source using two transistors (*curve a*) and of the controlled current source using three transistors (*curve b*).

This can be understood if we treat the whole circuit of fig. 4 as a single transistor with a current gain of $h_{\mathrm{FE}} = 1$ and consider that in general:

$$V_{(\mathrm{BR})\mathrm{CE0}} = V_{(\mathrm{BR})\mathrm{CB0}}\,(1 + h_{\mathrm{FE}})^{-1/N}$$

(where $V_{(\mathrm{BR})\mathrm{CB0}}$ is the collector-base breakdown voltage and N has a value between 2 and 4). For the individual transistor h_{FE} is about 100, and therefore $V_{(\mathrm{BR})\mathrm{CE0}}$ is two or three times lower than for our circuit as a whole, which has a breakdown voltage of $V_{(\mathrm{BR})\mathrm{CE0}} = 0.7$ to $0.8\ V_{(\mathrm{BR})\mathrm{CB0}}$. The breakdown effect is visible in the characteristics shown in *fig. 6a*.

Controlled current source using three transistors

By using a third transistor it is possible to make the output current I_{o} of the current source more accurately equal to the reference current I_{ref} (*fig. 7*). The circuit of fig. 7 operates by feeding back variations of the current through transistor *3* to the base of transistor *3* in the opposite sense by means of a current source like

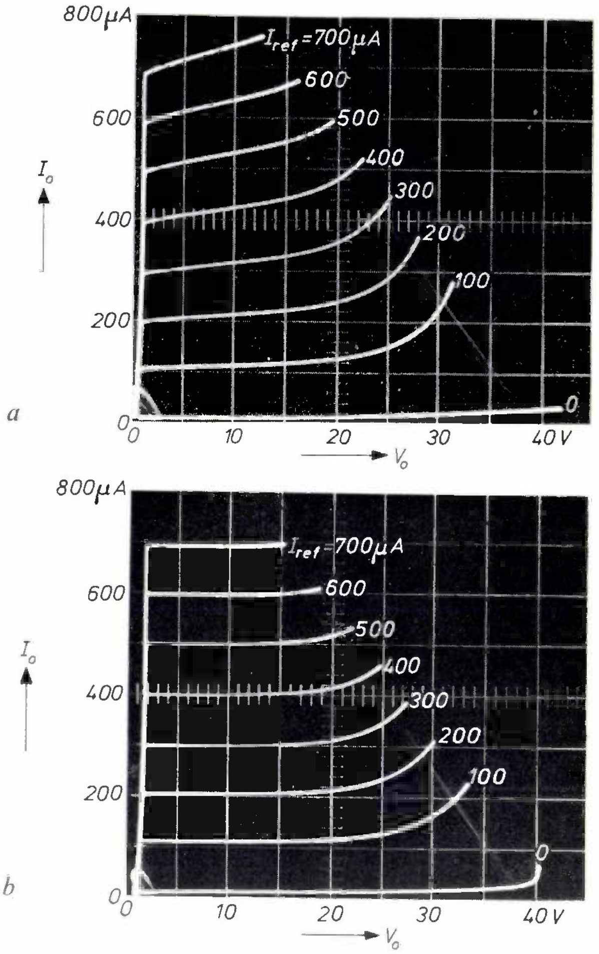


Fig. 6. Current-voltage characteristics of the controlled current source with two transistors (*a*) and of the controlled current source with three transistors (*b*). In both cases the breakdown voltage is about 0.8 times the collector-base breakdown voltage $V_{(\mathrm{BR})\mathrm{CB0}}$ of the output transistor. The curves in (*b*) are flatter because of the higher output impedance of the circuit with three transistors.

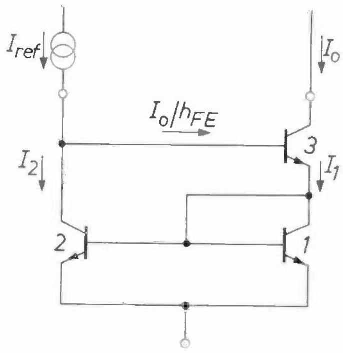


Fig. 7. Controlled current source with three transistors. I_{ref} reference current. I_{o} output current.

[4] G. Klein and J. J. Zaalberg van Zelst, General considerations on difference amplifiers, Philips tech. Rev. **22**, 345-351, 1960/61.

[5] G. Klein and J. J. Zaalberg van Zelst, Precision electronics, Philips Technical Library, Eindhoven 1967.

the one shown in fig. 4. For this current source eq. (2) shows that :

$$I_2 = I_1\left(1 - \frac{2}{h_{\mathrm{FE}}+2}\right).$$

From the circuit it also follows that:

$$I_1 = I_0 + I_0/h_{\mathrm{FE}},$$
$$I_2 = I_{\mathrm{ref}} - I_0/h_{\mathrm{FE}}.$$

From these three equations we arrive at the output current:

$$I_0 = I_{\mathrm{ref}}\left(1 - \frac{2}{h_{\mathrm{FE}}^2 + 2h_{\mathrm{FE}} + 2}\right). \qquad (5)$$

A comparison with equation (2) shows that the difference term here is about h_{FE} times smaller than in the case of the circuit with two transistors. There is no difference term of the order of $1/h_{\mathrm{FE}}$, since here I_{ref} and I_0 each deliver a single base current.

The output impedance of the current source is equal to $\frac{1}{2}h_{\mathrm{FE}}$ times the collector output impedance of a single transistor and therefore is $\frac{1}{2}h_{\mathrm{FE}}$ times greater than that of the circuit in fig. 4. The reference current source has the impedance of two diodes connected in series as its load.

An idea of the high-frequency behaviour of this current source can be obtained by substituting equation (3) in (5). Since the value of h_{fe0} is high, some terms can be neglected, and we obtain the following expression for the a.c. components of the currents:

$$I_0 \approx I_{\mathrm{ref}}\left(1 - \frac{2}{h^2_{\mathrm{fe0}} + 2\,h_{\mathrm{fe0}} + 2}\right) \times$$
$$\times \frac{1 + 2\,\mathrm{j}\,f/f_{\mathrm{T}}}{1 + 2\,\mathrm{j}\,f/f_{\mathrm{T}} + 2(\mathrm{j}\,f/f_{\mathrm{T}})^2}. \qquad (6)$$

Curve *b* in fig. 5 shows the variation of $|I_0/I_{\mathrm{ref}}|$ as a function of frequency. It can be seen from this that the current source with three transistors can be used up to higher frequencies than the one with two transistors.

The input and output impedances of the circuit can be calculated by determining the voltage variations at the base and collector of transistor *3* when I_{ref} and I_0 are varied. If I_{ref} changes by an amount ΔI_{ref}, then I_0 and I_1 change by the same amount and if S is the transconductance the current change ΔI_1 produces a voltage change $2\Delta I_1/S$ across the base-emitter junction of transistor *3* and across diode *1* [5]. The input impedance is therefore $2/S$, i.e. the impedance of two diodes in series.

At a variation of I_0 the value of I_{ref} remains constant, and since I_2 follows the variations of I_1 a variation of $\frac{1}{2}\Delta I_0$ occurs in both the emitter current and the base current of transistor *3*. The base current variation $\frac{1}{2}\Delta I_0$ gives a variation in the base-emitter voltage of $\frac{1}{2}\Delta I_0 h_{\mathrm{FE}}/S$. Transistor *3* amplifies this voltage μ times (μ is the amplification factor and is equal to the product of the output resistance R_0 of the transistor and the transconductance S), so that a voltage variation of $\frac{1}{2}\Delta I_0 h_{\mathrm{FE}}\mu/S = \frac{1}{2}\Delta I_0 h_{\mathrm{FE}} R_0$ appears at the collector. The output impedance of the current source is thus seen to be $\frac{1}{2}h_{\mathrm{FE}}R_0$, i.e. $\frac{1}{2}h_{\mathrm{FE}}$ times the output impedance of a single transistor.

It can be shown that the frequency characteristic should have the shape of curve *b* in fig. 5 by using Bode diagrams [6]. Using the current symbols to represent a.c. components again, we can write for the currents in transistor *3*:

$$I_1 = (h_{\mathrm{fe}} + 1)\, I_0/h_{\mathrm{fe}}.$$

Making use of equation (4), we can also write:

$$I_0/h_{\mathrm{fe}} = I_{\mathrm{ref}} - I_2 = I_{\mathrm{ref}} - \frac{I_1}{1 + 2\mathrm{j}\,f/f_{\mathrm{T}}},$$

from which it follows that

$$I_1 = \frac{I_{\mathrm{ref}}}{\dfrac{1}{h_{\mathrm{fe}} + 1} + \dfrac{1}{1 + 2\mathrm{j}\,f/f_{\mathrm{T}}}}. \qquad (7)$$

The Bode diagrams of $h_{\mathrm{fe}} + 1$ and of $1 + 2\mathrm{j}f/f_{\mathrm{T}}$ are given in *fig. 8a* and *8b*. Below $\frac{1}{2}f_{\mathrm{T}}$ the second term is dominant in the denominator of (7), since $h_{\mathrm{fe}} + 1$ is still $\gg 1$. Between $\frac{1}{2}f_{\mathrm{T}}$

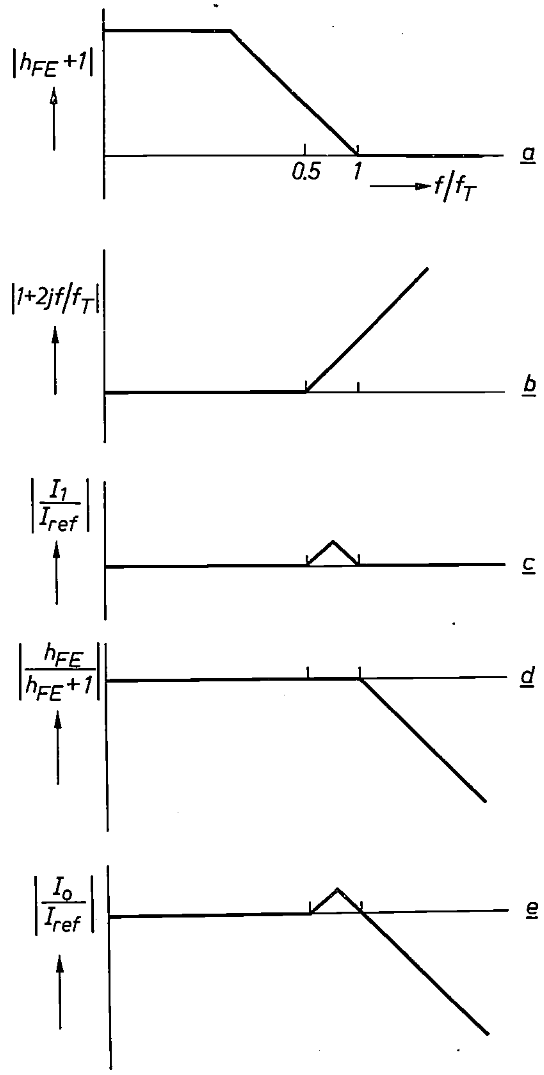


Fig. 8. Derivation of the frequency characteristic of the three-transistor current source represented by curve *b* in fig. 5, by means of Bode diagrams. Combining diagrams (*a*) to (*d*) gives diagram (*e*), which corresponds to the frequency characteristic.

and f_T both terms become equal to each other, and above this value the first term predominates. The Bode diagram of I_1/I_{ref} is therefore as shown in fig. 8*c*. What we are interested in, however, is not I_1/I_{ref} but I_o/I_{ref}, which is given by

$$I_o/I_{ref} = \frac{h_{fe}}{h_{fe}+1}\, I_1/I_{ref}.$$

The Bode diagram of $h_{fe}/(h_{fe}+1)$ is shown in fig. 8*d*; by adding fig. 8*c* and fig. 8*d* we get fig. 8*e*, which does indeed correspond to curve *b* in fig. 5.

The breakdown voltage of the current source with three transistors is about 0.8 $V_{(BR)CBO}$, just as in the case of the current source with two transistors. This can clearly be seen from the characteristics in fig. 6*b*, which, compared with fig. 6*a*, also show the higher output impedance of the circuit with three transistors.

Output current unequal to reference current

We have already seen that by giving reference diode *1* and transistor *2* of the current source in fig. 4 dissimilar emitter areas we are free to choose the ratio between reference and output current. This applies only within certain limits; ratios that are too high give large emitter areas and low cut-off frequencies.

A fixed ratio between output and reference current can also be obtained by incorporating a resistor in the emitter lead of diode or transistor. For the circuit shown in *fig. 9* it can be shown that

$$I_oR = (kT/e) \ln (I_{ref}/I_o). \qquad (8)$$

By adding relatively small resistances sources can be made that supply a very low current. If for example we have $I_{ref} = 100$ μA, then for an output current of $I_o = 10$ μA, (8) shows that R should have a value of 6 kΩ.

The use of emitters of different area does not upset the symmetry of the circuit, because the quantity eV_{BE}/kT is not dependent on the emitter area. The symmetry is however upset by the introduction of an emitter resistance, and the circuit no longer retains its basic insensitivity to temperature fluctuations. The temperature effect caused by the resistance is sometimes used for compensating other temperature effects.

It will be evident that the feedback current source of fig. 7 can also be designed with emitters of different area or with an emitter resistance. In that case, however, the expression for the output current again contains the difference term of the order $1/h_{FE}$ which did not occur in equation (5), since now all base currents no longer have the same magnitude.

Reference-current source

A reference-current source, which is an important element in the circuits dealt with here, is often obtained by deriving the current from the supply voltage via a large resistance.

A much more attractive current source for integration, which has only a small resistance and which is moreover independent of the supply voltage, can be obtained by combining two of the current sources described above (*fig. 10*). The resistor R serves for adjusting the output current.

The upper current source, consisting of *PNP* transistors, causes identical currents I_o to flow in both branches. To make the currents identical in the lower current source when there is a resistor R, transistor *2* is given a larger emitter area than transistor *1*. Equation (8) shows that R and the ratio p of the areas should then satisfy the condition:

$$I_oR = (kT/e) \ln p. \qquad (9)$$

For a given R and p the value of I_o is then fixed.

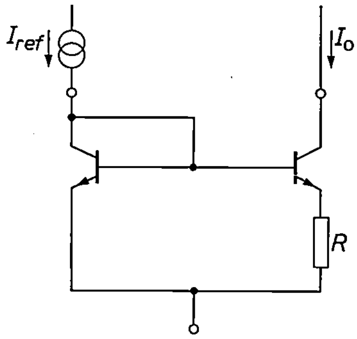


Fig. 9. Controlled current source in which the reference current and output current are unequal.

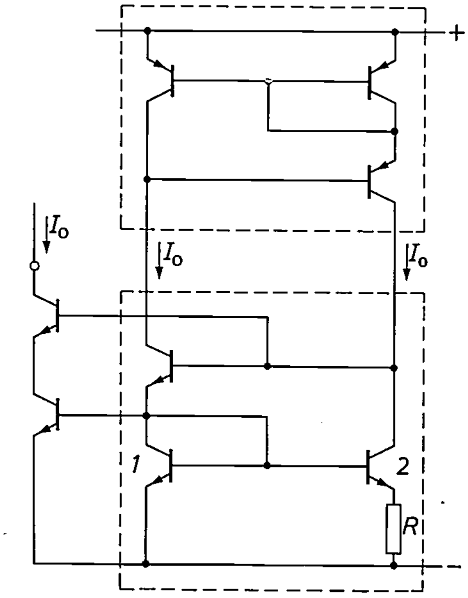


Fig. 10. Reference-current source which is independent of the supply voltage. A current source with three *PNP* transistors (*upper circuit inside dashed lines*) is connected with a current source using three *NPN* transistors (*lower circuit in dashed lines*); the output current of the one current source is the reference current for the other. An extra pair of transistors provides a constant current I_o.

[6] H. W. Bode, Network analysis and feedback amplifier design, Van Nostrand, Princeton, N.J., 1959.

If it is desired for example to bias the circuit to give a current of 100 μA, equation (9) indicates that for $p = 2$ we should choose $R = 180\ \Omega$.

The resistor does not have to be included in the current source with the unequal emitter areas. If it is not necessary to have identical currents in both branches, the resistor could be incorporated in the *PNP* current source.

Two extra transistors can be connected to the circuit to give a high-grade current source that could be used in a differential amplifier. Several such pairs can be connected to the circuit, enabling it to act as a common reference for a number of current sources, which can be necessary in a large amplifier circuit.

From equation (9) we see that I_0 is independent of the supply voltage but proportional to the absolute temperature, indicating that there might even be an application as a thermometer. This feature can be utilized for making the gain of a differential amplifier independent of temperature. The gain here is determined by the transconductance S of the transistors, which is given by $S = eI/kT$, where I is the current at the operating point. If I is obtained from the current source described here, we see from equation (9) that the transconductance and hence the gain is independent of temperature [7].

Amplifier circuits

The amplifier circuits we shall deal with are all based on the principle of the differential amplifier with a current source in the common emitter lead (fig. 3*b*). In the differential amplifier the input signal is the difference between the base voltages or currents of the two transistors. The difference between the collector currents is the output signal. Consequently, temperature and supply-voltage fluctuations, which cause the same variation in both collector currents, cancel out in the output signal, and the same is true for signals that appear in the same phase at the two bases [4] [5].

If a differential amplifier is followed by an output stage, the coupling is usually via a single-ended output; if it is followed by a second differential amplifier stage, then the coupling is balanced.

Differential amplifiers

When a controlled-current source is used as the collector load of a differential amplifier, the result is a circuit like the one shown in *fig. 11*. A voltage V_i between the input terminals gives rise to difference currents $\Delta I = \frac{1}{2}SV_\mathrm{i}$. The controlled current source causes the difference currents to be added at the output, so that a current SV_i appears there across an impedance which is equal to the output impedances of the differential amplifier and of the current source in parallel.

Since there are no collector resistances there is very little decrease in the collector voltage at a steep increase in the current I. This enables the circuit to handle signals appearing in phase at the two inputs even when the signals are almost equal to the supply voltage. A low impedance, e.g. a transistor, should be used for taking off the difference current $2\Delta I$.

If a circuit with a higher output impedance than the circuit in fig. 11 is desired, the transistors in the differential amplifier can each be replaced by a cascode configuration, which increases the output impedance of the differential amplifier h_FE times. In this case a current source with three transistors must be used, which as we saw earlier gives an output impedance $\frac{1}{2}h_\mathrm{FE}$ times higher than that of a single transistor. If all the transistors have the same parameters, the result is an output impedance $\frac{1}{4}h_\mathrm{FE}$ times that of a single transistor. The amplification factor of the circuit thus obtained, which is the product of this output impedance and the transconductance S, may in practice be as high as 10^5. The same high amplification factor can also be achieved with two amplifier stages connected in series, but the circuit indicated here has the advantage that only a single time constant is significant at high frequencies, so that any negative feedback present will not give rise to instability.

An example of how a balanced coupling can be made between two differential amplifiers is shown in *fig. 12*. The load for the collectors of the input stage is a double

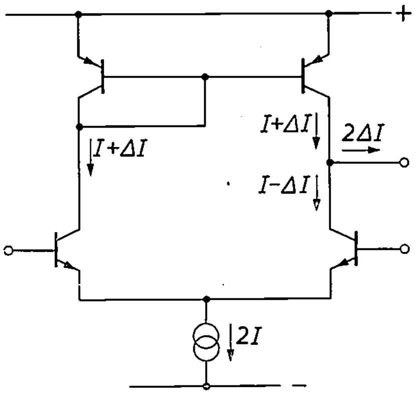


Fig. 11. Differential amplifier with single-ended output.

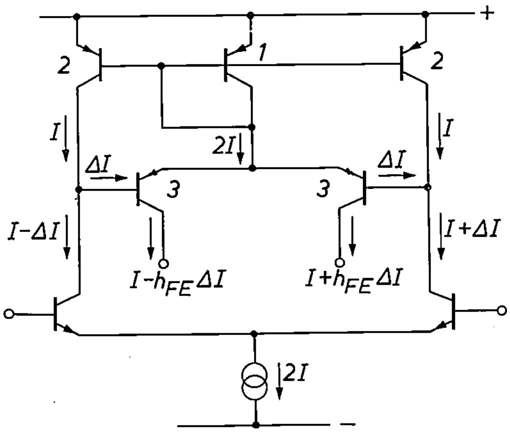


Fig. 12. Differential amplifier with balanced coupling.

controlled current source, which has two output transistors *2* and one combined reference diode *1* with an emitter area twice as large. The current in diode *1* is the sum current, which is independent of the drive and equal to $2I$, and since this diode keeps the base-emitter voltages of transistors *2* constant current I also flows through each transistor *2*. The difference currents $\pm \Delta I$ must therefore flow through the transistors *3*, and are therefore multiplied by the current gain h_{FE}.

The circuit can be extended similarly with a third amplifier stage by incorporating another such balanced circuit, now with *NPN* transistors, in the collector leads of the transistors *3*.

The collector voltage of the input transistors is equal to the supply voltage less two diode voltages. A positive in-phase signal at both inputs approximately equal to this voltage can be applied. The same applies to negative in-phase signals when the current source in fig. 7 is incorporated in the common emitter lead of the input transistors. This means that the differential amplifier with balanced coupling as described here is capable of handling exceptionally large in-phase signals.

Input circuits

After these general examples of differential amplifier stages we shall now examine the particular requirements which a differential amplifier must satisfy if it is to form the input stage of an integrated circuit.

When there is a d.c. coupling to an external circuit, both the input signal and the d.c. bias for the bases of both transistors have to be supplied from outside. In many applications it is desirable that the d.c. base currents should be small. This has led to the development of differential amplifiers with a low input current. A familiar example is the Darlington amplifier, a differential amplifier with series-connected emitter followers (*fig. 13*).

This configuration has a number of serious drawbacks. One is the considerable voltage drift, which is particularly undesirable in an input circuit. The current gain of the two inner transistors may differ appreciably, resulting in unequal currents through the outer transistors and thus causing an unbalance that leads to a marked voltage drift. Another drawback is that the output impedance of the outer transistors and the input capacitance of the inner transistors introduces an *RC* time constant which may be fairly high, since at the low emitter currents that flow the output impedance of the outer transistors is high.

The unbalance and the extra *RC* time constant can be reduced by taking an extra current from the outer pair (*fig. 14*). This minimizes the effect of the unequal base currents of the inner pair. The *RC* time constant becomes smaller because the emitter output impedance is reduced. The extra currents do not have to be high (e.g. 10 μA), but of course they cancel out to some extent the advantage of the Darlington amplifier.

The circuit in fig. 14 is not so attractive for integration because it contains fairly high resistances. In the version shown in *fig. 15* resistances ten times smaller can be used, and this circuit is therefore much more suitable for integration.

A small d.c. bias on the base is not the only requirement for the input stage of a differential amplifier. In some cases it may be necessary to stabilize the collector currents to keep the transconductance of the input transistors and the dissipation constant.

Fig. 16 shows a circuit that meets this requirement. The emitter leads of the differential amplifier in this circuit incorporate two *PNP* transistors *3* which perform three functions simultaneously. In the first place

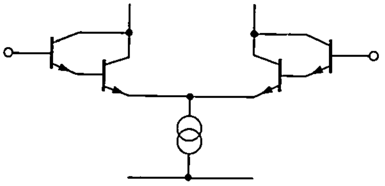


Fig. 13. Differential amplifiers with an emitter follower at the input to reduce the input currents (Darlington amplifier).

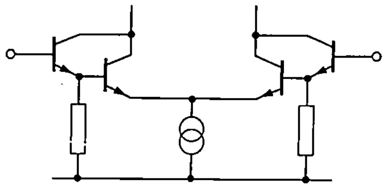


Fig. 14. Circuit as in fig. 13, in which extra currents are taken from the emitter followers to minimize the influence of the unequal current gain of the inner transistors.

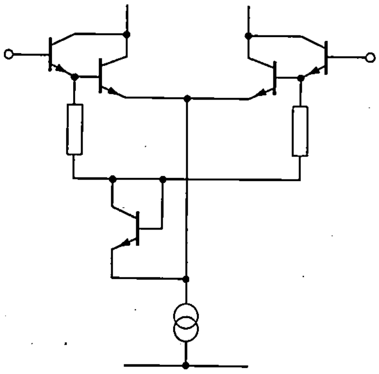


Fig. 15. Version of the circuit in fig. 14 with smaller resistors.

[7] A. J. W. M. van Overbeek and W. A. J. M. Zwijsen, Tunable integrated circuits, Philips tech. Rev. 27, 264, 1966.

they raise the breakdown voltage between the two input terminals to about 30 V; with the emitters connected the breakdown voltage would be equal to the Zener voltage of one of the base-emitter diodes, i.e. about 6 V. In the second place they give an output with a high internal impedance at the emitter end of the differential amplifier, i.e. at a d.c. voltage level close to that of the negative supply voltage. The advantage of this is that with an extra *NPN* transistor (shown dashed in fig. 16) an output is obtained at a d.c. voltage level midway between the supply voltages, i.e. at earth potential — a facility that is often required for the following stages. With an output at the collector end as in fig. 11 it is also possible to include an extra transistor to obtain an output at earth potential, but in this case the extra transistor must be of the *PNP* type, which has a lower cut-off frequency. Although the *PNP* transistors *3* in fig. 16 have a lower cut-off frequency, this does not matter so much since they are incorporated in a common-base configuration.

In the third place these same transistors help to stabilize the collector currents of the differential amplifier. This is because they form part of a controlled current source as in fig. 7, which also includes the transistors *1* and *2*. Unlike the configuration of fig. 7, there are two transistors *3*. The rather variable magnitude of the current gain $h_{\rm FE}$ of the *PNP* transistors is no drawback in this application, as eq. (5) shows, provided that $h_{\rm FE}$ is greater than about 5.

In this circuit, as in fig. 11, the difference signal is taken off by means of a current source consisting of two transistors. Here again, very large in-phase signals are permissible at both inputs.

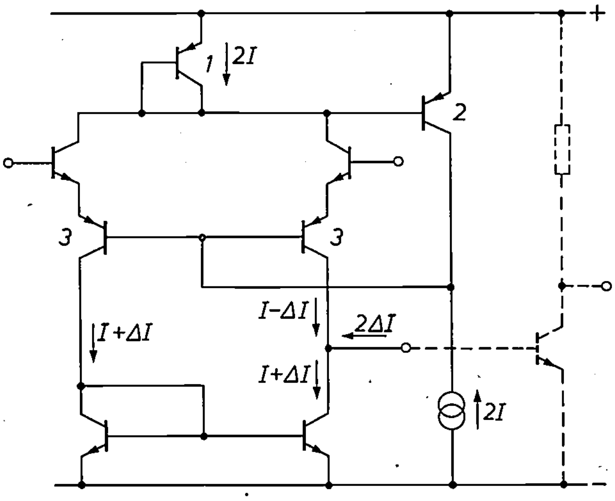


Fig. 16. Input differential amplifier in which the current source formed by transistors *1*, *2* and *3* keeps the collector currents constant. An *NPN* output transistor (*dashed lines*) is used to bring the d.c. level of the output signal to earth potential (midway between the positive and negative supply voltages).

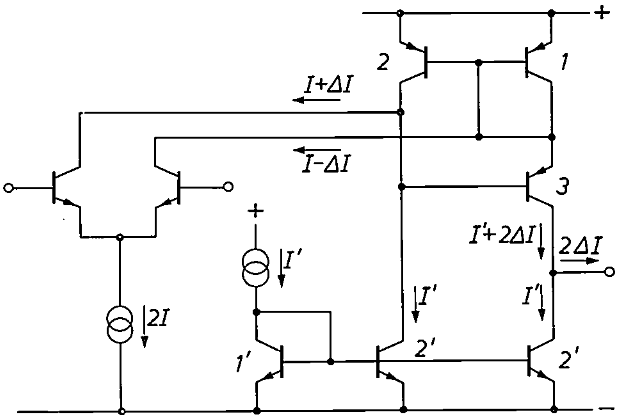


Fig. 17. Differential amplifier with d.c. level restorer, which brings the d.c. voltage level at the output midway between the positive and negative supply voltages, i.e. to earth potential.

D.c. level restorer

In the foregoing we have seen how the single-ended output of a differential amplifier is brought to a d.c voltage level between the positive and negative supply voltage, i.e. earth potential, by means of an extra transistor. This is desirable when this output has to be connected to an external load, whether or not via an output stage. In the circuit shown in fig. 12 we encountered a balanced output in which the amplifier stages shown, with possible extra ones, brought the difference signal to a level that lay alternately a few diode voltage levels above the negative or below the positive supply voltage. In this case a d.c. level shift is needed before the signal can be applied to an output or to an output circuit. *Fig. 17* shows a d.c. level restorer of this type. The signal is taken off this circuit by means of transistors in a common-base configuration. The voltage gain obtained with the circuit in fig. 16 is not obtained here, but on the other hand the bandwidth is greater.

The circuit consists of the two current sources *1*, *2*, *3* and *1′*, *2′*. The preceding differential amplifier is shown in simplified form on the left. The current through the source *1′*, *2′* with double output transistor *2′* is determined by the reference-current source I'; the reference current for transistor *2* of the upper current source is equal to the sum of I' and the current $I + \Delta I$ through the differential amplifier. The current through diode *1* is the same; the differential amplifier takes a fraction $I - \Delta I$ of this, and the remainder $I' + 2\,\Delta I$ flows through transistor *3*, and then divides between transistor *2′* which carries a current I', and the output.

The output impedance is equal to the parallel collector impedances of transistors *3* and *2′*. The circuit can be driven to within a few diode voltages of the supply voltage.

Class B output stage

Since a small load on the level shifter is sufficient to cause a loss of gain, it may be desirable in some cases to connect to the shifter an output stage that gives current gain only. A controlled current source can again successfully be used in such an output stage, as *fig. 18* shows. The current source here consists of the transistors *1* to *5*. A reference-current source *I* causes a constant current of about the magnitude of *I* to flow in the left-hand branch; this has the effect that the sum of the voltages across the base-emitter diodes of transistors *2* and *4* is constant.

This constant sum voltage also appears across the two base-emitter diodes of transistors *3* and *5*. This does not mean that the same current necessarily flows through these transistors. This is the case, though, when there is no output signal; a current approximately equal to *I* then flows in both transistors. However, if the voltage at the input of the circuit rises, the base-emitter voltage of transistor *5* rises with it, as does the current through this transistor. At the same time the base-emitter voltage of transistor *3* decreases by the same amount, so that a lower current flows in this transistor. Because of the exponential diode characteristic (1), an increase of the current in transistor *5* to γI ($\gamma > 1$) causes a decrease of the current in transistor *3* to I/γ. This is in fact a type of class B amplifier, but one in which the current in one branch never drops completely to zero.

If the voltage at the input of the circuit decreases, the current in transistor *5* also decreases and the current in transistors *3* and *1* rises. The base current for this is derived from the current source *I*.

The maximum output current during negative control is thus equal to the current *I* multiplied by the current gain of transistor *1*. If this output current is not sufficient, the circuit can be extended by adding a class C amplifier to it, as shown in *fig. 19*. The values of the resistors *R* are chosen so that transistors *7* and *8* do not conduct when the output current is small. When the output current is increased there comes a point at which the transistors start to conduct because the current supplied by transistors *3* and *5* increases the voltage across the resistors *R*. When this happens transistors *7* and *8* start to supply a large part of the output current.

Conclusion

We have seen from the treatment of basic and other circuits that in linear-circuits design today the trend is to adapt the circuit to the requirements of integrated-circuit technology in such a way as to obtain the optimum product. There are hardly any resistors. The circuits are as far as possible laid out symmetrically, thus minimizing temperature effects.

All this has been made possible through the successful application of the differential amplifier and a new element — a current source that can be controlled by a reference current. The introduction of this element is associated with a design approach in which current sources and current control are the dominant considerations. The voltages generated by the controlled currents are usually limited to diode voltages. A relatively low supply voltage is therefore sufficient, in spite of the stacking of transistors which characterizes linear integrated circuits today.

In this development the computer is an extremely useful tool. It is particularly useful in design calculations for determining high-frequency behaviour and the effect of spread in the elements and parasitic effects. The computer is also of great value in the drawing of masks and for circuit testing in production.

But it is the electronic engineer who will be responsible for the creation of new basic circuits. It is already

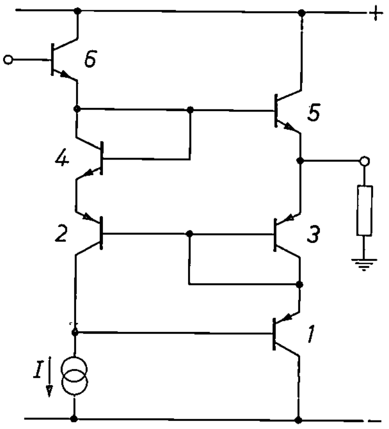


Fig. 18. Class B output stage.

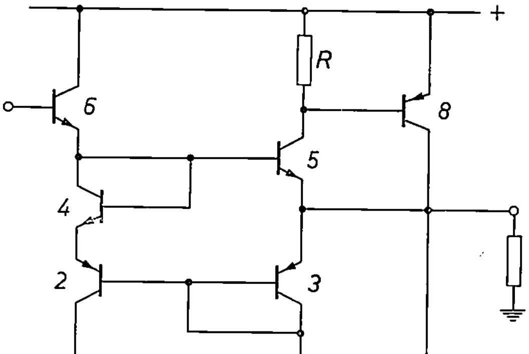


Fig. 19. Extension of the output stage in fig. 18 with a class C amplifier (transistor *7*, *8*) which only passes the peaks of the output current. This modification enables the circuit to supply higher output currents.

clear that circuit designers have accepted the challenge of integration with some enthusiasm and have found that there is much to be gained from the wealth of possibilities that it offers. Well may the layman look with some astonishment at integrated circuits, for in taking up the monolithic technique the electronic engineer — aided by the skill of the technologist — has produced circuits that are at least the equal of the traditional ones.

Summary. Integrated-circuit technology has both special capabilities and special difficulties for linear-circuit electronics. Transistors can be made that are identical and all operate at the same temperature; on the other hand, resistors and capacitors take up too much chip space and their use is to be avoided as far as possible. This is leading to a new electronics in which wide use is made of balanced circuits, e.g. differential amplifiers, and in which resistors are being replaced by current sources built up from transistors. Examples are discussed of current sources controlled by a reference current, differential amplifiers, and input and output stages. From these individual circuits complete integrated circuits such as operational amplifiers can be built up.

AVOID, a short-range high-definition radar

K. L. Fuller

Navigational aids for aircraft have been available for many years, and they have now reached such a state of perfection that aircraft can be landed blind. To exploit these aids to the full, the ground services at the airfield should be able to operate normally even in fog. With this kind of application in mind the author and his colleagues have developed a system in which established principles from radar technique are combined in an original way to produce a short-range high-definition radar. The system has been given the name AVOID (for Airfield Vehicle Obstacle Indication Device).

With the growing use of fully automatic landing systems at airfields there is an increasing need to drive vehicles on the airfield at fairly high speeds in conditions of very poor visibility. After a successful automatic landing the aircraft has to be guided from the end of the runway to the main terminal building. The "follow me" car, which normally shows the route to be followed, does not have sufficient visibility in dense fog to be able to guide in a large aircraft safely. Although the route could be indicated by means of signals on cables buried in the taxi-tracks, this does not guarantee that the route is free from obstacles. Moreover a system of this type involves considerable engineering installation work and is not very flexible. In the event of a crash landing it is obviously essential that fire tenders and ambulances should be able to reach the scene as soon as possible, without colliding with pieces of wreckage and survivors en route, and here a really effective aid to vehicle navigation in zero visibility is required.

The most practical solution to this problem appears to be a radar system with a two-metre resolution over a range from 3 to 160 metres, and scanning over a sector of 60° ahead of the vehicle. Such a radar has a resolution and short-range performance that is a great deal better than those of current radar systems. In addition it needs a rapid angular scan in order to avoid picture flicker and present a high enough information rate for moving objects to be followed accurately.

The requirements which this implies in current pulsed-radar techniques may be understood as follows. In order to be able to discriminate between two objects which differ in their distance by r, the length τ of the emitted radar pulse must be smaller than the time difference with which the echoes of these objects are observed, or, if c is the speed of propagation of the radar waves, $\tau < r/c$. For $r = 2$ m and $c = 3\times10^8$ m/s this

K. L. Fuller, B.Sc. (Eng.), is with Mullard Research Laboratories, Redhill, Surrey, England.

gives $\tau < 7$ ns. The microwave generator must therefore deliver pulses with a length of 7 ns and a power of several watts. Although not impossible, this would not be easy. To have a minimum range of 3 metres, the radar system would have to be capable of switching over from transmit to receive within about 5 ns, which, with the high receiver amplification required, would present problems.

Since the velocity of propagation of ultrasonic signals is 10^6 times smaller than for radar signals, an ultrasonic radar system appeared at first sight to offer a good solution. The pulse length can be 10^6 times larger. A drawback, however, of the much lower velocity of propagation is that the information rate is too small to produce an up-to-date picture. Secondly, the attenuation of ultrasonic radiation in air is so high that ranges in excess of 20 metres are difficult to obtain with a reasonable transmitter power. Finally the ultrasonic radar is very sensitive to interference generated by jet-engine noise.

Since a pulsed radar, whether microwave or ultrasonic, is less suitable for short-range operation, it was decided to develop a system in which range is measured by applying a linear frequency modulation to a continuous microwave signal [1].

Range measurement with frequency-modulated radar

The way in which range is measured with this system is shown in *fig. 1*. The line *a* gives the frequency of the transmitted signal as a function of time. The frequency of this signal increases linearly with time from a frequency f_0 to a value f_m, then decreases at the same rate to f_0, after which the frequency rises again, and so on. The frequency of the return signals will depend on time in the same way, but the phase of the modulation with respect to that of the transmitted signal will be displaced, and the phase shift will be greater the more distant the target (lines *b* and *c*). If these return signals from the targets are mixed with a sample of the transmitter output, a difference or beat frequency is obtained which is a measure of the target distance. The more distant the target the higher the difference frequency (*b′* and *c′* in fig. 1) becomes. In practice there will be a whole spectrum of difference frequencies. These frequencies will momentarily go to zero and return to their original value at the turn-round points on the main frequency sweep. This deviation is shorter, and hence less disturbing, the longer the period of the frequency modulation compared with the delay time corresponding to a target at maximum range.

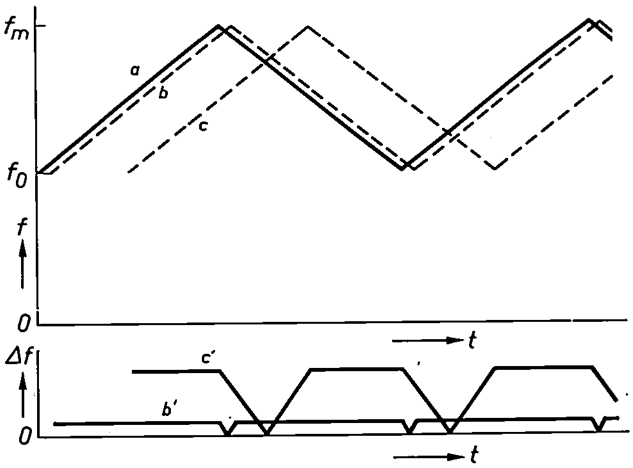


Fig. 1. Principle of range measurement with frequency-modulated radar. Curve *a* gives the frequency of the emitted signal as a function of time. Reflections from objects have the same modulation, but with a time delay proportional to the distance from the reflecting object (curves *b* and *c*). When these reflections are mixed with a sample of the transmitted signal, signals at the difference frequencies are obtained (*b′* and *c′*). The difference frequency increases with the distance of the object.

In order to make the most efficient use of the return signals it would be necessary to separate the associated difference frequencies with a large number of filters, whose pass-bands correspond to the range intervals we want to separate. The energy passed through each of the filters would then have to be integrated over a time equal to the time that the target spends in one range element. The complexity of a bank of filters is considerable, however, and it was therefore decided instead to use a single swept superheterodyne filter. This filter periodically scans the spectrum of difference frequencies and converts the parallel returned information into a serial range scan. The resultant loss of information has little effect on the performance of a short-range system of this type.

In order to have good range resolution it is necessary to have a very linear frequency sweep. For example, if it is required to resolve to one part in a hundred of the maximum range, the linearity of the sweep must be approximately 1%.

In a conventional pulse radar the range scan rate is determined by the velocity of propagation, but in the "AVOID" system any rate convenient to the system can be used, because it now depends on the tuning of the superheterodyne filter mentioned above. If the range is scanned from minimum to maximum and back again in a triangular form, and at the same time the aerial is slowly scanned in azimuth, the picture will be built up as shown in *fig. 2a*.

If it is desired to scan a picture built up in this way at a frequency of 50 Hz, the scanning beam will have to rotate at the same frequency. As it is very difficult to make mechanically scanned aerials rotate at such a speed, an electronically scanned beam must be used. There are advantages of simplification in scanning the beam by using a frequency-scanned aerial in conjunction with the same frequency modulation that is used for range measurement. If we compare the times needed for one azimuthal scan and for one range scan, it turns out that the frequency modulation can serve a dual purpose only when the azimuthal scan of the aerial is faster than the range scan of the superheterodyne filter. The picture is then built up as shown in fig. 2*b*. For one picture there are 80 sweeps of the beam so a 50 Hz picture frequency requires a beam frequency of 4 kHz.

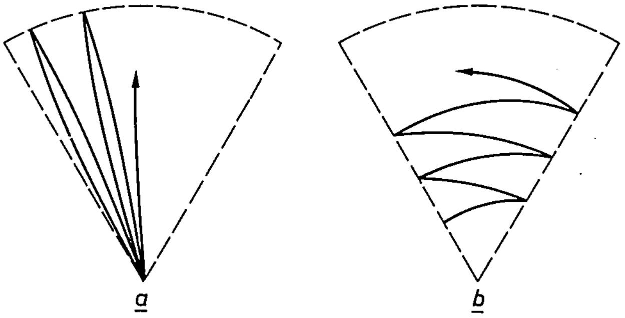


Fig. 2. Alternative ways of producing the radar display. *a*) Range scan faster than angular scan. *b*) Angular scan faster than range scan.

Speed of azimuth scan and range scan

In a system where the azimuth scan is derived from the frequency modulation of the signal by using a frequency-scanned antenna, each beamwidth of the total scan range must have a separate frequency bandwidth. This individual bandwidth is determined by that needed to give the desired range resolution. If the range resolution is r and the velocity of propagation of the radar signal is c then a range element corresponds to a time difference $2r/c$ and the bandwidth to carry the information for one range element is $b = c/2r$. For a horizontal beamwidth θ and a total angular scan ϕ the number of azimuth elements is $N = \phi/\theta$. The total bandwidth needed for one radar picture is now $B = N.b = (\phi/\theta)(c/2r)$. Now azimuth beamwidth is chosen so that angular and range resolution are equal at $\frac{1}{3}$ of the maximum range R. Hence $\frac{1}{3}R \tan\theta \approx \frac{1}{3}R\,\theta = r$, and the total required bandwidth $B = R\phi/3r \times c/2r = cR\phi/6r^2$. If now our frequency modulation is linear with time and can be written as $f = f_0 + f't$, then the time t_A needed for one complete azimuth scan is:

$$t_A = \frac{B}{f'} = \frac{cR\phi}{6f'r^2}.$$

This time will be compared to the time needed for one range scan.

The time difference for a reflection at minimum range r_0 is $\Delta t = 2r_0/c$ and with the frequency modulation as above this results in a difference frequency in the detector circuit equal to $f'\Delta t = f' \cdot 2r_0/c$.

For this frequency to be measured, at least two complete cycles must be observed. This takes a time $\tau \geqslant 2 \times c/(f' 2r_0) = c/(f'r_0)$. As the range scan is linear with time this time is available for each range element. The total number of range elements is $M = (R - r_0)/r \approx R/r$ since $R \gg r_0$, and consequently the total time t_R for one range scan is:

$$t_R = M\tau \geqslant \frac{R}{r}\frac{c}{f'r_0}.$$

There are now two possibilities:

a) Range scan faster than angular scan as in a conventional radar (fig. 2*a*). There must be N range scans for one angular scan: $Nt_R = t_A$ or:

$$N\frac{Rc}{f'rr_0} \leqslant \frac{cR\phi}{6f'r^2}.$$

This results in the relation:

$$2R \leqslant r_0$$

which is obviously not possible.

b) Angular scan faster than range scan (fig. 2*b*). Now there are M angular scans for one range scan, or $Mt_A = t_R$. With the expressions for t_A and t_R derived above this results in the equation:

$$\frac{R}{r} \times \frac{cR\phi}{6f'r^2} \geqslant \frac{Rc}{f'rr_0}.$$

This relation simplifies to $r_0 \geqslant 2r/N$, which is readily practicable.

The aerial

A familiar radar aerial system for varying the direction of the transmitted beam electronically is based on the interference principle. In this system the aerial does not consist of a single emitting element but of a number of elements in a row. Interference of the radiation emitted by these elements produces a beam in a direction which is determined by the position of the radiating elements and the phase relationship between them [2]. By making use of suitable frequency-dependent phase shifters it is then possible to use the same frequency sweep as that already being employed for range measurement to produce the angular scan.

For our purposes the simplest form of such an aerial is a piece of rectangular waveguide for 3 cm waves, with regularly spaced holes cut in the broad face. Microwave energy is fed in at one end of the aerial; as the energy passes through the waveguide a fraction is emitted through each hole. To ensure that the same power is radiated from all the holes, the diameter of the holes, starting at the input end of the waveguide, must steadily increase. The remaining energy at the other end of the guide is absorbed by a matched load. The microwave energy propagates in the waveguide in the form of an electromagnetic wave pattern. This has a wavelength λ_g greater than the wavelength λ_0 of electromagnetic radiation with the same frequency in free space. If the spacing of the holes in the waveguide is equal to the λ_g at a particular frequency, the power radiated from the holes will be in phase and the aerial will transmit a beam

[1] The design, construction and testing of the system and the aerial were carried out by K. Holford and A. J. Lambell. Much of the work was supported by the M.E.L. Equipment Company Ltd., Manor Royal, Crawley, Sussex, England.

[2] A. Meyer, Philips tech. Rev. 31, 2, 1970 (No. 1).

perpendicular to the waveguide. A change in the feed frequency causes a change of the value of λ_g, so that the holes now radiate with a certain phase difference between them, and the beam accordingly deviates from the perpendicular.

In the situation described above, where the spacing of the holes is greater than λ_0, the phase of the radiation from the holes being the same, side lobes will appear beside the main lobe broadside to the aerial. Assuming for simplicity that the aerial consists of only two elements with a spacing l, then the condition for the appearance of a beam at an angle ϕ to the normal on the aerial is $l \sin \phi = n\lambda_0$ ($n = 0, \pm 1, \pm 2, \ldots$). If $l > \lambda_0$ this equation will have real solutions not only for $n = 0$, which corresponds to the main lobe, but also for $n \neq 0$. These are the unwanted side lobes in the radiation pattern of the aerial. We can get around this difficulty by filling the waveguide with a dielectric, thereby making λ_g smaller and thus reducing the hole spacing. If the dielectric constant is sufficiently high, it is possible to make λ_g smaller than λ_0, so that $l \sin \phi = n\lambda_0$ has only one real solution for $n = 0$, the side lobes having been eliminated.

By filling the waveguide with a material of $\varepsilon_r = 2.54$, the condition $\lambda_g < \lambda_0$ was obtained for the whole frequency range of 8-11 GHz. The high dielectric constant of the material in the waveguide causes a decrease in the intensity of the radiation through the holes. To obtain better radiation from the holes the outside of the waveguide must also be covered with a layer of dielectric material.

With an aerial having more than two elements the situation is rather more complicated, and even when the spacing of the elements is smaller than λ_0 there will still be side lobes although their intensity will be much lower than in the case of $\lambda_g > \lambda_0$. The situation can be improved by applying an amplitude taper causing the elements to radiate with unequal intensity [3]. Since the radiation intensity from a hole depends on its size, this can be achieved by tapering the diameter of the holes towards the end of the aerial.

There was already a reason, however, for increasing the diameter of the holes towards the end of the aerial. These two arguments resulted in a particular pattern for the diameter of the holes along the waveguide. With a total length of 1.5 metres the aerial has about 80 holes. If we were to give each hole the optimum diameter this would mean having to use 80 different drills, most of which would be of non-standard sizes. It is found in practice that if the 80 holes are divided into ten groups, each corresponding to a standard drill size, the aerial works almost as well as it does when every hole has the calculated diameter. This of course considerably simplifies the construction of the aerial.

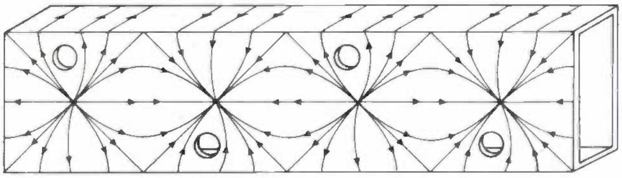


Fig. 3. Current pattern in the wall of the waveguide that forms the aerial. The holes disturb this pattern and hence radiate microwave energy. As the current patterns around the holes on both sides of the centre line are identical, both rows of holes will radiate in phase.

In the situation where the beam is broadside to the aerial, not only is the radiation from all holes in phase, but so are the minor reflections caused by the holes in the waveguide. This results in a large standing wave in the guide, which reduces the efficiency of the aerial. In all other beam attitudes the reflections are out of phase and their vector sum is so small that no troublesome standing wave arises. The aerial is therefore designed so that it scans from 5° to 65° to the normal through the frequency sweep from 11 to 8 GHz needed for range determination. The aerial must then be mounted at an angle of 35° to the front of the vehicle.

The currents in the wall of the waveguide follow the pattern given in *fig. 3*. The holes in the broad face of the waveguide will have to be offset from the centre-line of the face for maximum radiation. A second row of holes can also be made on the opposite side of the centre-line,

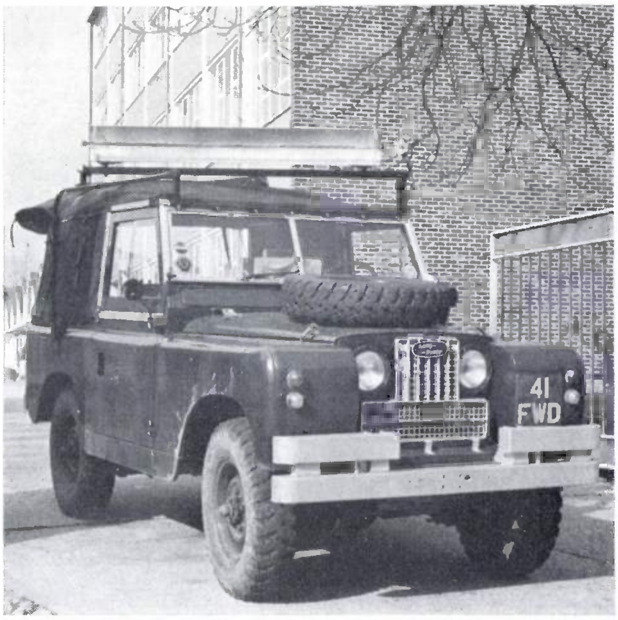


Fig. 4. The aerial mounted on top of a vehicle. The waveguide in the focal line of the parabolic cylindrical reflector can clearly be seen. The energy is fed in at the right and the fraction not radiated is absorbed in a load at the left. The portable television receiver for the display can be seen next to the steering wheel.

[3] See fig. 15 in Meyer's article [2].

each displaced by half the spacing with respect to the other row. The two rows of holes then radiate in the same phase and reinforce each other, resulting in an appreciable gain in aerial efficiency. To limit the vertical beamwidth, the waveguide is mounted in the focal line of a parabolic cylindrical reflector. The holes in the waveguide radiate backwards and the radiation is reflected as a narrow beam in the forward direction. The main lobe of this aerial has a vertical width of 10°. This would not be sufficient to cover the range of 3-160 m in front of a vehicle if the aerial was mounted at about 2 m above ground level, so the aerial is mounted in such a way that the distant part of the total range is covered by the main lobe, and the nearby part by vertical side lobes. The reduced gain of these side lobes is compensated for by the fact that nearby objects give stronger reflections. The aerial mounted on a vehicle is shown in *fig. 4*.

The display

The distance measured by a radar system is usually plotted on the display as a radius vector in the corresponding direction. In this way a plan of the surroundings is generated (*fig. 5a*). It is, however, possible to plot the distance vertically on the display screen while the azimuth is used as the horizontal coordinate. The scene of fig. 5*a* then appears as shown in fig. 5*b*; the picture now approximates to a perspective view. A slight correction to the range scan is sufficient to obtain an almost correct presentation. This type of display is easy to produce since the two signals that are a measure of angle and distance are already present in the circuitry of the system.

It would be of great advantage if a perspective view could be superimposed on the scene viewed through the windscreen. This would be very expensive to arrange, however, and true matching of scene and display would be possible for only one position of the driver's head. It was therefore considered preferable to produce a display on a cathode-ray tube which the driver can look at it by glancing slightly to one side.

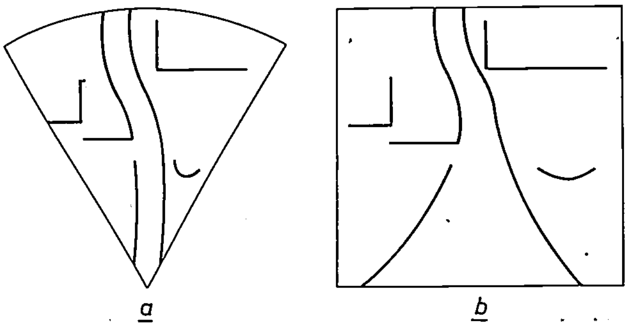


Fig. 5. The radar display. *a*) The conventional display in which the range of an object is plotted as a radius vector in the corresponding direction. This gives a plan of the surroundings. *b*) The type of display used in the AVOID system, in which the range of an object is plotted as a vertical coordinate, and the azimuth as the horizontal coordinate. A perspective view of the surroundings is obtained in this way.

It may happen that large metal objects reflect so strongly that the radar beam is reflected more than once between the object and the front of the radar vehicle. Every reflection is displayed as an object at a multiple of the range of the original object.

Also, strong reflections may overload the receiver and thus give rise to higher harmonics. These can give rise to spurious reflections, displayed at multiples of the original range. As all these spurious reflections are displayed as objects beyond the real object, they present no direct problem to the driver of the vehicle.

The experimental system

A block diagram of the AVOID radar system is given in *fig. 6*. The backward-wave oscillator Osc_1 is frequency-modulated over the range 8-11 GHz by modulation of its supply voltage. For this purpose the azi-

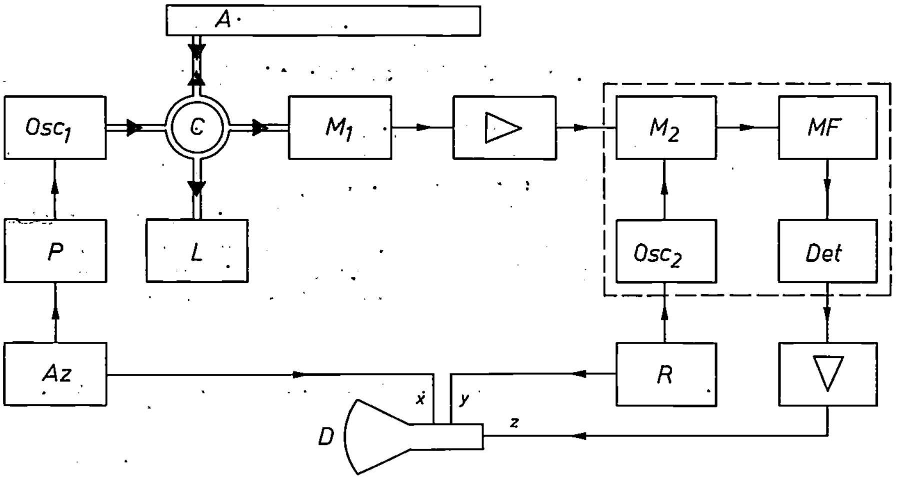


Fig. 6. Block diagram of the AVOID system. Osc_1 backward-wave oscillator. *P* power supply for the oscillator. *Az* triangular-wave generator for modulating the oscillator. *C* circulator. M_1 mixer, in which the reflected signal mixes with a sample of the oscillator signal that appears as "leakage" across the circulator. *A* aerial. The fourth port of the circulator is terminated by the load *L*. Osc_2 local oscillator, frequency-modulated by the triangular wave generated by *R*. M_2 mixer. *MF*. i.f. amplifier. *Det* detector. *D* display, with horizontal deflection *x*, vertical deflection *y* and intensity modulation *z*.

muth sweep generator *Az* produces a periodic triangular control signal with a frequency of 2.5 kHz. The relationship between supply voltage and frequency of a backward-wave oscillator is exponential, and therefore the power supply has a correction circuit to produce a linear frequency sweep. The power output from the oscillator is fed via a broadband circulator *C* to the aerial, and a small part goes directly to the detector as a reference signal for range measurement. The total transmitted power is about 10 mW, resulting in a mean power of 3 μW in any 1 MHz bandwidth. No frequency allocation is available at present for this broadband transmission but, in view of the very low power, its use for emergencies might be permitted.

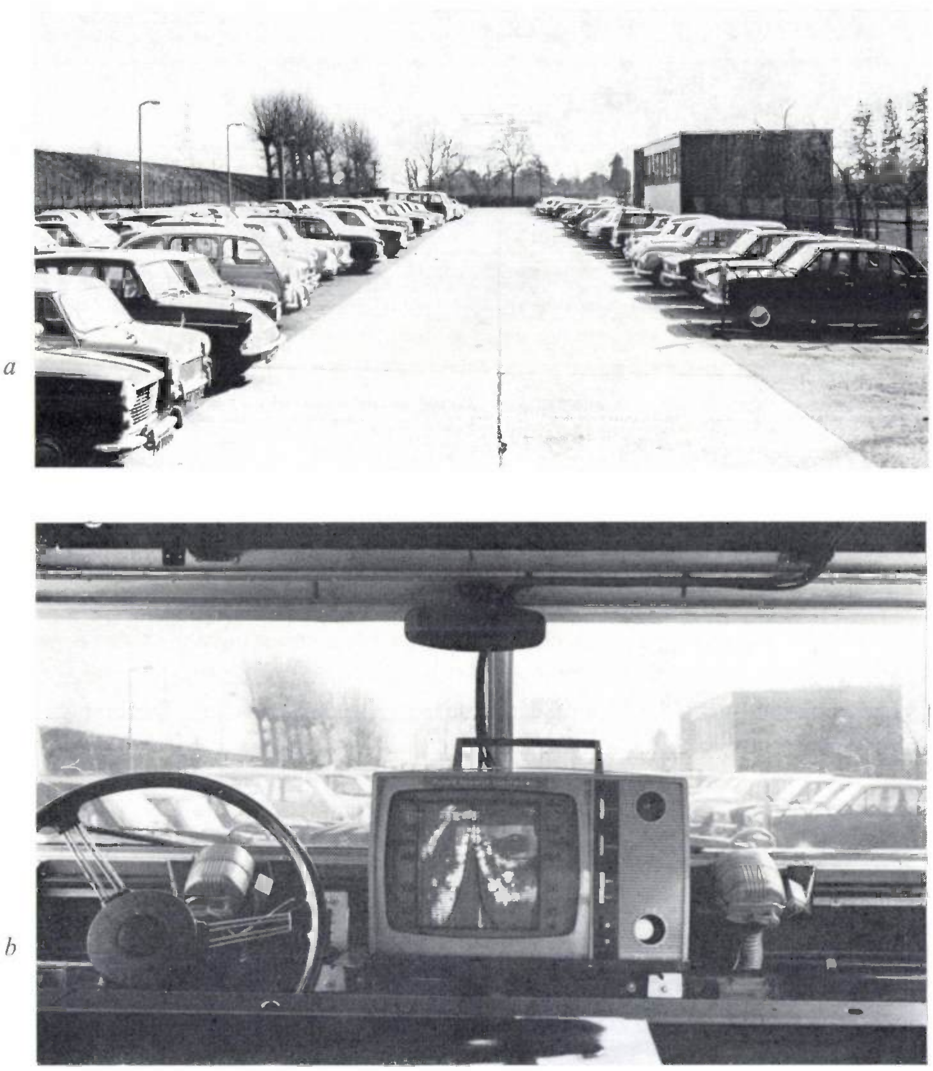


Fig. 7. *a*) View of a car park. *b*) The corresponding radar picture on the display in the vehicle. In front of the display screen there is a graticule giving a range scale and two converging lines indicating the free space required in front of the vehicle.

Return signals from targets go from the aerial *A* through the circulator into a mixer stage M_1, where they are mixed with the reference signal. The resultant difference frequencies are amplified, the high frequencies corresponding to long-range targets being amplified more than the low difference frequencies.

The next four blocks in the diagram comprise the swept-superheterodyne filter, which scans with a triangular waveform through the frequency spectrum, which is in fact a range spectrum. After detection and amplification, the signal is used for brightness modulation of the display *D*. The horizontal and vertical deflection signals for the cathode-ray tube are obtained from the azimuth and range sweep generators *Az* and *R*.

The frequency of the azimuth generator is 2.5 kHz and that of the range sweep generator 50 Hz. The complete picture scan rate is 50 Hz. The target resolution is 2° in azimuth over a 60° scan, so that there are thus 30 picture elements in the azimuth direction. The range resolution is 2 metres over a maximum range of 160 metres, i.e. 80 elements, so that the complete picture is built up from 2400 elements. The same system can also produce a picture rate of 25 per second, which doubles the number of lines on the screen without changing the resolution. The effect is to produce a picture which appears to have a better definition, but at the expense of some flicker. The display system used for our experiments consists of a modified portable television receiver. *Fig. 7* shows the driver's view ahead.

An extensive programme of trials with the system showed that the driver needs a short period of familiarization, after which the radar picture is found very useful. Blind driving, with the windscreen completely obscured, was tried in two locations, a fenced empty carpark and a deserted airfield. Although the driver completely lost his sense of direction, having no visual or compass information, the vehicle did not collide with any of the numerous obstacles, and it was easy to drive through a route marked by corner reflectors. The title photograph shows another blind-driving demonstration.

Summary. To allow vehicles to be driven on an airfield under conditions of poor visibility a radar system has been developed with a range of 3-160 m, 2 m resolution and a 50 Hz picture-frequency display. Range measurements are made with a frequency-modulated c.w. signal. Electronic angular scan is obtained by using an aerial consisting of a length of waveguide with a large number (about 80) of radiating holes. In this way the use of a slow, mechanically scanned aerial is avoided and a picture frequency is obtained that is adequate for a flicker-free daylight-viewing display.

By filling the waveguide with a dielectric, unwanted side lobes in the radiation pattern of the aerial are suppressed. Comparing the time needed for one azimuth scan and one range scan shows that azimuth scan has to be faster than the range scan if the same frequency modulation is to be used for both range measurement and azimuthal beam scanning.

Parametric amplifiers for a radio-astronomy interferometer

R. Davies and R. E. Pearson

Microwave parametric amplifiers can now be designed that give diode noise temperatures not far above 10 K when cooled by liquid nitrogen. For applications such as radio astronomy, these devices are more practical than masers, which are more complex in themselves and require a more elaborate cryogenic system since they are operated at liquid-helium temperatures. The article below describes a cooled parametric-amplifier system for the radio interferometer at Defford, England.

Introduction

In radio astronomy it is often important to be able to locate a faint source of radio noise and measure its angular diameter. To detect such faint sources it is clearly desirable to have a low-noise input stage for the receiver system, and the angular diameter of the source can be measured by using an interferometer arrangement. A radio interferometer has two aerials, each with its own receiver system; the diameter of the source is derived by training the aerials on it and measuring the correlation of the output noise as a function of aerial spacing [1].

The radio noise that reaches the Earth from such sources is in the microwave range of frequencies, so the radio interferometer must be a microwave system. Since the background noise from outer space is very low indeed, a great deal may be gained by using a low-noise amplifier in the system. The low-noise amplifier first used in such systems was the maser [2], which can have an effective input noise temperature [3] of about 6 K. However, although the maser has such a low noise temperature, it has two great disadvantages. It can only be used with complicated and expensive support equipment, in particular the cryogenic equipment associated with the liquid-helium cooling; and its bandwidth can never be very large. The maser introduced the microwave systems engineer to the spectacular benefits of low-noise devices, and at the same time convinced him that there would be much to be said for a simpler and perhaps broad-band device, even if the noise temperature were a little higher.

In fact, work on such a device had begun in the mid 1950s. This was the varactor-diode parametric amplifier, based on the varactor or variable-capacitance diode [4] [5]. Its reactive method of energy transfer [6] promised a low noise level and the varactor diode itself was tiny and very suitable for use in microwave circuits. Since those days there has been a good deal of progress and today the parametric amplifier has superseded the maser for almost all applications in which a low-noise microwave amplifier is required. Even with fairly simple tuned circuits the bandwidth of the parametric amplifier is better than that of the maser, and it can be increased further by using simple filter techniques [7]. The parametric amplifier can also give quite good noise performance at ordinary ambient temperatures. However, in applications where maser-like performance is required cooling is still necessary, but the temperatures required are usually not so low that liquid helium has to be used.

R. Davies, Ph.D., and R. E. Pearson, B.Sc., Grad. Inst. P., are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] This principle and its application to optical stars have been treated earlier in this journal: R. Hanbury Brown and A. Browne, The stellar interferometer at Narrabri, Australia, Philips tech. Rev. **27**, 141-159, 1966.

[2] Masers for the radio-astronomy interferometer at Defford, near Malvern, England, have been described earlier 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.

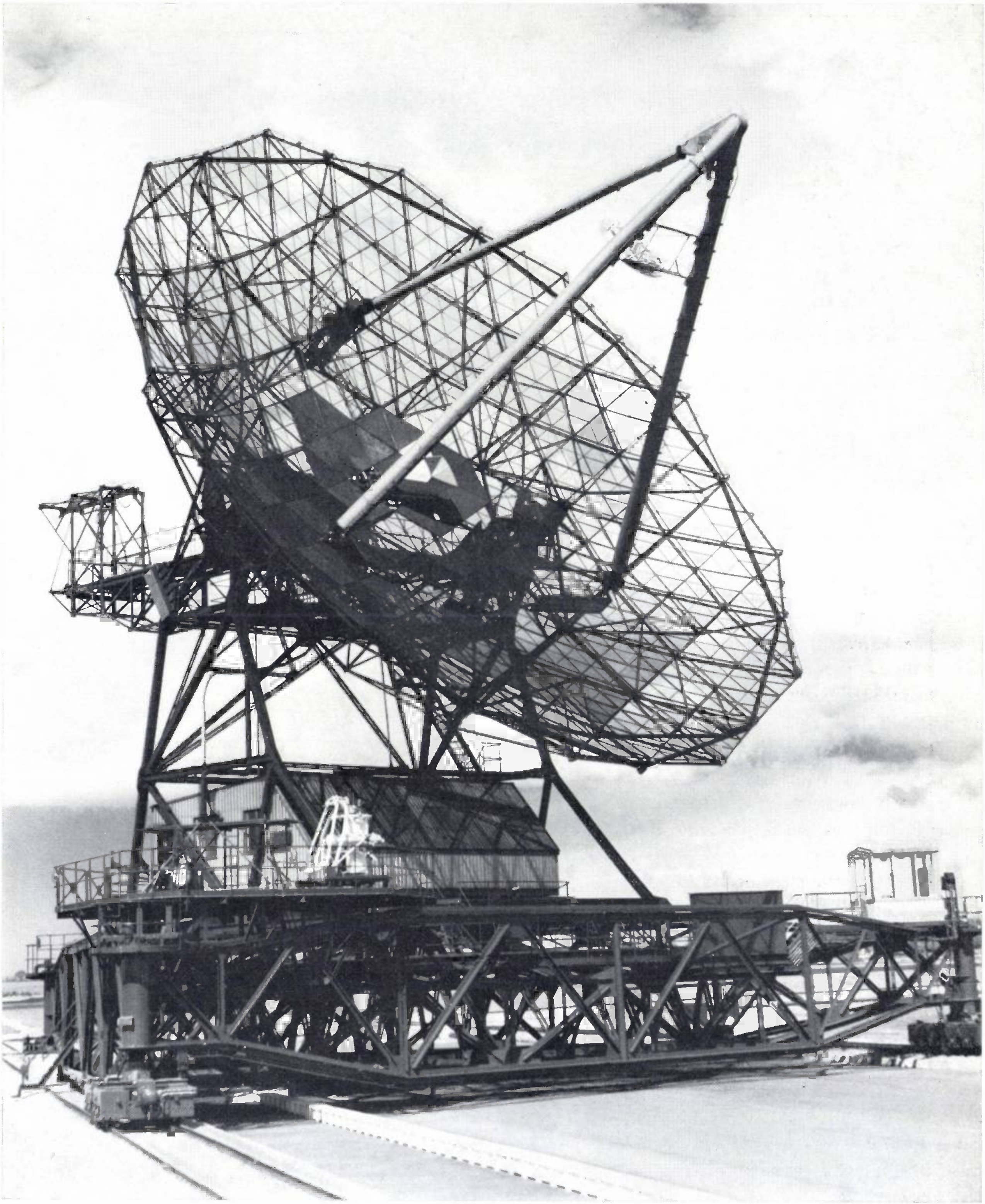
[3] The effective input noise temperature (often called the amplifier noise temperature) is the temperature at which the input termination must be held to produce an output noise power, per unit bandwidth, double that which would occur if the termination were cooled to absolute zero.

[4] H. Heffner and G. Wade, Gain, bandwidth and noise characteristics of the variable-parameter amplifier, J. appl. Phys. **29**, 1321-1331, 1958.

[5] C. S. Aitchison, Low noise parametric amplifiers, Philips tech. Rev. **28**, 204-210, 1967.

[6] One of the first to recognize the low-noise potentialities of non-linear capacitances was A. van der Ziel, then working at Philips Research Laboratories. He published the classic paper: On the mixing properties of non-linear condensers, J. appl. Phys. **19**, 999-1006, 1948.

[7] B. L. Humphreys, Characteristics of broadband parametric amplifiers using filter networks, Proc. IEE **111**, 264-274, 1964.



Copyright Controller HBMSO

Fig. 1. One of the two aerials for the radio interferometer at Defford, near Malvern, England. To vary the length and direction of the base line each aerial can be moved along its own special double railway track: the two sets of tracks intersect at an angle of 67°. The two aerials are trained on a radio source and the noise received is amplified in a separate channel for each aerial. Receiver noise is kept low by using a varactor-diode parametric amplifier as the first stage. Measuring the correlation of the output noise as a function of the aerial spacing will give the angular diameter of the noise source. The tripod structure carries the extra reflector of the Cassegrain feed.

In the present article we shall describe the parametric amplifiers that have been developed to replace the masers in the Defford radio interferometer (*fig. 1*). There are two parametric amplifiers, one for each receiver channel, and they operate at a frequency of about 2.7 GHz. The amplifiers are cooled to 80 K by liquid nitrogen to bring the effective noise temperature of the receiver below the specified value of 35 K.

On the face of it, it would seem that a microwave receiver with a noise temperature of 35 K for its first stage is going to give a much worse performance than one with a maser first stage whose noise temperature is about 6 K [2]. However, we have to remember that there are noise contributions from the parts of the system that precede the first stage. In the Defford radio interferometer the aerial itself has an effective noise temperature of 20 K and the aerial feeder system has a noise temperature of 10 K. Noise contributions after the first amplifying stage will not be significant because of the gain of the first stage, so we can take the total system noise temperature to be 36 K with the maser and 65 K with the parametric amplifier.

Now the sensitivity of the receiver system can be expressed in terms of a minimum detectable source temperature ΔT, which is equal to $T_s/\sqrt{2B\tau}$, where T_s is the receiver-system noise temperature, B is the bandwidth of the receiver and τ is the post-detector integration time. The relation shows that when the parametric amplifier is used the post-detector integration time has to be increased by just over three times to keep the minimum detectable source temperature at the same value (assuming the same bandwidth). This is a disadvantage, but one that is more than outweighed by the advantages of a system that, because it requires no liquid helium, is simpler and less expensive to run.

The parametric amplifier

Let us begin by indicating the main elements of the varactor-diode parametric amplifier used in this system. The most vital element is the *varactor diode*. This is a *P-N* junction that behaves as a capacitance that varies with applied voltage [8]. The varactor diode is driven by a local source called the *pump*, which applies a voltage at a high frequency f_p across the diode. Under these conditions the diode appears to small signals as a *time-varying capacitance*. Pumped in this way, the diode forms the link between two circuits as shown schematically in *fig. 2*, where C is the time-varying capacitance. For clarity the pump circuit is not shown. The circuit to the left is known as the signal circuit, because it is connected to the source E of the signal at frequency f_s; the right-hand circuit, connected only to the diode, is known as the *idler* circuit. The reactance X_s is provided to make the signal circuit resonant at the frequency f_s, and similarly the reactance X_i is provided to resonate the idler circuit at f_i. Band-pass filters F_s and F_i are also included to confine the signal- and idler-frequency currents to the appropriate circuits.

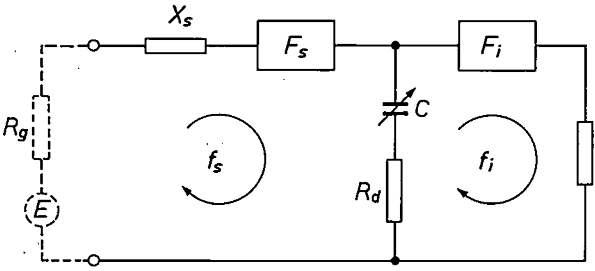


Fig. 2. Equivalent circuit of parametric amplifier. The left-hand circuit is the signal circuit and the right-hand one is the "idler" circuit. They are coupled by the varactor diode, which is driven by a pump signal to make it behave as a time-varying capacitance C. (The pump circuit is not shown.) Loss in the diode is represented by the spreading resistance R_d. The reactance X_s is provided to make the signal circuit resonant at the frequency f_s, and reactance X_i resonates the idler circuit at the frequency f_i. Band-pass filters F_s and F_i confine the currents to the appropriate circuits. Parametric amplification is achieved when $f_s + f_i = f_p$, where f_p is the pump frequency.

If the resonant frequencies and the pump frequency are so chosen that

$$f_s + f_i = f_p, \qquad (1)$$

then [4] [9] the capacitance C varying at the pump frequency is able to transfer energy from the pump to appear at the frequency f_s (or at f_i). The circuit will then function as an amplifier. The gain can be considered to arise because the circuit behaves as a negative resistance for an input signal of frequency f_s. A simple physical explanation of the action is given in the Appendix.

A circuit like the one shown in fig. 2 is not particularly suitable as it stands for use as an amplifier, since the output signal will appear at the same terminals as the input signal. It is much more convenient to include a *circulator*, which separates the incident and amplified signals, as shown in *fig. 3*. (Port *1* couples to port *2*, port *2* to port *3*, etc.). In this configuration, the negative resistance presented by the amplifier at the signal frequency gives a reflection coefficient greater than unity, and thus a gain since input and output signals are separated. The circulator also prevents the gain from being affected by variations in source impedance.

Since the method of energy transfer is based on a time variation of a reactance, which gives no thermal noise, it would appear that the amplifier should have a low input noise temperature. In fact, the diode is not quite a perfect reactance; there is a small resistive loss (the "spreading resistance") caused by the resistance of the ionic lattice to the movement of the electrons and holes. This spreading resistance is represented by R_d in the schematic diagram of fig. 2. By carrying out a theoretical analysis in which sources of resistive noise are taken into account, it can be shown [9] that for high gain (say 10 dB or more) the effective input noise temperature T_{amp} of a parametric amplifier of this type is given by:

$$T_{\mathrm{amp}} = T_{\mathrm{d}} \left[\frac{1}{p} + \frac{f_{\mathrm{s}}}{f_{\mathrm{i}}}\left(1 + \frac{1}{p}\right)\right]. \qquad (2)$$

Here T_{d} is the physical temperature of the diode and $p = R_{\mathrm{g}}/R_{\mathrm{d}}$ where R_{g} is the source impedance at the signal frequency f_{s}. The quantity p is known as the *overcoupling ratio.*

It would appear from (2) that all that has to be done to give a low value of input noise temperature is to make p high and $f_{\mathrm{s}}/f_{\mathrm{i}}$ low. However, p and f_{s} are already related: it can be shown [10] that for high gain

$$f_{\mathrm{s}} f_{\mathrm{i}}(1 + p) = (\gamma f_{\mathrm{c}})^2, \qquad (3)$$

where γ is a measure of the *capacitance variation* of the diode defined by $\gamma = (C_{\mathrm{max}} - C_{\mathrm{min}})/2(C_{\mathrm{max}} + C_{\mathrm{min}})$ and f_{c} is the cut-off frequency defined by $f_{\mathrm{c}} = 1/2\pi C R_{\mathrm{d}}$[11]. The relation (3) is derived by setting the effective negative resistance in the signal circuit at signal frequency equal to the total positive resistance, a condition which is met to a good approximation for high gain. From (2) and (3) it can be shown that there are theoretical optimum values for both p and f_{i}, giving a minimum noise temperature for the amplifier. This is

$$T_{\mathrm{amp\ min}} = 2 f_{\mathrm{s}}\, T_{\mathrm{d}}/\gamma f_{\mathrm{c}}. \qquad (4)$$

The quantity γf_{c} is thus a figure of merit for the varactor diode, for it indicates the lowest input noise temperature that it will give in an "optimum" amplifier. The optimum idler frequency is equal to this figure of merit:

$$f_{\mathrm{i\ opt}} = \gamma f_{\mathrm{c}}. \qquad (5)$$

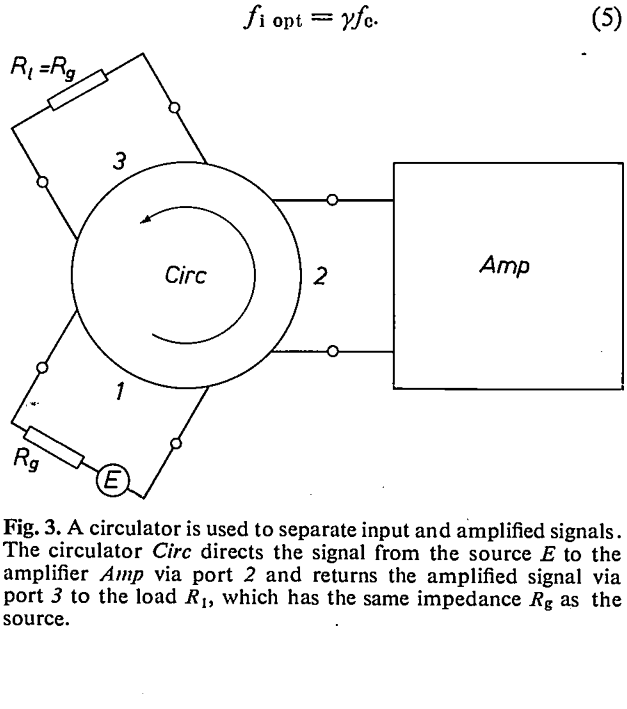


Fig. 3. A circulator is used to separate input and amplified signals. The circulator *Circ* directs the signal from the source *E* to the amplifier *Amp* via port *2* and returns the amplified signal via port *3* to the load R_{l}, which has the same impedance R_{g} as the source.

The best commercially available diodes (e.g. the Mullard CXY10) have a γf_{c} of 40 GHz or more.

The power gain G and operating bandwidth B for a parametric amplifier with simple tuned circuits as in fig. 2 are related by:

$$G^{\frac{1}{2}}B = 2/(B_{\mathrm{s}}^{-1} + B_{\mathrm{i}}^{-1}), \qquad (6)$$

where B_{s} is the bandwidth of the unpumped signal circuit and B_{i} is the bandwidth of the unpumped idler circuit. As indicated in the Introduction, greater bandwidths can be achieved, at the same gain, by using simple filter techniques [7] [12].

Design and performance of the amplifier system

Design of basic amplifier

As we saw earlier, the effective input noise temperature had to be less than 35 K for radio astronomy. The lowest operating temperature for a diode cooled by liquid nitrogen would be about 80 K, and the amplifier was to operate at about 2.7 GHz.

With γf_{c} taken as 40 GHz, eq. (4) shows that the minimum noise temperature that can be obtained at this frequency from a diode cooled to 80 K is 10.5 K. The relations (5) and (1) show that to achieve this minimum value the idler frequency should be about 40 GHz and the pump frequency should be about 43 GHz. However, it was not as easy to design a practical parametric amplifier with a noise temperature below 35 K as these figures might suggest. There was no convenient source that would deliver pump power (about 150 mW) at a frequency of 43 GHz, and because of the stray elements of the varactor diode it was inconvenient to provide an idler circuit resonant at 40 GHz. Moreover, the figure of 35 K for the noise temperature would have to include noise contributions from other parts of the parametric amplifier unit such as the circulator and the microwave feed paths. As we shall see below, the circulator alone can contribute some 10 K of noise.

The highest frequency for which there was a convenient pump source available was 40 GHz. As (1) shows, the idler frequency should then be about 37 GHz. In practice it was not possible to achieve this, but the

[8] See pp. 206-207 of the article by Aitchison [5].

[9] L. A. Blackwell and K. L. Kotzebue, Semiconductor-diode parametric amplifiers, Prentice-Hall, London 1961.

[10] See for example C. S. Aitchison, R. Davies and P. J. Gibson, A simple diode parametric amplifier design for use at *S*, *C*, and *X* band, IEEE Trans. **MTT-15**, 22-31, 1967.

[11] C_{max} is usually taken as the capacitance at 1 μA forward current and C_{min} as the capacitance at −6 V reverse bias. The cut-off frequency f_{c} is probably best defined for the value of C corresponding to zero bias, but manufacturers often quote cut-off frequencies for −6 V reverse bias.

[12] C. S. Aitchison, R. Davies and C. D. Payne, Bandwidth of a balanced micropill-diode parametric amplifier, IEEE Trans. **MTT-16**, 46-47, 1968.

CXY10 diode does have a resonance at about 27 GHz, due to stray inductance in series with the junction capacitance. Equations (2) and (3) show that the noise temperature of the resulting non-optimum amplifier will only be about 3 K higher than the minimum value of 10.5 K. From eq. (3) it can be shown that the overcoupling ratio should be about 20.

Some earlier work had already been done on a design principle that could be used with the CXY10 diode to provide resonant circuits for the signal, idler, and pump frequencies required. In this approach [5] [12] a *second diode* is introduced. The final design for the amplifiers for the radio interferometer is based upon this arrangement. We shall now look at the design of the amplifier with the aid of the schematic drawing shown in *fig. 4*.

The two diodes D are located inside the pump waveguide W with their top ends connected by the flat bar B, and in opposite polarity. (An enlarged sketch of a single diode is shown on the right.) The centre of the bar is joined to the coaxial input system C_1 by the post P. The lower end of each diode is connected to a coaxial circuit C_2. The pump waveguide W is terminated in a matched load behind the diode structure, and there is a tunable filter in the pump waveguide (neither of these is visible in the figure). As the figure shows, the coaxial input system consists of a cascaded series of lengths of coaxial line l_1 and l_2, stepped in the way shown. At the top of the coaxial system there is a connection to one port of the circulator (port *2* of fig. 3).

Let us now try to explain how the arrangement corresponds to the schematic diagram of fig. 2. At first glance it seems as though it cannot do so, for there are two diodes. However, at the signal frequency the two diodes may be considered to form a single unit, since the bar connecting the diodes is much smaller than the wavelength, and the diodes are therefore effectively in parallel. The filter F_s is not present as a separate item; its function is achieved by making C resonant with X_s, which therefore has to be an inductance (see *fig. 5a*). An inductance L_1 is provided by the bar-and-post structure, which also serves to couple the pump power into the two diodes. The tunable filter mentioned above (but not shown) is used to match the diode to the pump source.

At the idler frequency, the two diodes are again combined, but in such a way that they form a closed resonant loop (see fig. 5*b*), connected to the outside world only by the parametric action of the varying capacitance. In each diode the stray inductance L_2 and the junction capacitance C give the series resonance mentioned above, so that the complete idler loop thus formed is resonant at about 27 GHz. The element F_i of fig. 2 therefore represents the filtering effect of the two simple tuned circuits that form the loop; X_i represents the reactance due to diode stray inductance. With identical diodes connected in opposite sense this idler circuit is balanced, i.e. no idler voltage will appear across the terminals AA, which correspond to the lower end of the post P in fig. 4. The diodes cannot be connected directly together since they are not small compared with the idler wavelength λ_i ($\approx$ 1 cm) and so are connected instead by a transmission line that is half a wavelength long at the idler frequency. (A half-wave line has no transforming effect.) Physically, this transmission line is a microstrip line formed by the bar B and the lower wall of the waveguide.

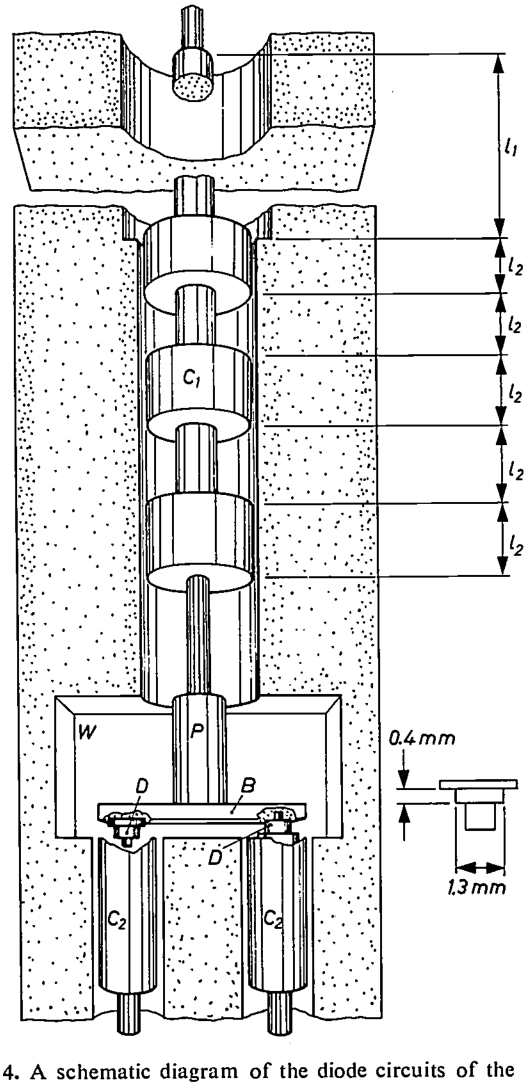


Fig. 4. A schematic diagram of the diode circuits of the parametric amplifier. The diodes D are located in the pump waveguide W (axis into the drawing) and linked by the bar B. The diodes are connected to the bar in opposite sense and are connected to the signal input and output (i.e. to the circulator) by the inductive post P and the coaxial-line system C_1. The stepped coaxial lines form a double quarter-wave transformer at the signal frequency f_s and a stop-filter at the idler frequency f_i; $l_1 = \lambda_s/4$ and $l_2 = \lambda_i/4$ where λ_s is the signal wavelength and λ_i is the idler wavelength. The distance between the two diodes is made equal to $\lambda_i/2$. The coaxial circuits C_2 are low-pass filters for the diode current monitoring circuit. An enlarged sketch of a single diode is shown on the right.

The coaxial structure C_1 has two functions: it transforms the input impedance to the correct value and also provides a low-pass filter that will pass the desired signal but reject signals at the idler frequency (such signals may arise if the diodes are not quite identical). We saw above that a single diode should be fed from a signal source whose impedance is about 20 times greater than the spreading resistance R_d. With *two* diodes effectively in parallel the source impedance should be about 10 R_d, and at the signal frequency the coaxial structure acts as a double quarter-wave transformer to transform the standard 50-ohm input impedance to the appropriate value. The line of length l_1, equal to a quarter of the signal wavelength λ_s, is the first stage of the transformer and the second stage is formed by the alternate low- and high-impedance lines each of length l_2. This second stage of the transformer also acts as the low-pass filter; it is designed with l_2 equal to a quarter of the idler wavelength and therefore rejects signals at the idler frequency but passes the desired signal.

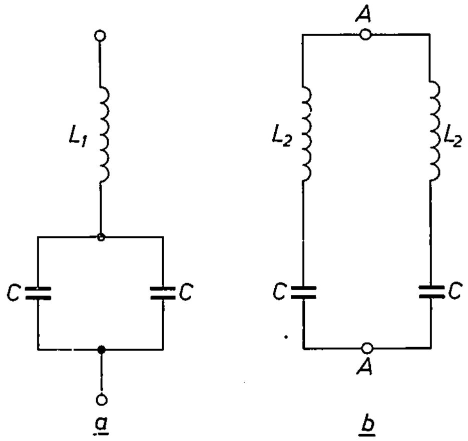


Fig. 5. *a*) At the signal frequency f_s the diodes are effectively in parallel, since the bar connecting the diodes is much shorter than λ_s. The inductance L_1 due to the bar-and-post structure corresponds to the tuning reactance X_s of fig. 2, and the series resonance of L_1 and the junction capacitances C provides the filter action of F_s. *b*) At the idler frequency f_i the two diodes are combined in such a way that they form a closed resonant loop. In each diode the stray inductance L_2 gives a series resonance with the junction capacitance C. For the CXY10 diode this resonance is near 27 GHz, so that the complete idler loop is resonant at this frequency. Here the stray inductance corresponds to the tuning reactance X_i of fig. 2, and the series resonance inside the loop provides the filter action of F_i. With identical diodes connected in opposite sense this circuit is balanced, and no idler voltage will appear across the terminals AA.

Since the idler circuit is well decoupled from other circuits by the balanced arrangement and the low-pass filter it has quite a large bandwidth (about 3 GHz). This makes it easier to obtain a sufficient signal bandwidth (see eq. 6), and also makes it easier to stabilize the signal phase against pump-frequency variations. The coaxial circuits C_2 are also low-pass filters: they provide the connections to the monitoring circuit.

The amplifier-circulator unit

We saw above that a microwave circulator is used with the parametric amplifier to separate the input and output signals. The circulator characteristics of importance are the *isolation* between ports that should not be coupled, and the *loss* between coupled ports. If the isolation is too low there will be an appreciable reflected signal, and hence mismatch, at the input port. The loss in the circulator is undesirable because it reduces the signal and has the effect of increasing the effective input noise temperature of the receiver.

Both coaxial and waveguide circulators were available when the equipment was designed, but the one whose performance was most suitable was a waveguide version. This was a four-port device with an isolation of 40 dB and an insertion loss of 0.15 dB. With this degree of isolation the reflected signal at the input is sufficiently small (at 20 dB gain the voltage reflection coefficient is less than 0.13). A loss of 0.15 dB at room temperature corresponds to an increase in noise temperature of nearly 12 K, and this is acceptable in our system provided that the waveguide-to-coaxial transition that connects amplifier and circulator is cooled: The effect on the noise temperature of the various losses in the amplifier-circulator unit is discussed below. A schematic arrangement is shown in *fig. 6*.

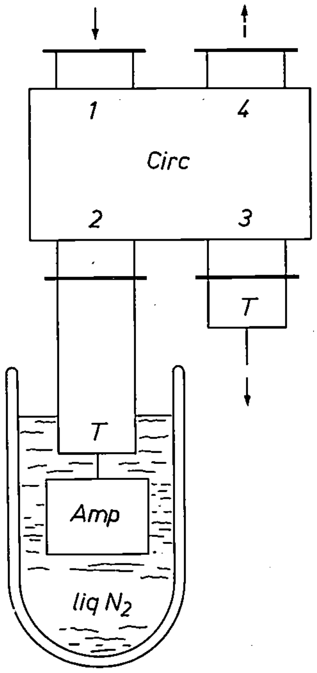


Fig. 6. A schematic diagram of the amplifier-circulator unit. The signal enters at port *1* of the four-port waveguide circulator *Circ*, which is at room temperature. It emerges at port *2*, which is terminated in the parametric amplifier *Amp*, cooled by liquid nitrogen. The amplified reflected signal arriving at port *2* emerges at port *3*. Port *4* is terminated in a matched load at room temperature. *T* indicates a waveguide-to-coaxial transition.

Calculated noise performance of the amplifier system

Table I lists the physical temperature, loss and theoretical noise contribution [9] for each component of the parametric-amplifier system. (The noise contribution from the amplifier is of course calculated from eq. 2.) Adding up the noise contributions in the right-hand column shows that at a system gain of 23 dB the noise temperature of the amplifier system should be 31.K. Some of the components in the Table are listed twice since they are encountered by the signal both before and after amplification. The rather high operating temperature of 320 K for the uncooled components was chosen because of the heat generated by the pump klystron: this will be discussed later. .

Table I. System noise temperature at 23 dB system gain in terms of the contribution of each component.

Component	Physical temperature K	Loss dB	Noise contribution K
Input waveguide	320	0.02	2.25
Circulator (input)	320	0.15	11.8
Downlead	80	0.05	1.34
Transition	80	0.10	2.3
Amplifier	80	—	13
Transition	80	0.10	0.004
Uplead	80	0.05	0.002
Circulator (output)	320	0.30	0.09
Transition	320	0.1	0.03
Coaxial output lead	320	0.3	0.09
Total noise temperature			31 K

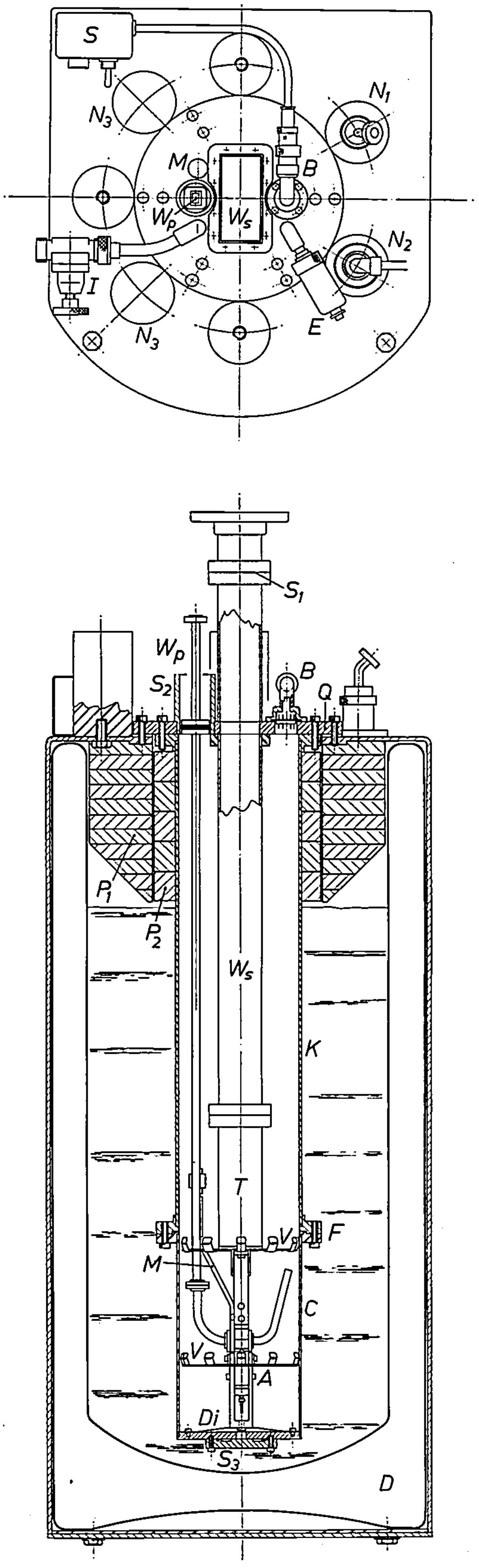


Fig. 7. Sectional drawing and top view of the cryogenic system. Outer diameter 31.8 cm, inner container diameter 10.2 cm and length 58.4 cm. *D* is a standard 34-litre liquid-nitrogen dewar made of stainless steel. The amplifier *A* is mounted in an atmosphere of helium gas inside the sealed container *K*, whose upper part is made of glass-fibre bonded resin to reduce the nitrogen boil-off rate. The lower part C is made of copper, and is joined to the upper part by steel flanges *F* with a gold O-ring vacuum seal. *T* waveguide-to-coaxial transition. W_s signal waveguide: this has very thin walls of copper-plated stainless steel. S_1 vacuum seal of thin plastic sheet mounted across the signal waveguide. The pump waveguide W_p is made from silver-plated nickel, and has a sliding vacuum seal S_2 to allow for differential contraction. *M* remotely adjusted matching unit to match pump power to the diodes. To keep down the boil-off rate plugs P_1 and P_2 of expanded PVC are mounted beneath the top plate *Q*; the horizontal lines represent annular aluminium-foil radiation shields. The thermal connection between the liquid nitrogen and the amplifier is formed by the copper base *C*, the copper fingers *V* and the diaphragm spring *Di*, which allows for differential contraction. Further heat transfer is also provided by the helium gas. The baseplate nitrogen-vacuum seal S_3 is demountable, with a locking sealant, so that the amplifier can be removed from the inner container.

On top of the inner container: W_p pump waveguide, W_s signal waveguide, *M* pump-match control, *E* helium safety-valve, *B* diode bias connector, *I* helium inlet valve. In the filling process the inner container is initially evacuated through *I*, then flushed with helium gas and evacuated again. Finally it is filled with helium gas at a pressure of about 1.1 atm. With the helium supply still connected via *I* the dewar is then filled with liquid nitrogen through N_1. When the fill is complete *I* is closed and the helium supply disconnected. N_2 sensing point for liquid-nitrogen level; this can be used with an automatic nitrogen-replenishing system. Gaseous nitrogen can escape through the nitrogen vents N_3, which have heated release valves. Switch *S* will connect a meter into the varactor-diode circuit, which is short-circuited under operating conditions.

Thermal considerations

The cryogenic system

The cryogenic system, which contains the cooled components of the amplifier complex, is shown in *fig. 7.* To keep the volume occupied by the cooled components as small as possible, the waveguide-to-coaxial transition *T* is of the "end-fire" type [13]. The amplifier *A* and the signal and pump waveguides are suspended from a top plate *Q* and are housed in a thermally designed sealed container *K*, which is filled with helium gas. The container keeps the microwave components free of water vapour and liquid nitrogen, which would introduce losses and also variations in performance as the nitrogen level changed.

The container must introduce only negligible thermal loss to the outside while providing a good thermal connection between the amplifier and the liquid nitrogen. The thermal loss is minimized by making the upper part of the container of glass-fibre bonded resin, and good thermal contact between the amplifier and the liquid nitrogen is ensured by making the lower part *C* of the container of copper. The resin tube and the copper section are joined by a pair of special stainless-steel flanges *F*, one brazed to the copper and one bonded to the resin. Between the two flanges, which are bolted together, there is a gold O-ring to form the vacuum seal. The seal is unaffected by repeated temperature variations between 77 K and 300 K. Good thermal contact between the amplifier and the copper wall is ensured by copper fingers *V* and a diaphragm spring *Di* which allows for differential contraction. Further heat transfer is provided by the helium gas inside the container.

A view of the end-fire transition and the amplifier assembly is shown in *fig. 8.*

The waveguide feed W_p for the 29.5 GHz pump signal (internal dimensions 3.56×7.12 mm) is electroformed from nickel and plated inside with 5 μm of copper to give good electrical conduction. The larger signal-waveguide feed (3.404×7.214 cm inside) could not be made in this way, however, since the thermal conduction would then be too large. This waveguide is fabricated from 0.038 mm stainless steel sheet in two halves, which are welded together. This waveguide is also plated inside with 5 μm of copper. There is a row of six holes along the centre of one broad face of this waveguide to prevent it from collapsing when the vacuum is released during the filling process (see caption to fig. 7). *Fig. 9* shows a photograph of the sealed container and the amplifier complex.

The sealed container is suspended in liquid nitrogen contained in a stainless-steel dewar of standard dimensions. To reduce the boil-off rate the underside of the dewar top plate is lined with a 15 cm thick layer of expanded polyvinyl chloride (PVC). Heated nitrogen vents are included in the top plate to prevent a dangerous build-up of pressure. A safety valve (10 lbs/in², i.e. about 0.7 kg/cm²) is also fitted for the helium gas in the sealed container.

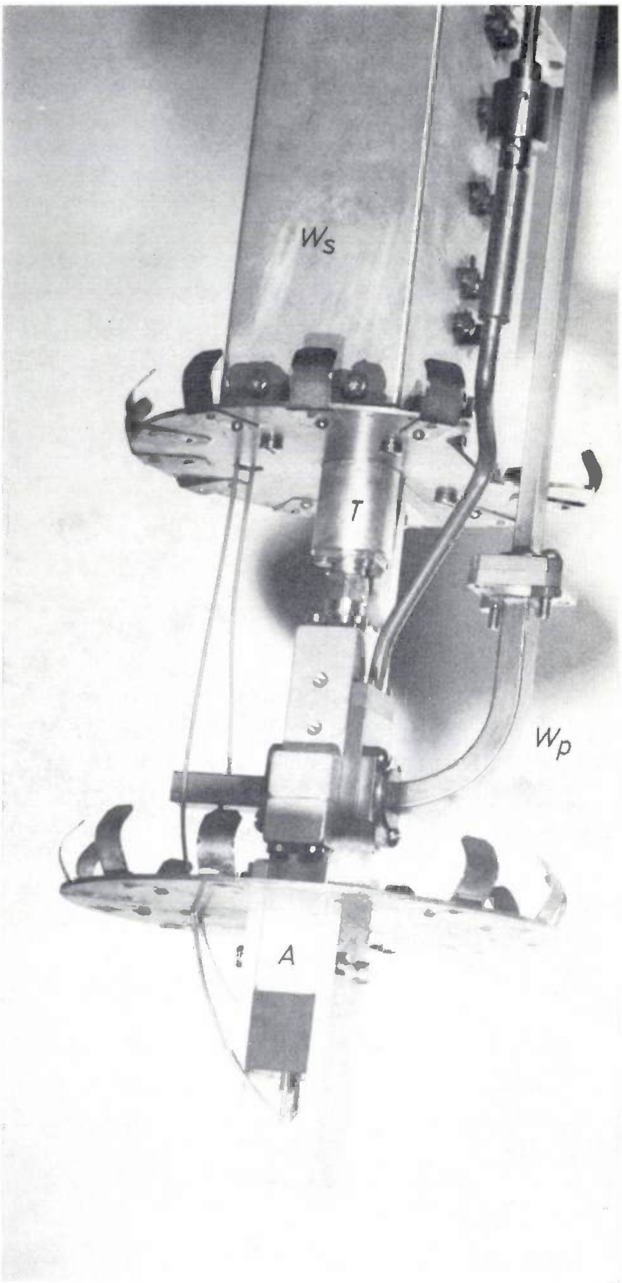


Fig. 8. For a compact design the parametric amplifier *A* is connected to an end-fire waveguide-to-coaxial transition *T* at the end of the signal waveguide W_s. The pump waveguide W_p is on the right.

In the design of the cryogenic system great care was taken to allow differential contraction of the various components while keeping the structure rigid enough

[13] J. C. Dix, Design of waveguide/coaxial transition for the band 2.5-4.1 Gc/s, Proc. IEE **110**, 253-255, 1963.

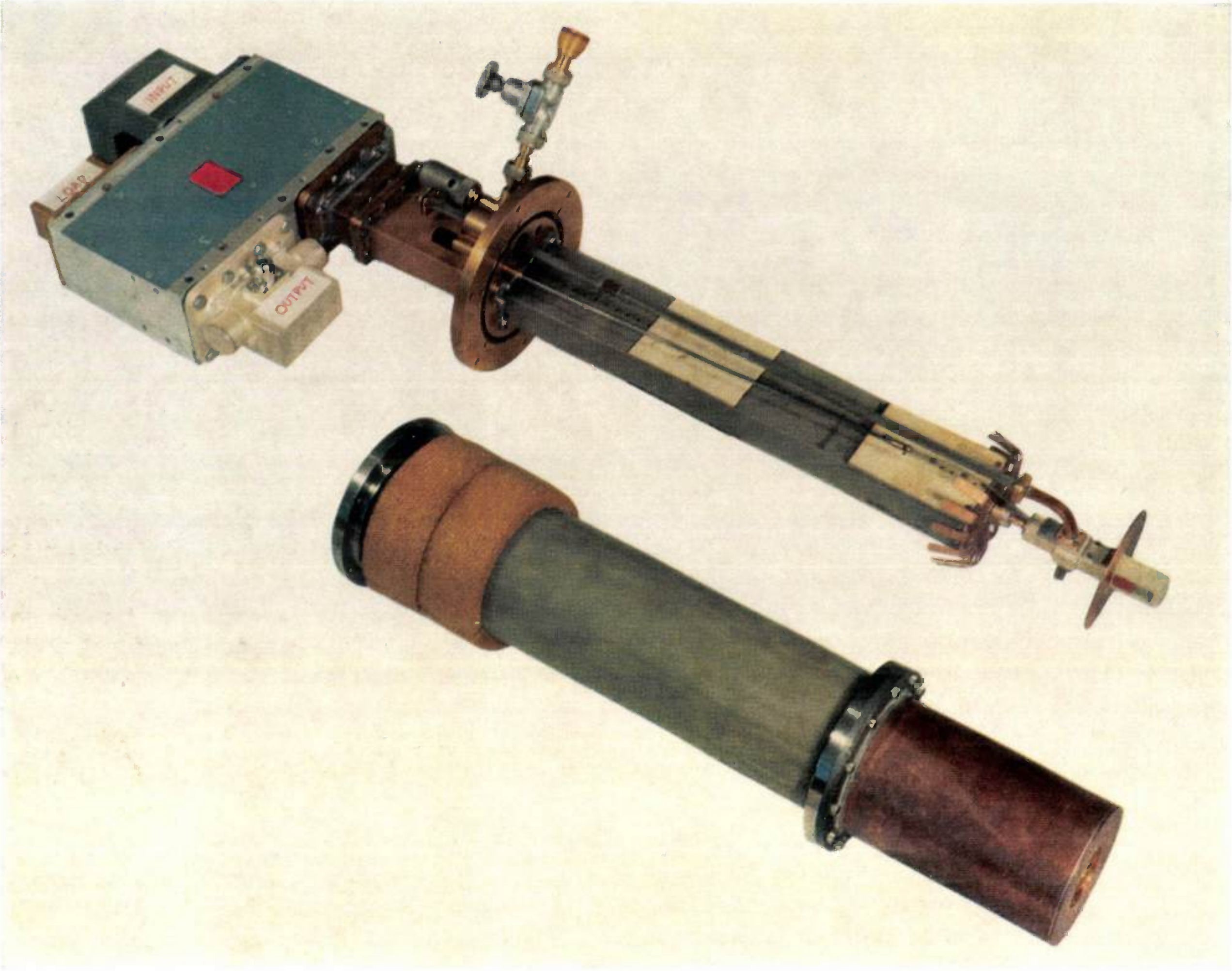


Fig. 9. The sealed inner container of the cryogenic system (foreground) has an upper part of glass-fibre bonded resin and a lower part of copper. The PVC rings (P_2 in fig. 7) can be seen at its upper end. The amplifier complex, which fits into the container, is shown behind it.

to give stable operation. The design was also arranged so that the number of demountable seals immersed in liquid nitrogen was as small as possible.

A single filling of liquid nitrogen will keep the system adequately cooled for 100 hours when the dewar is kept vertical. However, the complete cooled amplifier system is rigidly attached to the aerial, and may therefore tilt through ± 45°. Under these conditions a single filling will keep the system cooled for 72 hours.

An automatic nitrogen-replenishing system has also been added which can be used if there is a need for a long period of continuous operation. It also allows the amplifier system to be operated more easily at the primary focus of the aerial.

The temperature-stabilized enclosure

The temperature of the varactor circuits is stabilized at about 80 K by the liquid nitrogen, but to achieve the stability necessary in interferometry the temperatures of the pump klystron and the circulator also have to be stabilized. These components are therefore housed in a temperature-stabilized enclosure mounted on top of the dewar.

The temperature-stabilized unit is shown in the photograph of *fig. 10*. Since the klystron gives out nearly 60 watts of heat it is convenient to stabilize the enclosure to a temperature above the highest ambient value. This is done by using air blowers with heaters that are automatically controlled by a mercury-contact thermometer.

The outer surface of the unit, through which the heat is dissipated, is finished in a white enamel paint that has an emissivity of about 0.8. The unit measures 50 × 37.5 × 25 cm. A weather-proofing cover, finished with the same paint, is also available: this enables the stabilizer-dewar unit to be used out of doors.

Performance

The main details of the performance of the amplifier system are given in *Table II.*

The noise temperature was measured by the *Y*-factor method, used earlier with the masers [2] that the parametric amplifiers replace. In this method matched loads at room temperature and at 77 K are connected in turn to the input and the difference in noise output is recorded. In use, however, the noise performance of the system is a function of the aerial match: there is a room-temperature matched load at the fourth port of the circulator, and any mismatch in the input circuit will therefore cause noise to be reflected back into the system.

Table II. The main performance figures for the parametric amplifiers for the Defford radio interferometer.

Signal centre frequency	2.695 GHz (11.12 cm)
Gain	23 dB
Effective input noise temperature	31 ± 3 K
3 dB bandwidth	40 MHz
Input voltage reflection coefficient	< 0.13
Gain stability	± 0.25 dB/hr ± 0.5 dB/day
Phase stability	± 2°/day
Pump frequency	29.5 GHz
Pump power	150 mW
Operating temperature	77 K
Operational time (vertical)	100 hours
Operational time (with tilting through ± 45°)	72 hours

The phase stability was measured by means of a phase bridge and the results are shown graphically in *fig. 11*. The measurements were made for a wide range of ambient temperatures and indicate a stability of better than ± 2° per day [14].

The authors would like to thank the United Kingdom Ministry of Defence (Navy Department) for permission to publish this article, which refers to work carried out under a CVD (Coordination of Valve Development) contract.

Appendix: Physical explanation of parametric amplification

The explanation of parametric amplification given some years ago in this journal [15] was limited to a rather simple case. Here we shall extend that explanation a little further in the way indicated by H. Mooijweer [16].

We consider a simple lossless *LC* circuit (*fig. 12*) in which there is a current of frequency $f = 1/2\pi\sqrt{LC}$. The charge on the capacitor and hence the voltage across it then vary sinusoidally with time. The capacitance is the reactive element that is varied, and it is assumed that the capacitance variation shown in fig. 12 is obtained by moving the plates of the capacitor. Every time that the charge reaches a maximum value, the plates are pulled sharply apart, and they are moved together again at the zero points of the charge curve. The first movement requires work to be supplied, but the second requires no work. There is therefore an overall flow of energy to the circuit in each period of the oscillation; this energy goes to aid the oscillation. The amplitude of the

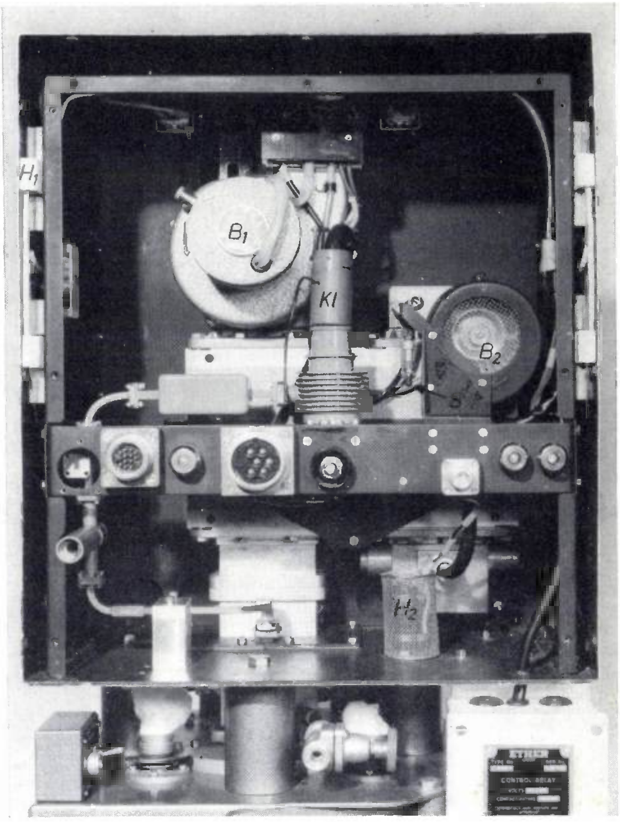


Fig. 10. Photograph of the temperature-stabilized enclosure. Air is sucked over the components by the main blower B_1, then forced over the low-thermal-capacity heaters H_1 between the two skins of the case. The klystron *Kl* is blown by the small blower B_2; the sensor *S*, which is a mercury-contact thermometer, is situated between B_2 and the klystron. The anti-condensation heaters H_2 are switched on when the amplifier is switched to standby.

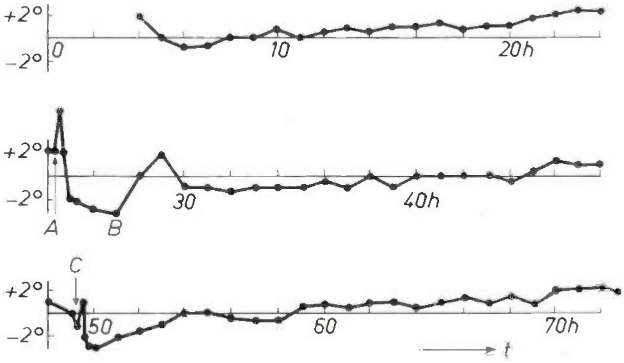


Fig. 11. Phase stability of the parametric-amplifier system over three days' continuous operation. *A* indicates nitrogen fill by pouring. At *B* there was a disturbance in the measuring system. *C* marks the start of automatic nitrogen filling. The phase stability is ± 2° per day.

[14] The phase-stability information was made available by Mr. G. Moule of the Royal Radar Establishment, Malvern, England.

[15] B. Bollée and G. de Vries, Experiments in the field of parametric amplification, Philips tech. Rev. **21**, 47-51, 1959/60.

[16] H. Mooijweer, Ned. T. Natuurk. **30**, 145, 1964.

voltage across the capacitance consequently increases in steps, since the charge tends to remain constant when the capacitance is altered because of the presence of the inductance. The effect is like a sort of "negative damping". In this way spontaneous oscillations in the circuit can grow from a small disturbance because of the periodic variation in capacitance (the pumping action) at a frequency equal to twice the resonant frequency f of the circuit. This implies that a small signal introduced into the circuit at its resonant frequency will also grow in magnitude, i.e. be amplified.

For a circuit with losses, the situation would have been much the same except that some of the energy supplied by pumping would have been dissipated in the losses.

components are equal, but this is not essential. The total charge is then $Q = Q_1 + Q_2 = 2Q_0 \sin \frac{1}{2}(\omega_1 + \omega_2)t \cos \frac{1}{2}(\omega_1 - \omega_2)t$, an oscillation of frequency $\frac{1}{2}(\omega_1 + \omega_2)$ modulated in amplitude at a frequency $\frac{1}{2}(\omega_1 - \omega_2)$. (The carrier is suppressed.)

The pump source again provides a capacitance variation in the form of a square wave. The capacitance is increased at the zeros of $\sin \frac{1}{2}(\omega_1 + \omega_2)t$ and reduced half-way between the zeros. This means that the pump frequency is now equal to $f_p = f_1 + f_2$. Moreover, the energy flow from the pump source to the capacitance is no longer the same at every decrease in capacitance because of the amplitude-modulated character of the charge. However, it is never negative, since the capacitance is always increased at the zeros of the charge curve.

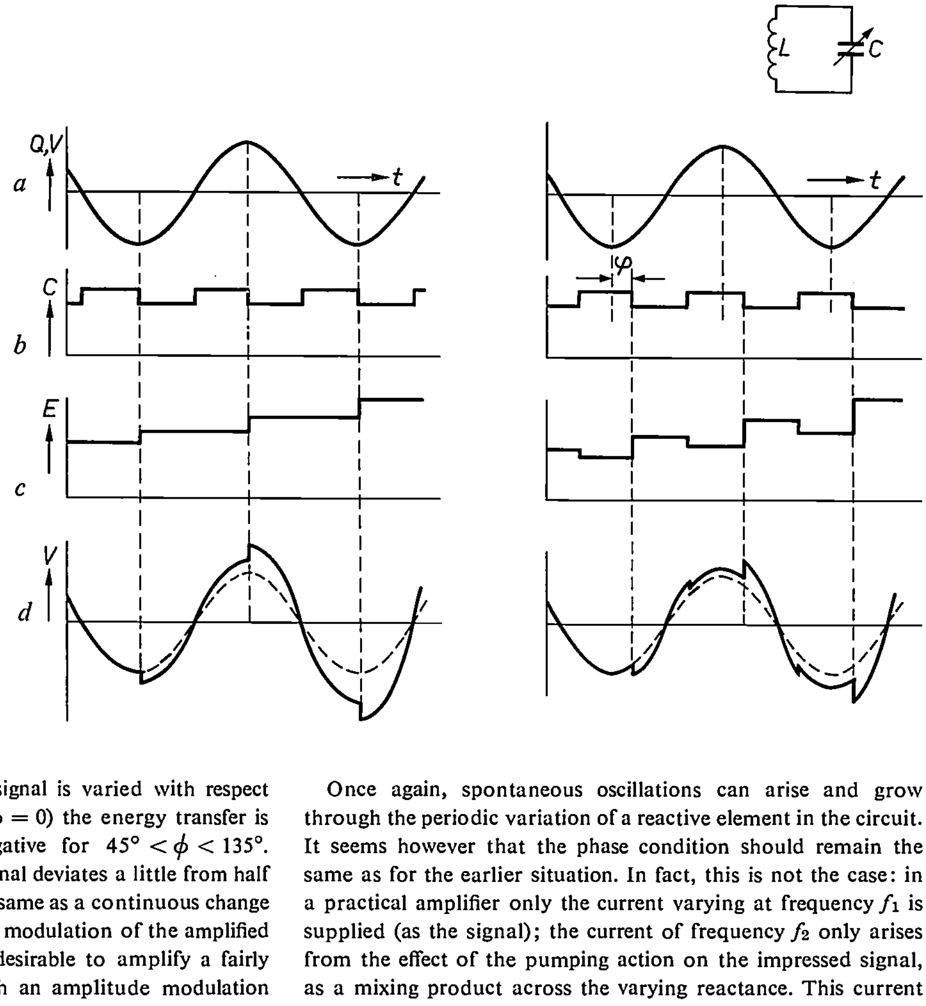


Fig. 12. Parametric excitation in a simple LC circuit. The capacitance C is the reactance that is varied; it is assumed that the variation is caused by moving the plates apart and together. *a*) Original variation with time of the voltage V and charge Q on the capacitance. *b*) Periodic variation of the capacitance C. *c*) Increase in the circuit energy E caused by the pumping action. *d*) Growth of the voltage V across the capacitance ("negative damping").

If the phase ϕ of the pumping signal is varied with respect to the situation shown in fig. 12, ($\phi = 0$) the energy transfer is reduced [9] and can even be negative for $45° < \phi < 135°$. Now if the frequency of an input signal deviates a little from half the pump frequency the effect is the same as a continuous change in phase: there will be an amplitude modulation of the amplified signal. In many applications it is desirable to amplify a fairly wide band of frequencies and such an amplitude modulation would then arise.

This difficulty can be avoided by using a slightly more complicated circuit. Instead of a single oscillatory circuit with periodically varying capacitance or inductance, a circuit is used that has two (or more) resonant frequencies f_1 and f_2, with the coupling at the periodic reactance. It is found that in this configuration the phase condition described above is no longer present. We shall explain this with the aid of *fig. 13*. Again, it is assumed that the circuit is lossless and that C is the reactive element that is varied. It is also assumed that one loop only supports current at frequency f_1, and the other loop only supports current at frequency f_2. The total charge on the capacitance then consists of a component $Q_1 = Q_0 \sin \omega_1 t$ and a component $Q_2 = Q_0 \sin \omega_2 t$, where ω_1 and ω_2 are the angular frequencies corresponding to f_1 and f_2. For convenience we have assumed that the amplitudes of the two

Once again, spontaneous oscillations can arise and grow through the periodic variation of a reactive element in the circuit. It seems however that the phase condition should remain the same as for the earlier situation. In fact, this is not the case: in a practical amplifier only the current varying at frequency f_1 is supplied (as the signal); the current of frequency f_2 only arises from the effect of the pumping action on the impressed signal, as a mixing product across the varying reactance. This current generated in this way at f_2 then has the correct phase automatically. If the wrong phase were to appear, it would be quickly damped out anyway, since this energy would be extracted from the circuit by the pumping action. The mixing product, the charge or current varying at $f_p - f_1 = f_2$, is the idler signal.

In this arrangement, with the phase condition no longer relevant, we can choose the frequencies f_1 and f_2 far enough apart to separate them with filters.

The arrangement with the two resonant circuits can be considered as a more general case; the arrangement with the single resonant circuit and $f_1 = f_2 = \frac{1}{2}f_p$ is therefore referred to as the "degenerate" case.

In an ordinary LC circuit with loss represented by a resistance R in the loop a disturbance will give rise to a damped waveform of the form $\exp(-R/2L)t \cos(\omega t + \phi)$. But in the pumped

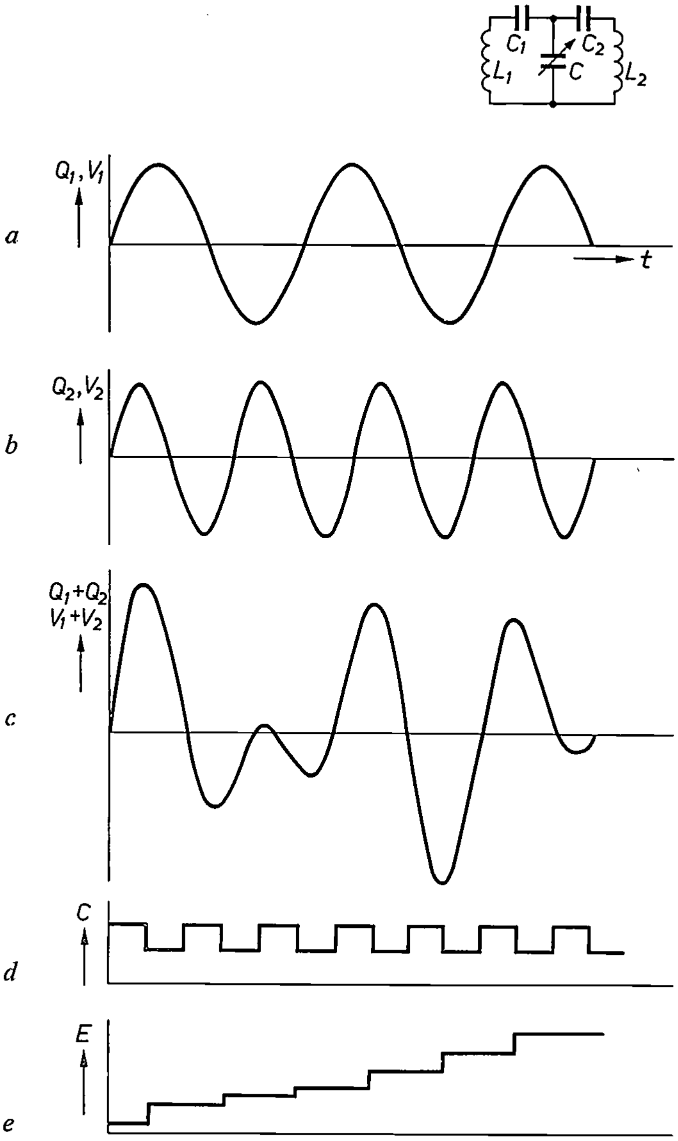


Fig. 13. Parametric excitation in a circuit with two resonant frequencies. *a*) Component Q_1 of the charge varying at frequency f_1. *b*) Component Q_2 of the charge varying at frequency f_2. *c*) Total charge Q_1+Q_2 on the varying capacitance. *d*) Periodic variation of the capacitance *C*. *e*) Energy increase in the circuit, due to the pumping action.

circuits we have been considering here, the oscillatory waveform *grows* with the pumping action (before it is eventually limited by non-linearity). Now assuming that the waveform grows exponentially, the coefficient of the exponential would have to be positive — which would imply an effective *negative resistance*. In analysis it is frequently convenient to consider the gain action as being due to the introduction of a negative resistance into a circuit.

The pumping action in the above simple description was assumed to be a square-wave variation of capacitance, caused by mechanical action. In practice, of course, the time variation of capacitance is obtained by driving a varactor diode with a large sinusoidal voltage.

Summary. A parametric-amplifier system has been developed by Mullard Research Laboratories for the radio-astronomy interferometer at Defford, England. Two of these systems replace the two masers previously used. A very stable and sensitive receiver system is required in radio astronomy, and such requirements can be met by using a varactor-diode parametric amplifier, cooled with liquid nitrogen as the first receiver stage. A brief recapitulation of parametric-amplifier principles is followed by an account of the design approach of a two-diode amplifier and a description of the complete system. Special attention is given to the effect of loss and other imperfections in the signal-feed system, and the resulting noise degradation is evaluated. Stability is ensured by careful temperature stabilization of the uncooled microwave components. The amplifiers operate at 2.695 GHz and are pumped at 29.5 GHz. The noise temperature is 31 ± 3 K and the 3 dB bandwidth for 23 dB peak gain is 40 MHz. Gain stability is ± 0.25 dB over 1 hour, ± 0.5 dB over 24 hours, and the phase stability is ± 2° over 24 hours. The system will operate for 100 hours on one filling of nitrogen, and automatic replenishment can also be used.

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 Forschungslaboratorium Aachen GmbH, Weißhausstraße, 51 Aachen, Germany	*A*
Philips Forschungslaboratorium Hamburg GmbH, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (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: Recombination and trapping in epitaxial *n*-type gallium arsenide.
Philips Res. Repts. **26**, 261-278, 1971 (No. 4). *E*

V. Belevitch & Y. Genin: Cascade decomposition of lossless 2-ports.
Philips Res. Repts. **26**, 326-340, 1971 (No. 4). *B*

J. Burmeister: Crystal growth of tellurium by chemical transport.
Mat. Res. Bull. **6**, 219-223, 1971 (No. 4). *A*

J. P. Deschamps: Maximal classes of solutions of Boolean equations.
Philips Res. Repts. **26**, 249-260, 1971 (No. 4). *B*

G. Eschard, J. Graf & R. Polaert: High-speed shutter tubes of biplanar and proximity focusing design: improvements through introduction of a microchannel wafer.
IEEE Trans. **ED-17**, 986-989, 1970 (No. 11). *L*

Z. van Gelder: Calculation of non-linearities of the calibration curves in atomic absorption flame spectroscopy.
Spectrochim. Acta **25B**, 669-681, 1970 (No. 12). *E*

D. Gossel: Meßsysteme und Regelungen mit Frequenzsignalen.
Messen Steuern Regeln **14**, 22-28, 1971 (No. 1). *H*

J.-P. Hazan, J. Haisma, G. Marie & J. Nussli: Real-time oscilloscope observation of an ultrafast photodiode response to mode-locked laser pulses.
IEEE J. **QE-6**, 744-745, 1970 (No. 11). *L*

B. Hill: Verknüpfung von Nachrichtenkanälen mit digitalen Lichtablenkern.
Nachrichtentechn. Z. **23**, 549-552, 1970 (No. 11). *H*

M. Jung: Untersuchungen an elektrophotographischen Selen-Schichten mit Röntgenstrahlung.
Thesis, Karlsruhe 1970. *A*

D. Kasperkovitz: A new monolithic tristable element.
Solid-State Electronics **13**, 715-716, 1970 (No. 6). *E*

D. Kasperkovitz: The hook-emitter transistor: a proposed amplifier with high gain-bandwidth product.
Solid-State Electronics **13**, 1025-1031, 1970 (No. 7). *E*

K. Mouthaan: Characterization of nonlinear interactions in avalanche transit-time oscillators, frequency multipliers, and frequency dividers.
IEEE Trans. **MTT-18**, 853-862, 1970 (No. 11). *E*

K. Mouthaan: Low-frequency multiplication noise in silicon avalanche transit-time diodes.
Philips Res. Repts. **26**, 298-325, 1971 (No. 4). *E*

D. H. Paxman & R. J. Tree: Characteristics of bulk indium-phosphide microwave oscillators.
Electronics Letters **7**, 240-241, 1971 (No. 10). *M*

J.-C. Richard, P. Saget & G.-A. Boutry (Conservatoire National des Arts et Métiers, Paris): Contribution à l'étude des propriétés photoélectriques du sodium pur préparé et conservé dans l'ultravide.
C. R. Acad. Sci. Paris **271B**, 1098-1100, 1970 (No. 22). *L*

P. J. Severin: Four-point-probe resistivity measurements on silicon heterotype epitaxial layers with altered probe order.
Philips Res. Repts. **26**, 279-297, 1971 (No. 4). *E*

C. H. F. Velzel: Measurement of surface roughness by interferential contrast — an application of shearing interferometry to the study of phase objects.
Optical Instruments and Techniques, Proc. Conf. Univ. of Reading 1969, pp. 238-248; 1970. *E*

J. G. Wade & D. W. Parker: Stability of a wideband gyrator circuit.
Electronics Letters **7**, 224-225, 1971 (No. 9). *M*

F. F. Westendorp: Nonaging cold-pressed $SmCo_5$ magnets.
IEEE Trans. **MAG-6**, 472-474, 1970 (No. 3). *E*

Volume 32, 1971, No. 1 pages 1-32 Published 5th November 1971

PHILIPS TECHNICAL REVIEW

VOLUME 32, 1971, No. 2

The SO_2 monitoring network in the Rhine estuary region

H. J. Brouwer, S. M. de Veer and H. Zeedijk

In the industrial countries a great deal of effort is being put into projects for monitoring and controlling pollution of the environment. The article below describes a network designed by Philips for monitoring air pollution in the industrial zone of the Rhine estuary.

Introduction

Since the beginning of human life on Earth man has produced waste substances and discharged them into the atmosphere. With the advance of technological development their quantity has rapidly increased. The general public has now come to know many of these waste products by name: carbon dioxide, carbon monoxide, oxides of nitrogen, sulphur dioxide, ozone, unburned hydrocarbons, to mention only a few. Contrary to popular belief, the production of most of these substances by man is an order of magnitude smaller than their natural production. It is mainly because artificial production usually takes place in geographically limited areas that we are now having to pay so much attention to it. Considerable local concentrations can arise, particularly if the weather conditions change for the worse. Unpleasant odours, irritation of the eyes and of the respiratory tract, headaches, and damage to plant growth are frequent causes of complaint. There may even be an increase in the death rate, particularly among people more susceptible to air pollution because they have diseases of the respiratory organs.

In most industrial countries air pollution has become a matter of particular concern to the public authorities. The Dutch Government gave Philips a contract to develop a fully automatic network for monitoring the concentration of sulphur dioxide in the Netherlands. The first regional monitoring system, in the industrial belt of the Rhine estuary (Rijnmond), has meanwhile been completed and was put into service in October 1969 [1]. This will be the subject of the present article. Similar monitoring systems have now also been completed in other industrial areas of the Netherlands, in particular in the Provinces of Twente and Zeeuws-Vlaanderen. In the future these regional systems will be coupled to a nation-wide monitoring network with a wider "mesh". Negotiations are still in progress on plans to extend this future Dutch network to cover Belgium and part of West Germany. Preparations are being made at the same time to make the network suitable for monitoring other air pollutants, such as ozone, oxides of nitrogen, and carbon dioxide.

What is the point of measuring only a few of the many pollutants and why should sulphur dioxide be the first to be chosen? The weather conditions partly responsible for the formation of heavy concentrations have in principle the same effect on all air pollutants. An increase in the concentration of a particular compound is therefore a warning indication that the other compounds may also be present in larger concentrations. Moreover, experience has shown that air pollutants produced by sources of the same type — e.g. the carbon monoxide, oxides of nitrogen, and hydrocarbons from the exhaust of motor vehicles — are present in fairly constant ratios. For a given collection of sources the concentration of one compound therefore provides information about the concentration of the others. Sulphur dioxide is one of the most representative air pollutants in the sense that it is produced and

Drs. H. J. Brouwer and Ir. S. M. de Veer are with the Philips Industrial Equipment Division, Eindhoven; Ir. H. Zeedijk is on the staff of Eindhoven University of Technology.

[1] The monitor was designed in cooperation with the Government Institute for Public Health (RIG). Mention should also be made of the contributions from Prof. Ir. J. G. Hoogland of Eindhoven Technical University and from G. W. Meijerman of the Netherlands Institute of Public Health, with whom many valuable discussions were held.

discharged in relatively large quantities nearly everywhere in urban and industrial areas. What is more, it is a substance that rarely occurs naturally. A high SO_2 concentration is therefore a clear warning sign. For these reasons SO_2 is sometimes referred to as a tracer or indicator.

As we saw above, meteorological conditions have a considerable influence on changes in the concentrations of air pollutants. The vertical temperature gradient in the atmosphere is a particularly important factor. Normally the temperature decreases with increasing height. If, for reasons which we shall presently discuss, the temperature drop falls below a particular value or, even worse, if the temperature increases with increasing height ("temperature inversion") the air then becomes heavier in the lower layers than in the higher ones. The air pollutants then accumulate in the lower layers because there is no upward movement in the atmosphere.

Temperature inversion often takes place during a cloudless night as a result of cooling due to radiation from the Earth's surface. The inversion layer is usually broken up during the day by the sun and wind. It may happen, however, for example if there is no wind, that an inversion layer persists for a few days at a height of between 100 and 300 metres. These are the days when there is a greater likelihood of an increase of air pollution.

The monitoring network installed in the Rijnmond area takes continuous readings of the concentration of SO_2, changes in pollutant levels being immediately recorded and passed to a computer for processing. This rapid recording and processing makes it possible to detect at an early stage the weather conditions favouring air pollution [2], so that industry can be alerted in good time to take preventive measures. This early-warning facility is a special and unique feature of our system.

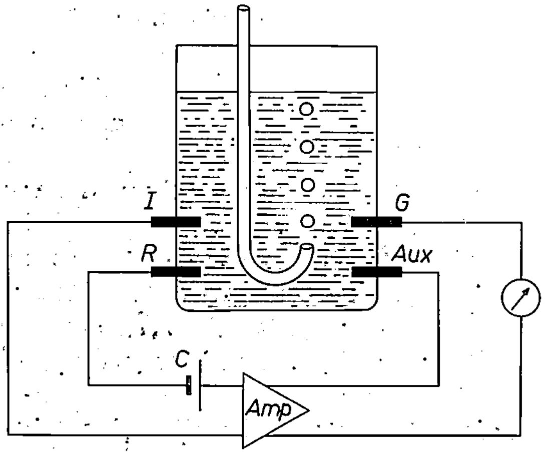


Fig. 1. Principle of SO_2 measurement by the "coulometric" method. *I* indicator electrode of platinum. *R* reference electrode of Ag-AgBr. *G* generator electrode. *Aux* auxiliary electrode. *C* voltage source. *Amp* amplifier. The air sample is passed through a measuring cell filled with a solution of 0.1 M potassium bromide and 2 M sulphuric acid, together with a small amount of free bromine. The SO_2 in the air reduces the bromine. The resultant decrease of the bromine concentration changes the potential difference between *I* and *R*, and this induces a current between *G* and *Aux*, generating a quantity of bromine that compensates for the amount consumed. The current between *G* and *Aux* is a measure of the quantity of SO_2 supplied per unit time.

Principle of the measurement

The SO_2 concentration is measured on the principle of continuous "coulometric" titration [3]. In the monitor that we have developed the air sample is passed through a measuring cell filled with a solution of 0.1 M potassium bromide in 2 M sulphuric acid, which also contains a small concentration of free bromine (*fig. 1*). The SO_2 in the air sample causes a decrease in the quantity of bromine by reduction of the free bromine to bromide ions. The concentration of bromine is kept constant by means of a control circuit consisting of two pairs of electrodes, a voltage source and an amplifier. The first electrode pair *I* and *R* "measures" the bromine concentration as a (redox) potential difference, and at the desired value of the bromine concentration this voltage is compensated by the voltage source *C* so that the output from the amplifier *Amp* is zero. A decrease in the bromine concentration as a result of reaction with SO_2 produces a change in the input signal of the amplifier, causing a current to flow between the second electrode pair *G* and *Aux*. Bromine is then formed by electrolysis until the bromine concentration has reached its initial value. The current that forms the bromine is recorded as a measure of the SO_2 concentration in the air.

An important characteristic of this principle of measurement is that no reagents are used. This is important because the monitor is required to work for long periods without supervision or inspection (for at least three months), and to some extent this determined the whole concept of the monitoring system.

The monitor

Air is drawn in by the SO_2 monitor at the top of a pole three metres high, known as a "sniffer" through which it passes to the measuring cell contained in a sturdily built box at street level (*fig. 2*). This box contains two removable compartments, one for the chemical

[2] L. A. Clarenburg, A telemetered system to predict unfavourable weather conditions, paper (No. 68-55) presented at the 61st Annual Meeting of the Air Pollution Control Association, St. Paul (U.S.A.), 1968.

[3] P. A. Shaffer, Jr., A. Briglio, Jr., and J. A. Brockman, Jr., Anal. Chem. 20, 1008, 1948; H. Landsberg and E. E. Escher, Industr. Engng. Chem. 46, 1422, 1954.

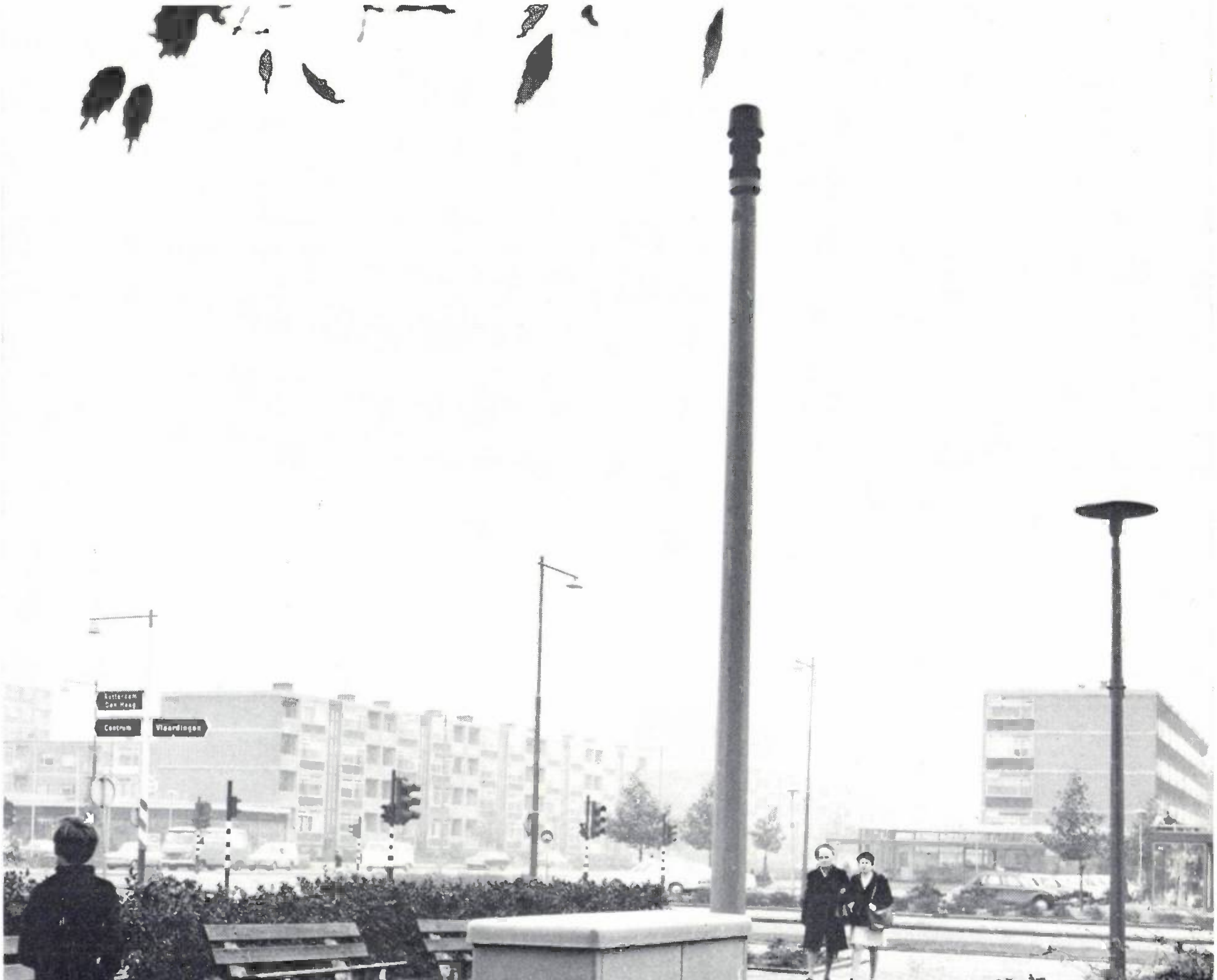


Fig. 2. An SO_2 monitor and "sniffer" pole at Schiedam.

equipment, which includes the measuring cell, and the other for the electronic equipment (*fig. 3*). The complete monitor is shown schematically in *fig. 4*.

The top of the "sniffer" pole is fitted with a dust filter *1*, electrically heated to a temperature which fluctuates from about 100 to 120 °C, depending on the outside temperature. At this high temperature no SO_2 is adsorbed on trapped insects or particles of dust, and in fog the filter is protected from drops of moisture, which could also take up SO_2. There is no oxidation of

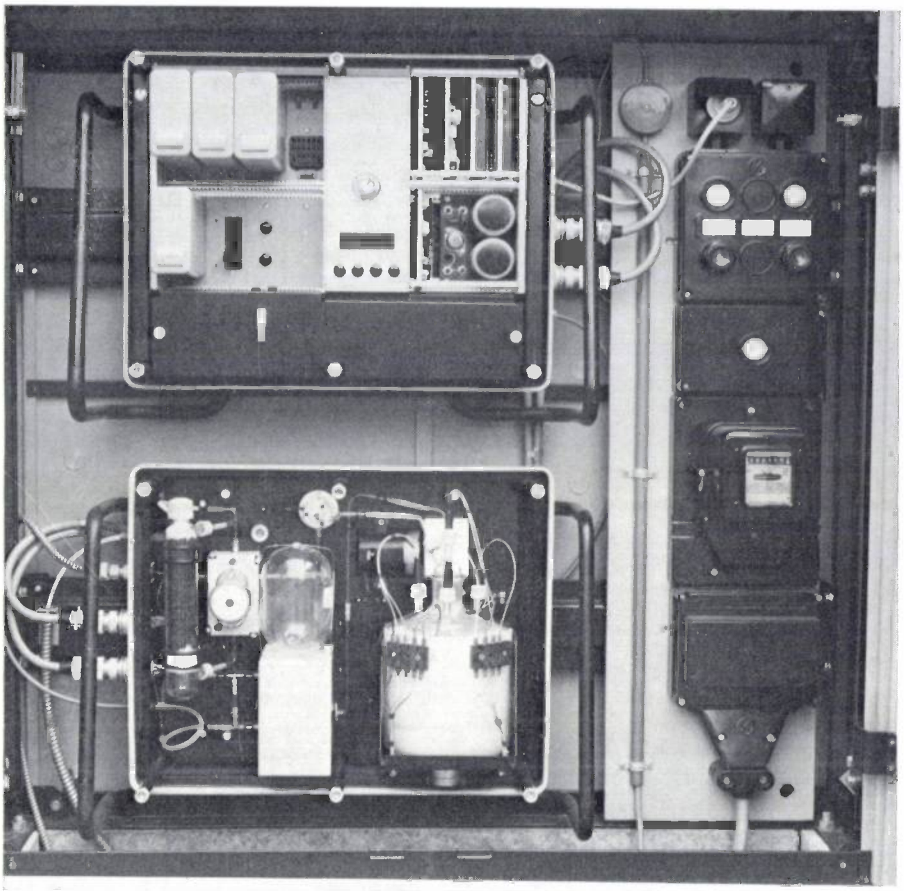


Fig. 3. View inside the SO_2 monitor. Below, the chemical compartment; above, the electronic compartment.

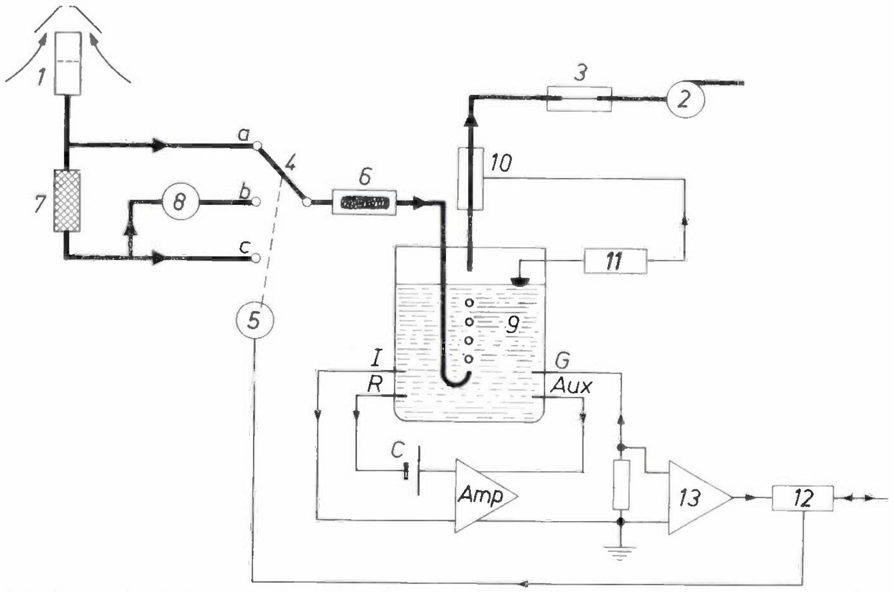


Fig. 4. Diagram of SO_2 monitor (see fig. 2). The pump *2* sucks in air, and the capillary *3* keeps the air current at a constant value. The filter *1* removes dust, insects, etc. from the air stream, and filter *6* removes many gaseous impurities in the air (except SO_2). On arrival in the measuring cell *9* the SO_2 in the air sample induces an electric signal in the cell, which is made compatible for transmission by telephone line by the output amplifier *13* and the telemetering unit *12*. The signals from the monitor are received in the central control room, where a computer is installed. Using the same telephone link the computer sends control signals to the motor *5*, which controls the positions of the three-way valve *4*. The valve is shown here in the measuring position (*a*). In the calibration and zero-mode positions of the valve the air passes over an activated-charcoal filter *7*, in which all SO_2 is adsorbed. In the calibration position (*b*) the SO_2 source *8* adds a known quantity of SO_2 to the SO_2-free air. By comparing the measuring signal with the calibration and zero signals the computer calculates the SO_2 concentration of the input air. *10* Peltier cooling element. *11* circuit for switching this element on and off (see text).

SO_2 or decomposition of dust particles at this temperature. Other measures to prevent adsorption of SO_2 are a nylon coating for the aluminium parts of the inlet and the use of a "Teflon" tube for transporting the air to the measuring cell. The pump *2* sucks in the air, while the capillary *3* ensures that the air flow remains constant at about 150 cm^3 per minute. The tube which conducts the air to the measuring cell contains a "Teflon" three-way valve *4*, whose position is controlled by a motor *5* for three modes of operation. In one position the system is in the measuring mode; the other two positions are for the "zero" mode of operation and for calibration. In the measuring mode (*a*) the air is conducted direct to the measuring cell through a filter *6*, consisting of a roll of silver gauze heated to about 120 °C. This filter reacts with or decomposes constituents that could interfere with the SO_2 measurement such as H_2S and ozone. In the zero and calibration positions the air passes over an activated-charcoal filter *7*, in which all the SO_2 is adsorbed. Moreover, in the calibration position (*b*) a constant quantity of SO_2 gas from a source *8* is added to the air stream. During regular maintenance this added SO_2 is compared with that from another SO_2 source, which in its turn is calibrated by means of a photometric determination with pararosaniline [4]. In the Rijnmond system the valve is switched to the different positions twice a day by telemetering from a central control room, the measuring signal then being compared with the zero signal and the calibration signal. *Fig. 5* shows a recording obtained in this way.

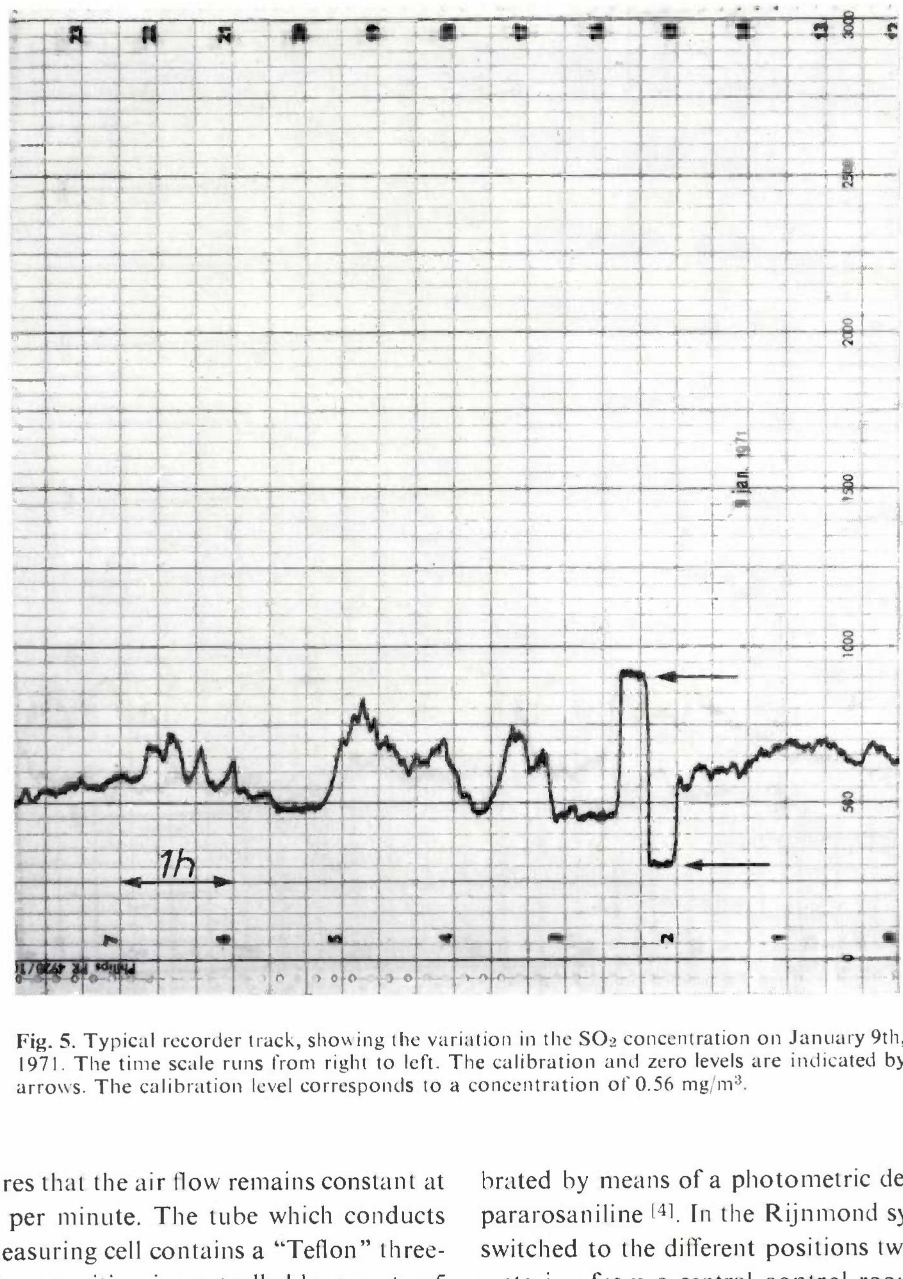


Fig. 5. Typical recorder track, showing the variation in the SO_2 concentration on January 9th, 1971. The time scale runs from right to left. The calibration and zero levels are indicated by arrows. The calibration level corresponds to a concentration of 0.56 mg/m^3.

The measuring cell *9* consists of an inner and an outer vessel (not shown in the diagram) each of which contains two electrodes (*fig. 6*). The air is passed to the inner vessel, where the principal reactions take place. Firstly, the SO_2 in the air is oxidized by the free

[4] F. P. Scaringelli, B. E. Saltzman and S. A. Frey, Anal. Chem. **39**, 1709, 1967.

bromine. Secondly a redox potential which depends on the bromine concentration is generated across the platinum electrode *I*. The corresponding electrode in the outer vessel is an Ag-AgBr electrode *R*, whose potential with respect to the electrolyte is practically constant since there is a large excess of the Br^- ions, which determine the potential of this electrode. The third important reaction in the inner vessel is the release of bromine from the electrode *G* as a result of an electrolysis current. The associated electrode *Aux* is in the outer vessel.

The inner vessel of the measuring cell is made fairly small, because the time constant of the control circuit that keeps the bromine content constant is determined by the volume of the electrolyte in this vessel. The inner vessel is connected to the outer one by a small aperture.

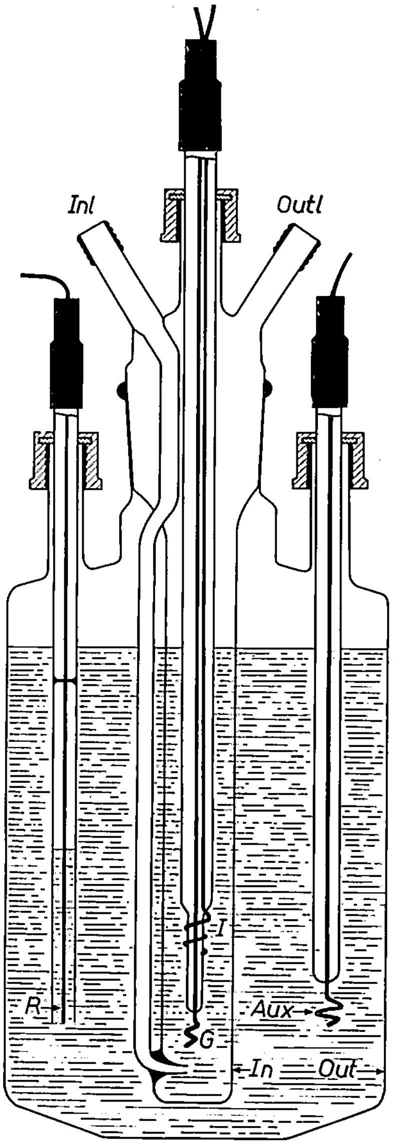


Fig. 6. The measuring cell. The inner vessel *In* contains the indicator electrode *I* and the generator electrode *G*; the outer vessel *Out* contains the reference electrode *R* and the auxiliary electrode *Aux*. *Inl* air inlet. *Outl* outlet aperture.

The electrolyte in the outer vessel dilutes any contaminants that may arise in or enter the inner vessel.

It is of the utmost importance to keep the temperature of the measuring cell constant. The temperature affects the measured potential differences, the volatilization and other parameters. The temperature chosen for the instrument described is 37 °C, which under normal circumstances is higher than ambient temperature. The measuring cell does not therefore have to be cooled, which would have required rather complicated and expensive measures. At the temperature selected the volatilization of bromine and evaporation of water is not excessive. The volatilization of bromine is automatically compensated by the control circuit. Thus, even when the air contains no SO_2, a small "zero current", flows between *G* and *Aux*. The SO_2 concentration is then measured as an increase above this zero current. The evaporation of water is compensated by means of a Peltier cooling element *10* situated in the air outlet of the measuring cell. As soon as the liquid level in the outer vessel falls below a critical value, the cooling element is automatically switched on. Water then condenses at this point in the line and drips back into the measuring cell.

We shall now take a closer look at the SO_2 calibration source *8*. This consists of a spherical glass reservoir filled with a mixture of sulphur dioxide and air at atmospheric pressure. The source has its outlet in a small capillary tube filled with silicone rubber. This substance is permeable to SO_2, so that there is a continuous flow of SO_2 out of the reservoir. The diameter and length of the capillary tube are matched to the required outflow rate. As this is slightly temperature-dependent, it is important to keep the source at constant temperature. After each three months' use the outflow rate shows a tendency to decrease, and the source has to be recalibrated. Since the continous outflow of SO_2 could cause an undesirable accumulation of contaminants in the instrument when the three-way valve is in the zero or measuring position, a flushing air-stream is passed through the system in these positions to remove the excessive SO_2 (not shown in the diagram).

The compartment with the electronic equipment contains the circuits *11* for switching the Peltier cooling element on and off and the controls for the thermostats, together with the other electronic circuits, the power transformer, the pump and the telemetering equipment *12*.

In the telemetering unit, a Philips Multi-Tone Telesupervision System, a signal is generated in which the magnitude of the output current from the amplifier *13* is coded in the form of a variable frequency [5]. In this form the signal can be transmitted by telephone line;

it only occupies a small part of the available bandwidth. The telemetering unit also contains two receivers, which are supplied with control signals from the central control room by the same telephone line. These receivers determine the position of the three-way valve *4*. The use of two receivers, which can be fed through two different channels, reduces the susceptibility of the system to any line interference there may be in the telephone link.

Before describing the monitoring network in the Rhine estuary, we shall first discuss the preliminary experiments with the monitors and the trials carried out with them, partly in the laboratory and partly in the Rijnmond region.

Table I. The ratio I_X of the measuring signal with 0.5 ppm of the substance X to the signal with 0.5 ppm SO$_2$. The dosing and calibration methods are also shown. (Dosing for the last four pollutants was done with a calibrated injection syringe, making calibration unnecessary. A description of the method of dosing and calibrating with ethylene will be found in the literature quoted.)

Pollutant X	I_X	Dosing	Calibration
Nitrogen dioxide	— 0.05	permeation source	photometric
Nitric oxide	0.00	,, ,,	,,
Ozone	0.00	UV irradiation	with SO$_2$ monitor
Ethylene [6]	0.02	—	—
Hydrogen sulphide	0.00	permeation source	with SO$_2$ monitor
Methyl mercaptan	1.8	,, ,,	photometric [7]
Benzene	0.00	injection syringe	—
Chloroform	0.00	,, ,,	—
Carbon disulphide	0.00	,, ,,	—
Propionaldehyde	0.01	,, ,,	—

Experiments, trials and initial practical experience

In the experiments on the measuring cells tests were first conducted to determine the extent to which they react to air pollutants other than SO$_2$. The interference they cause is expressed by I_X, the ratio of the measuring signal with 0.5 ppm of the substance X to the signal with 0.5 ppm of sulphur dioxide. The interference caused by a substance was investigated by adding a known quantity of that substance to an air-stream containing a known quantity of SO$_2$. The decrease of the SO$_2$ signal (for an oxidizing substance) or the increase (for a reducing substance) determines the magnitude of I_X. The results of these experiments, and the method of dosing and calibration used, are presented in *Table I*. The interference from most air pollutants, such as nitric oxide, nitrogen dioxide, ozone, hydrogen sulphide, ethylene, benzene and aldehydes is small. Appreciable interference is caused only by methyl mercaptan, but this is not important in practice since this compound usually only occurs in extremely low concentrations. The other substances investigated only interfere with the SO$_2$ measurement if their concentration is many times greater than that of the SO$_2$, which it very seldom is in practice.

We have now had more than four years experience with the SO$_2$ monitors, two of them in the Rijnmond area. The experience has been very satisfactory. The instruments have met the requirements for a constant zero current and reproducibility of the calibration signal. The accuracy of the monitors has been checked by determining the average daily values of the SO$_2$ concentration by the pararosaniline method [4]. The differences in the values found have never exceeded 8 %. During the entire period of the investigation the measuring cells have never shown any sign of cumulative contamination. After three months the dust filter at the top of the "sniffer" pole was still quite serviceable, in spite of considerable deposition of dust. After this time the air flow resistance had hardly increased at all, and analyses showed that scarcely any SO$_2$ had been adsorbed in the filter. One of the conclusions we came to was that simple maintenance servicing once every three months is in fact sufficient to guarantee reliable measurements of the SO$_2$ concentration in the atmosphere.

The monitor now in use has the following features. It can be set for various sensitivities, the minimum measuring range being zero to 0.1 ppm (0.3 mg SO$_2$/m^3) and the maximum range zero to 3 ppm (10 mg SO$_2$/m^3). The response time for a stepped change in the SO$_2$ concentration is about 1½ minutes; after this time the monitor indicates 63 % of the final value. The zero-point drift is less than 0.01 ppm per day (not cumulative), and the reproducibility of the calibration is better than 0.01 ppm. With its built-in calibrating facility the instrument can readily be calibrated once a day, and the accuracy of measurement depends mainly on the accuracy of the SO$_2$-source calibration performed by means of the pararosaniline method.

[5] J. J. Wilting, Ingenieur **82**, G 75, 1970 (No. 27) (in Dutch).

[6] H. W. Washburn and R. R. Austin, Proc. Nat. Air Pollution Symp. **1** (1949), 69, 1951.

[7] H. Moore, H. L. Helwig and R. J. Graul, Amer. Industr. Hyg. Assoc. J. **21**, 446, 1960.

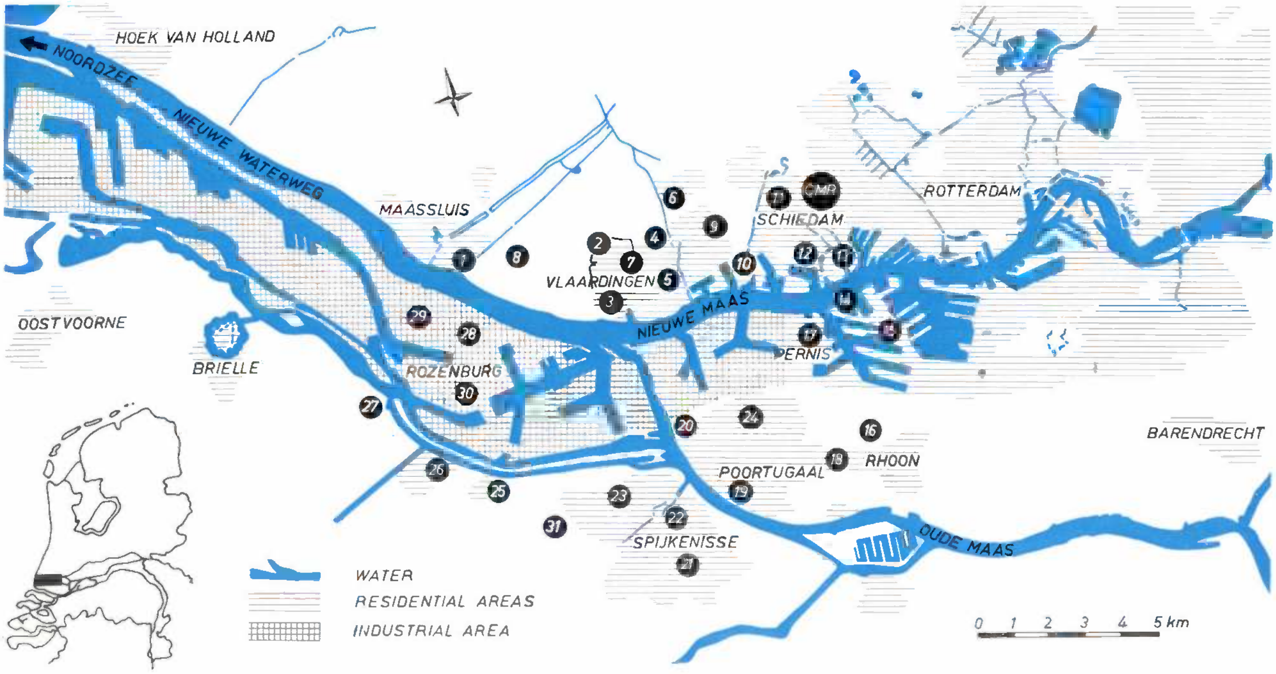


Fig. 7. Sites of the 31 SO_2 monitors in the Rijnmond area. *CMR* is the central control room.

The Rijnmond system

The monitoring network in the Rhine estuary consists of 31 monitors. This number was chosen on the basis of a statistical investigation carried out by Dr. L. A. Clarenburg of the Rijnmond Authority [8]. The monitors are distributed over the Rijnmond area as shown in *fig. 7*, and are connected to a Philips P9201 computer set up in the central control room at Schiedam. At intervals of a minute the computer scans all 31 monitors. The SO_2 concentrations are calculated from the measuring signals received and the zero signal and calibration values, which are stored in the computer memory. As already mentioned, the zero and calibration signals are recorded twice a day.

The computer performs various other calculations which serve in particular for correcting the measured concentrations for the effects of site, time, wind direction and season. To begin with, the hourly mean SO_2 concentrations $\bar{c}$ are determined. Next, another mean ($\bar{\bar{c}}$) is calculated from all the hourly mean concentrations $\bar{c}$ relating to the same site, the same place and the same wind direction. This is done using a sliding scale over a period of about three months, i.e. a weighted mean is calculated, using a weighting factor which is smaller the farther back one goes in time, and nearly zero when the time difference is more than three months. In this way a correction is made for seasonal effects.

The values $\bar{c}$ are divided by the values $\bar{\bar{c}}$ and the computer prints out the "delta values": $\Delta = (\bar{c}/\bar{\bar{c}}) - 1$. If Δ is greater than zero, then the concentration is greater than "normal" for the relevant site, time, wind direction and season. If the delta values, averaged over the whole monitoring network, rise above threshold values determined by statistical calculations, the computer triggers an internal alert (*fig. 8*). Should the local meteorological station expect the prevailing weather conditions to continue during the next six hours, an external alert is put out, warning all industrial plants in the Rhine estuary region by means of a permanent telecommunication link between the central control room and the individual industrial plants. If not only the delta values but also the actual measured concentrations exceed a certain threshold, a further warning is given to the industrial plants. Since the monitoring network was first put into operation in 1969 some twenty alerts have been given.

Following the first warning, plants suspend all odour-generating activities that are not directly connected with production, such as cleaning work like blow-through of pipelines and chimneys. After the second alert plants switch to the burning of low-sulphur fuels [9].

[8] L. A. Clarenburg, Precision of atmospheric sampling for air pollution levels in cities and in industrial areas, paper (No. 68-42) presented at the 61st Annual Meeting of the Air Pollution Control Association, St. Paul (U.S.A.), 1968.

[9] An extra warning has recently been given if complaints of smarting eyes and sore throats are received. Industrial plants are then requested to reduce to a minimum all activities that could release pollutants that have this effect.

The monitoring system is also used for tracing strong sources by means of cross-bearings, and trend measurements are made to determine the extent to which the average concentrations in the atmosphere show a tendency to change. These trends are not reflected in the Δ values but in the measured values themselves, which for this purpose are corrected for the meteorological conditions.

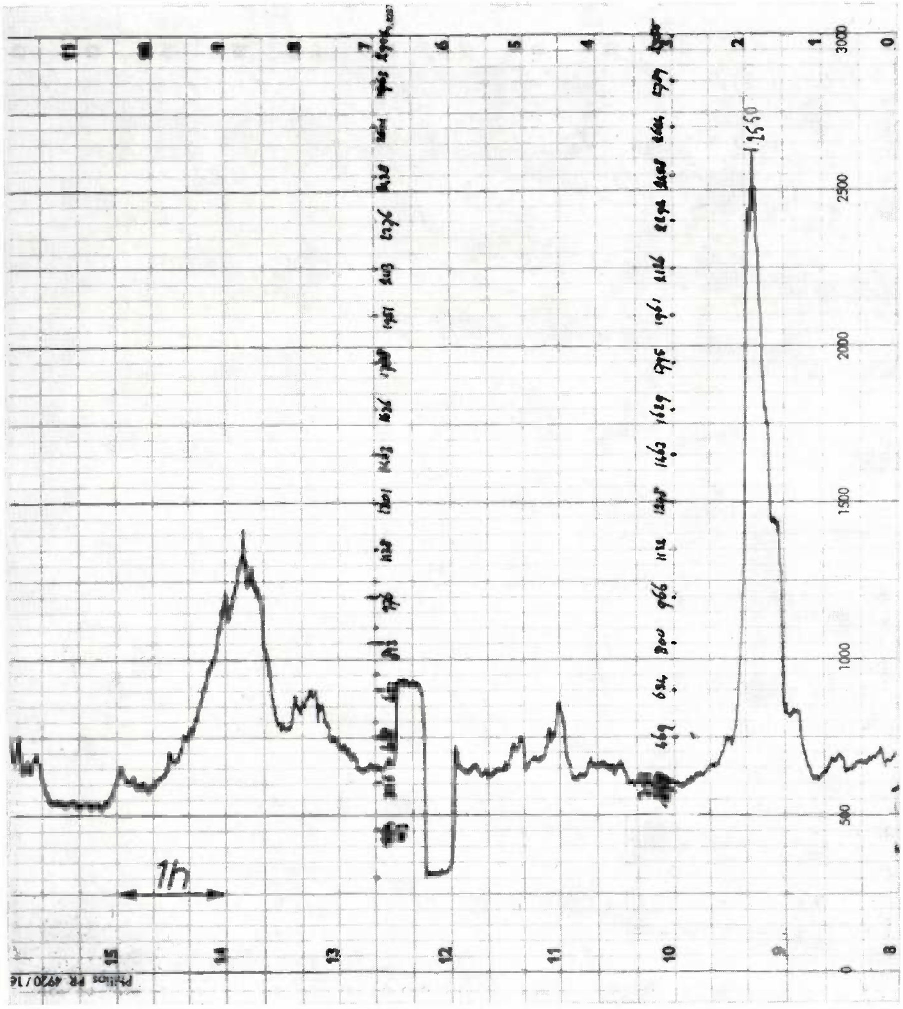


Fig. 8. The variation in SO_2 concentration of Februry 11th, 1971, a day on which a first warning was given. The two vertical rows of figures are calculated values of the concentration in $\mu g/m^3$; the right-hand row is based on a previous zero level and calibration level, and the left-hand one is based on the zero and control levels seen in the trace.

Summary. The article describes a network for monitoring SO_2 concentrations in the Rhine estuary region. The network consists of 31 monitors, connected by telephone lines to a Philips P9201 computer. The monitors operate without supervision, and require simple servicing only once every three months. The monitor works on the coulometric principle. A constant air stream is passed through a measuring cell containing a solution which includes free bromine. The sulphur dioxide in the air reacts with the bromine. The bromine content is kept constant by a feedback system in which the bromine used up by the SO_2 is replenished by means of an electrolysis current. This current is a measure of the quantity of SO_2 supplied per unit time. Every minute the computer receives a measuring signal from the monitor. It can also receive a calibration signal and a zero signal. The position of a three-way valve in the monitor determines which signal is received; the positions of the valve are controlled by the computer. The computer calculates the SO_2 concentration by comparing the measuring signal with the calibration signal and the zero signal. The computer also carries out various calculations for correcting the measured concentrations for the effects of site, time, wind direction and season. If the values thus corrected, averaged over the whole Rijnmond area, rise above a threshold value determined by statistical calculations, an alert signal is put out.

Transmission of simple pictures

L. G. Krul and P. Reijnierse

Some offices are already using an internal television system for verifying cheques, but the transmission costs soon become high if such a system is to be linked up with other offices. However, since the pictures to be transmitted are quite simple — numerals and signatures — systems of smaller bandwidth can be used making transmission over existing telephone links financially more attractive. The authors have made a study of a method of transmission that could be used for pictures of this type; it requires a bandwidth of no more than 10 kHz.

To transmit a signal from one place to another a transmission channel of adequate bandwidth is needed. If the bandwidth is too small the signal received will be distorted. In carrier telephony, for example, a signal with frequencies between 300 and 3400 Hz can be transmitted over a single telephone channel. This is sufficient for transmitting speech but it is not enough for music. For transmission of music a bandwidth of at least 15 kHz is required, and in fact a music channel has a bandwidth of 16 kHz, so that it occupies as much of the frequency band as four telephone channels (the total bandwidth of a telephone channel is 4 kHz).

The transmission of a 5 MHz television signal over telephone circuits obviously requires a much larger number of channels. Using microwave links 960 telephone conversations can be transmitted in a frequency band of 4 MHz, or one television signal in a band of 6 MHz. In the 12 MHz system, using coaxial cables, three groups of 900 conversations can be transmitted in a frequency band of 12 MHz, or one group of 1200 conversations in a frequency band from 300 Hz to 6 MHz plus one television signal in the band between 6 and 12 MHz. This means that the television signal is taking the place of no less than 1500 telephone channels. For video transmission it is therefore important to investigate methods that can work with a smaller bandwidth.

One possible answer is the "slow-scan" method, in which the whole picture is not, as in television, scanned in 1/25 second but in say 20 seconds. The result of this is that the required bandwidth is reduced by a factor of 500 from 5 MHz to 10 kHz. A disadvantage is of course that it now takes 20 seconds to build up a complete picture, so that a camera or a storage tube has to be used to enable the picture to be seen as a whole.

If the picture contains less information than an ordinary television picture (so that it is coarser in structure and gradation) then there are fewer details to be transmitted, and a smaller bandwidth can therefore be used without lengthening the scanning time. This is the case with the kinds of picture that this article is concerned with: simple line drawings, numerals, letter characters and signatures. These could be transmitted in a readily recognizable form by a system working with fewer than the usual 625 picture lines.

We have investigated another method, which is suitable only for transmitting these simple pictures. This method records only the transitions between light and dark, which are well defined here and contain the essential information for pictures of this type. With this technique a bandwidth of 10 kHz has been found to be sufficient, while the scanning time per picture is only 2/25 second — twice as long as in conventional systems.

Principle of the method

To explain our method we shall begin with the simple example shown in *fig. 1*. This picture was taken with a television camera scanning horizontally from left to right, but without the interlacing normally used in television. A complete scan therefore consists of half the normal 625 lines and lasts 1/50 second. The figure shows two transitions from light to dark, *a* and *c*, and two from dark to light, *b* band *d*. From now on we shall refer to light-dark transitions as "positive" and to dark-light transitions as "negative". The positive transitions are numbered in the scanning direction in ascending order 1^+, 2^+, 3^+, etc., the negative transitions 1^-, 2^-, 3^-, etc. This procedure leads to the numbering shown in fig. 1 and in the rather more complicated example of *fig. 2*. In the second figure the numbering of transition *b* changes at several places; at a dashed line transition 1^- always changes to transition 2^-, or

L.G. Krul and P. Reijnierse are with Philips Research Laboratories, Eindhoven.

vice versa. This is no disadvantage, however, as we shall shortly see.

In our transmission method the *transitions* are reproduced at the receiving end as *lines*. *Fig. 3* and *fig. 4* are the displays corresponding to fig. 1 and fig. 2 when *all* the positive and negative transitions are transmitted. As we shall see, each transmitted transition requires a complete field scan. The time needed to transmit a picture therefore increases with the number of transitions to be transmitted. If the picture is a figure consisting solely of lines, then it is sufficient to transmit only one of the two transitions (a positive and a negative) yielded by each line, since each *transition* at the transmitting end becomes a complete *line* at the receiving end. The time necessary for transmitting a figure is thus halved.

As we said earlier, our work was mainly directed towards the transmission of signatures and numerals. Now when a line of handwriting is scanned *perpendicular to the direction of writing*, it is very unusual to find more than four intercepts with the lines that form the characters. In the experimental arrangement that we have built the number of transitions to be transmitted is therefore limited to four. The transmission of a complete picture then takes four times 1/50 second, i.e. 80 milliseconds. We have found that it is important to always transmit the first white-to-black transition (1^+) and the last black-to-white transition; these form the "envelope" of the figure when they are combined. Since the last black-to-white transition consists of differently numbered sections (see fig. 2), we introduce a separate notation, l^-, for this transition.

Fig. 5a shows a script letter *e* and the transition numbering that results when this letter is scanned vertically. Fig. 5*b* shows the picture that results when the transitions 1^+, 2^+, 3^+ and l^- are transmitted in a cycle of four

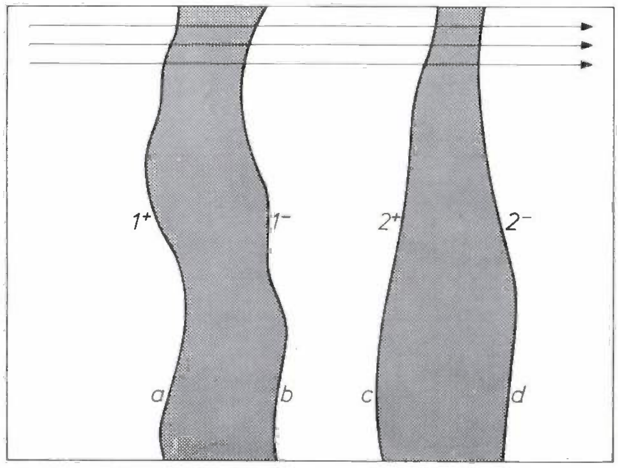


Fig. 1. The numbering of the light-to-dark transitions for an arbitrary black-and-white figure. The arrows indicate the direction in which the television camera scans the picture. The light-to-dark transitions *a* and *c* are indicated as positive, 1^+ and 2^+, the dark-to-light transitions *b* and *d* as negative, 1^- and 2^-.

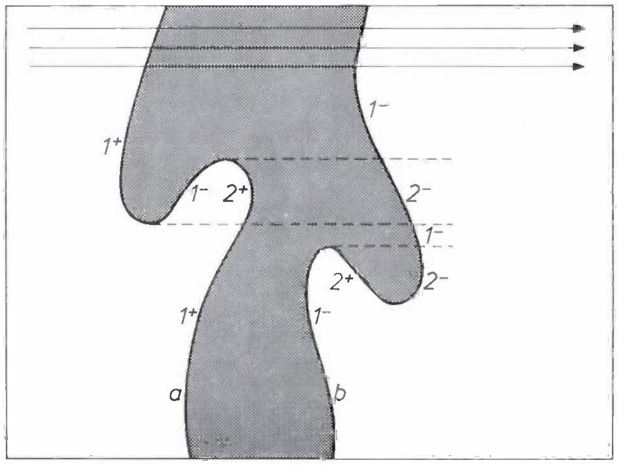


Fig. 2. Numbering of the transitions for a rather more complicated figure than in fig. 1. It can be seen here that parts of one transition can have different numbers, depending on the location relative to the scanning beam. The dashed lines indicate places where the numbering changes.

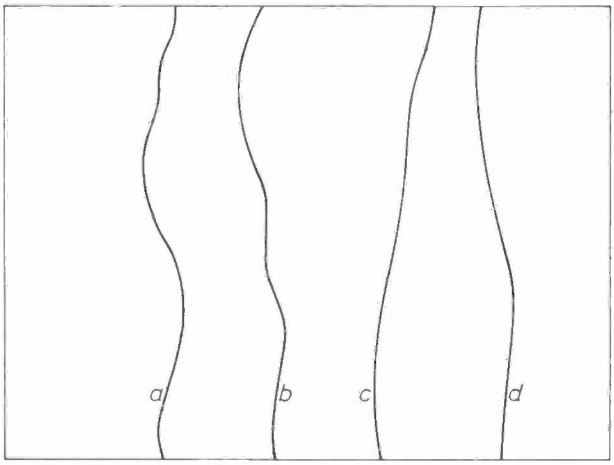


Fig. 3. Reproduction of fig. 1 at the receiving end. The light-to-dark transitions are displayed as lines.

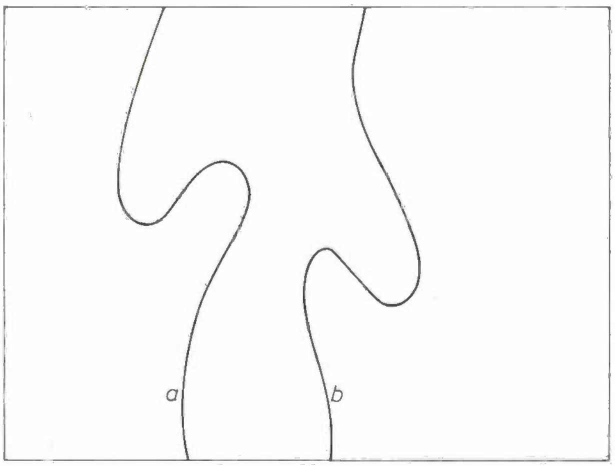


Fig. 4. As fig. 3, but now for the picture in fig. 2. The change in the numbering cannot be seen in the picture.

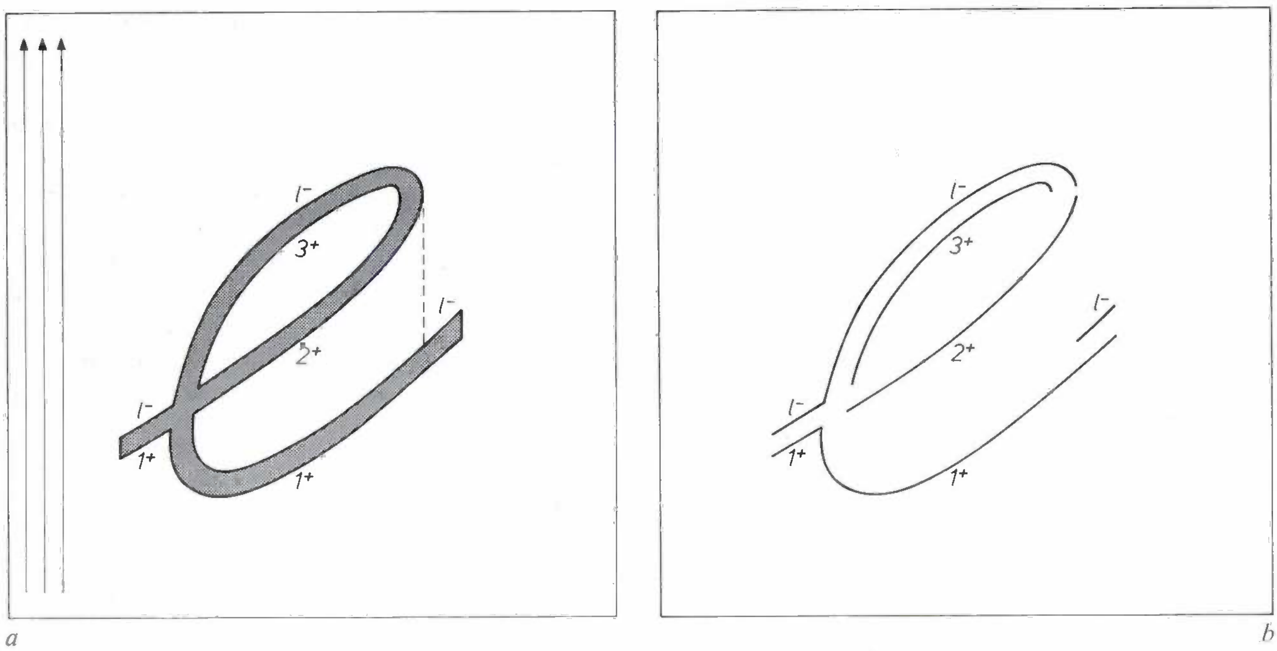


Fig. 5. *a*) A script letter *e* with the numbering for the light-to-dark transitions. The scan is now vertical, from the bottom upwards. *b*) Display of the letter when the transitions 1^+, 2^+, 3^+ are transmitted, together with the last black-to-white transition 1^- (here composed of the transition 3^- and parts of the transition 1^-). When the thickness of the lines is small compared with the size of the letter, the double lines will merge.

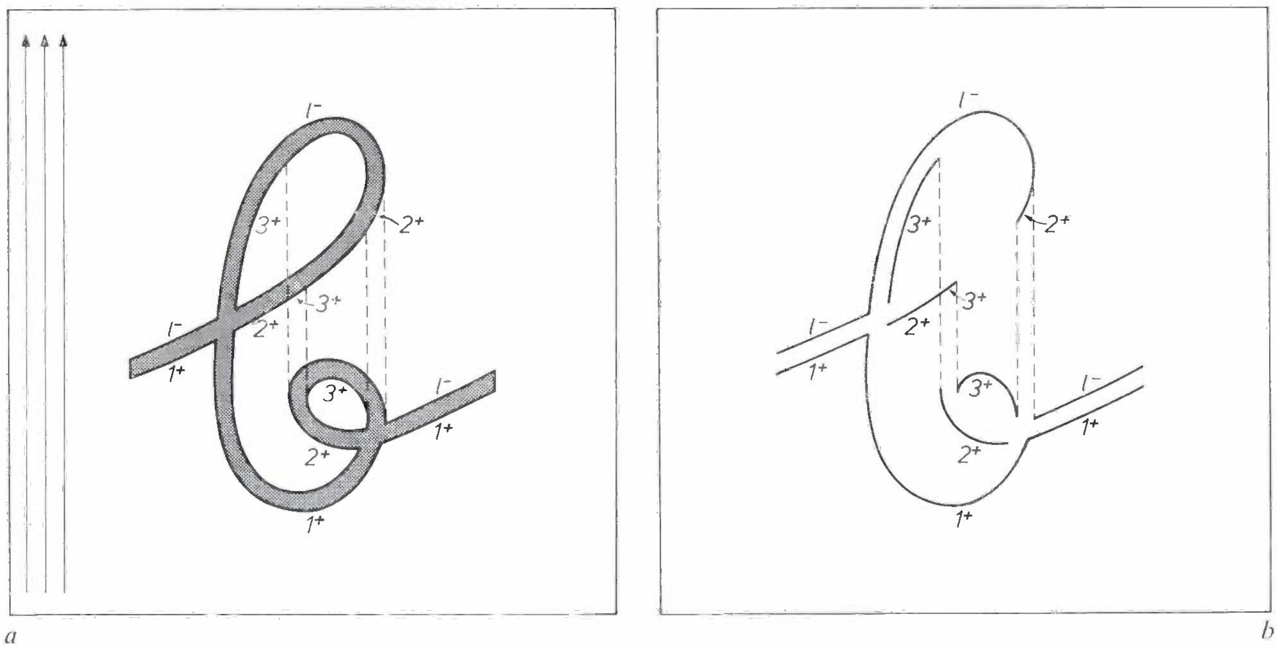


Fig. 6. As fig. 5, but now with a script letter *b*. Because only four transitions are transmitted, part of the information is lost. Nevertheless the representation is readily recognized as a letter *b*, particularly if the double lines merge.

successive scans. It can be seen that these four scans produce a very clear picture of the original letter. If the line of the letter is very thick, as here, the picture will partly consist of double lines. If, however, the thickness of the letter is small compared with its dimensions, which is usually the case in handwriting, the double lines will tend to merge at the receiving end. As a second example *figs. 6a* and *b* show the situation for the letter *b*. We see here that because of the limitation to four scans a few parts of the transitions are not reproduced; nevertheless the result is still recognizable as a letter *b*.

There are two techniques for determining the positions of the transition points to be transmitted, an analogue method and a digital method. In the analogue method a sawtooth voltage is used, which begins at zero at the start of every scanning line. The value of this voltage at the instant that the electron beam passes a transition point is a measure of the position of that point.

In the digital method a counter is used which is reset to zero at the beginning of each line and which counts up to say 256 in the duration of a line. The number in the counter at the instant when the scanning beam passes a transition point establishes the position of that point. This information is transferred by means of a number of registers to a digital-analogue converter, which produces an analogue signal again. In this article we shall confine ourselves to the analogue method.

The equipment

The picture is scanned by a simple black-and-white type camera (*Cam* in *fig. 7*), which is merely rotated through 90° to give a vertical scan. The video signal *vid* is passed through a level detector *LD* and a differentiator *Di*, which extract only the transitions (both positive and negative) from the original picture information. Voltage pulses indicating these transitions appear at the outputs *P* and *N*. Normally these outputs are at the voltage level 1, but when the scanning beam in the camera passes a positive transition, *P* switches briefly to the level 0, and so does *N* when a negative transition is passed.

The circuit in fig. 7 contains a field counter *FC* and a transition detector *TD*. The field counter has four outputs, corresponding to the four picture scans of a cycle. At the beginning of a cycle the field-synchronizing pulse *fs* sets all outputs of *FC* to zero except the output corresponding to the transition to be measured, which is at the level 1. The transition detector has three outputs, which correspond to the transitions 1^+, 2^+ and 3^+. At the beginning of each line these outputs are set to 1 by the line-synchronizing pulse *ls*. When the scanning beam in the camera passes the transitions 1^+, 2^+ or 3^+, the outputs 1^+, 2^+ or 3^+ are momentarily switched to 0. To describe the other parts of the circuit we shall take two different picture scans by way of illustration.

We begin with the case of the fourth scan, i.e. the one for the last negative transition 1^-. In fig. 7 the levels 1 or 0 that are present when a line scan begins are shown next to the outputs of the counters *FC* and *TD*. We see that the AND gates *B*, *C* and *D* all have a 0 at their input connected to the field counter. Consequently these logic circuits are blocked during the whole of the fourth scan. The transition detector and the output *P* connected to it for the positive transitions therefore have no effect. The OR gate *E* has a 0 at all its inputs and hence at its output. The lower input of the OR gate *F* is therefore also at 0. The AND gate *A* has a 1 at both inputs and thus delivers a 1 at its output. This arrives at the upper input of the OR gate *F*, so that the output of *F* is at 1.

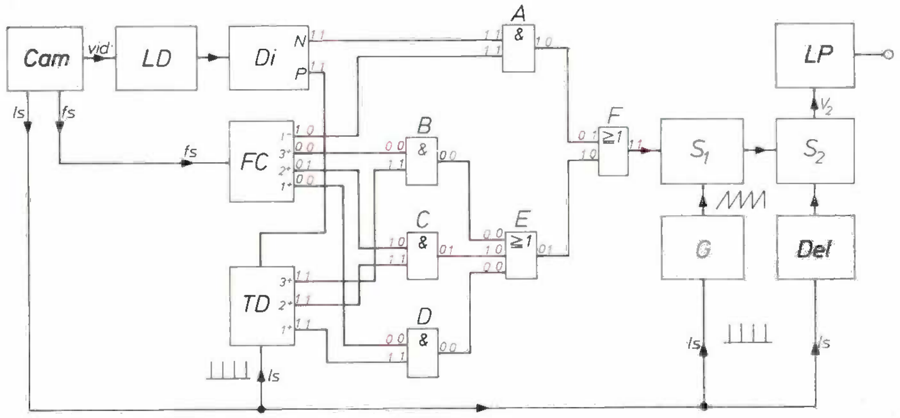


Fig. 7. Diagram of the equipment at the transmitting end. *Cam* camera. *vid* video signal. *ls* line synchronizing pulse. *fs* field synchronizing pulse. *LD* level detector. *Di* differentiator. *FC* field counter. *TD* transition detector. *A*, *B*, *C* and *D* are AND gates. *E* and *F* are OR gates. S_1 first sampling circuit, S_2 second sampling circuit. *G* sawtooth generator. *Del* delay circuit. *LP* low-pass filter. The numbers 1 and 0 in black at the various outputs and inputs correspond to the situation at the beginning of the fourth picture scan, and the numbers in red correspond to the situation at the beginning of the second picture scan.

We have already seen that positive transitions can produce no response. A negative transition, on the other hand, brings the output *N* — and hence the upper input of the AND gate *A* as well — momentarily to 0. As a result the output of *A* is momentarily at zero level as is also the output of the OR gate *F*. This output pulse from F now triggers the first sampling circuit S_1. At the instant that a pulse arrives from F this circuit measures the instantaneous value of a sawtooth voltage which it receives from the generator G and which is triggered by the pulse *ls*. The sampling circuit contains a storage device (in this case a capacitor), which always stores the last measured voltage value. At the end of a line scan it therefore contains the value which the sawtooth voltage had at the instant when the scanning beam crossed the last negative transition. Just before the end of a line scan the voltage value contained in the storage capacitor of the first sampling circuit is taken over by a second sampling circuit S_2, which also contains a storage capacitor. This stores the voltage sample until the next line has been scanned. The second sampling circuit is triggered by the pulse *ls* via a delay circuit *Del* with a delay time of 60 µs. During the fourth picture scan a voltage that follows the position of the transition l^- on the successive scanning lines in a series of steps therefore appears at the output of the second sampling circuit. If at the receiving end this voltage is applied to the vertical deflection of an oscilloscope whose time base is driven by a sawtooth voltage at 50 Hz, we obtain a picture which is a faithful copy of the transition

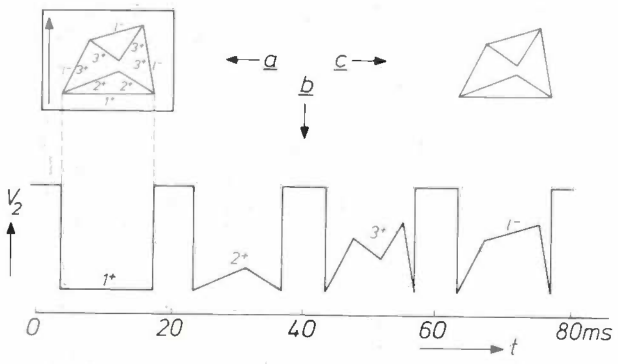


Fig. 8. During the scanning of the figure (*a*) the second sampling circuit S_2 (see fig. 7) delivers an output voltage V_2 (*b*) from which the figure is reconstructed at the receiving end (*c*). The reconstructed picture is seen to be identical with the transmitted picture.

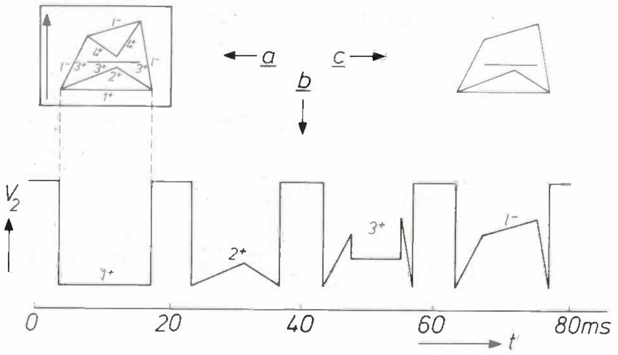


Fig. 9. As fig. 8, but because only four transitions are transmitted (*c*) is now not identical with (*a*). The line indicated by the transition 4^+ is not reproduced.

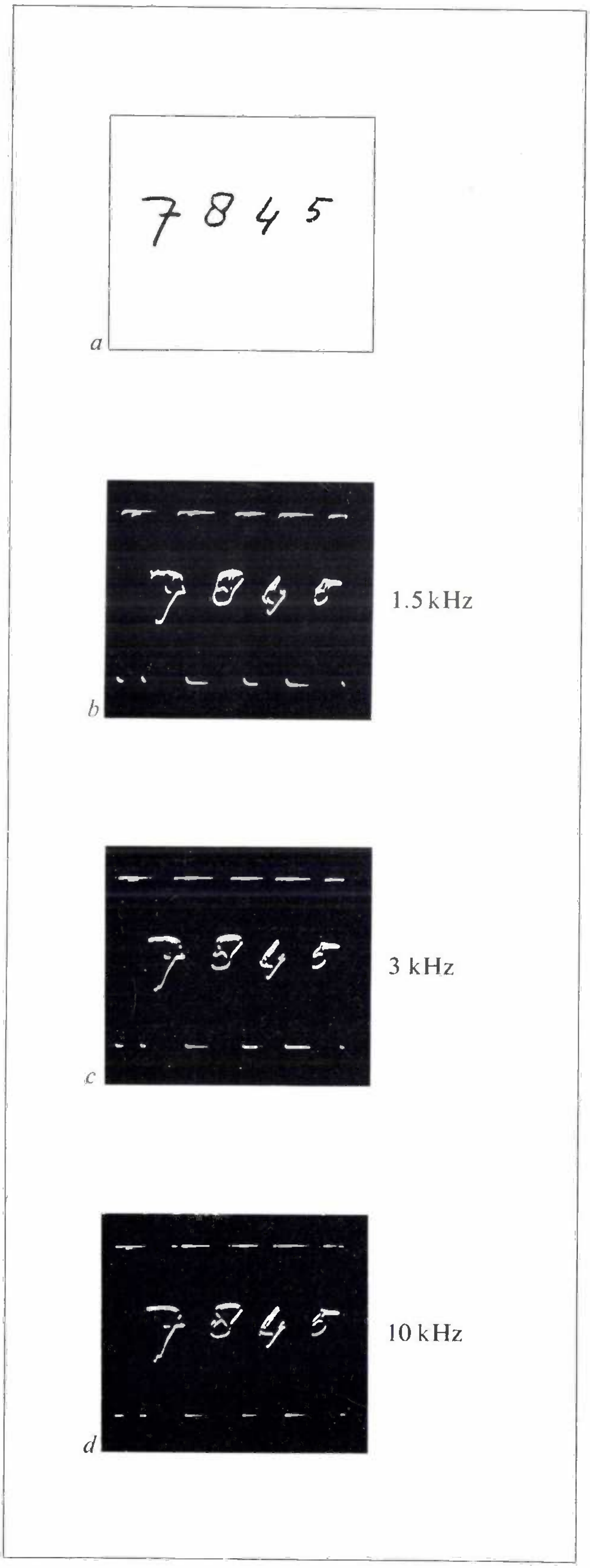


Fig. 10. Some handwritten numerals, of which (*a*) shows the original, transmitted at different bandwidths, 1.5, 3 and 10 kHz. The resultant pictures at the receiving end are (*b*), (*c*) and (*d*). The double lines, as shown in fig. 5*b*, can clearly be seen. For these very simple pictures a bandwidth of 3 kHz would be just sufficient for a recognizable display.

1^-. The low-pass filter *LP* cuts off the higher frequencies generated in the stepwise shifts of the voltage value.

As an example of a case where a positive transition is to be reproduced, we shall consider the scan during the second field, during which the transition 2^+ is recorded. The levels that are now present when a line scan begins are indicated by the red numbers in fig. 7. The "zeros" at the outputs 1^+, 3^+ and 1^- of the field counter block the AND gates *A*, *B* and *D*; the only way in which the OR gate *F* can be affected is through the AND gate *C*. This happens when a 2^+ transition is passed, when output 2^+ of the transition detector will go momentarily from 1 to 0. The outputs of gates *C*, *E* and *F* take this pulse over, causing the first sampling circuit to trigger. The voltage passed to the second sampling circuit at the end of each line scan then corresponds to the instant at which the 2^+ transition is passed. The output voltage of the second sampling circuit now gives a faithful copy of the 2^+ transition. The same reasoning applies to the transitions 1^+ and 3^+.

Figs. 8 and *9* show two examples of line figures (*a*), with the associated output voltages V_2 from the second sampling circuit as a function of time (*b*), and the pictures reproduced at the receiving end (*c*). It can be seen from the second example that when only four fields are used, part of the information (here the transition 4^+) is lost when the figure is too complicated.

In the curves (*b*) of figs. 8 and 9 it can be seen that the sections indicating the transitions are separated by flat sections in which the voltage has a constant high value. These flat sections arise because the scanning electron beam encounters no transition at these places, so that the output voltage of S_1 rises to the peak value of the sawtooth voltage. These voltage peaks can appear on the display as short lines, but they are above the picture and do not interfere with it. The voltage transients before and after these horizontal lines can interfere with the picture, however, particularly if the low-pass filter *LP* has a small bandwidth. To avoid this, measures have been taken to suppress the electron beam in the display tube when the first time derivative of the signal to be displayed exceeds a particular value. Because of this very steeply rising lines in the picture cannot be reproduced. This does not matter, however, since in any case only one or two points on these are recorded by the vertical scan.

Some results

Fig. 10 and *fig. 11* give two examples of picture transmission with our experimental arrangement. The numerals in fig. 10*a* are reproduced in fig. 10*b*, *c* and *d* for three different cut-off frequencies of the low-pass

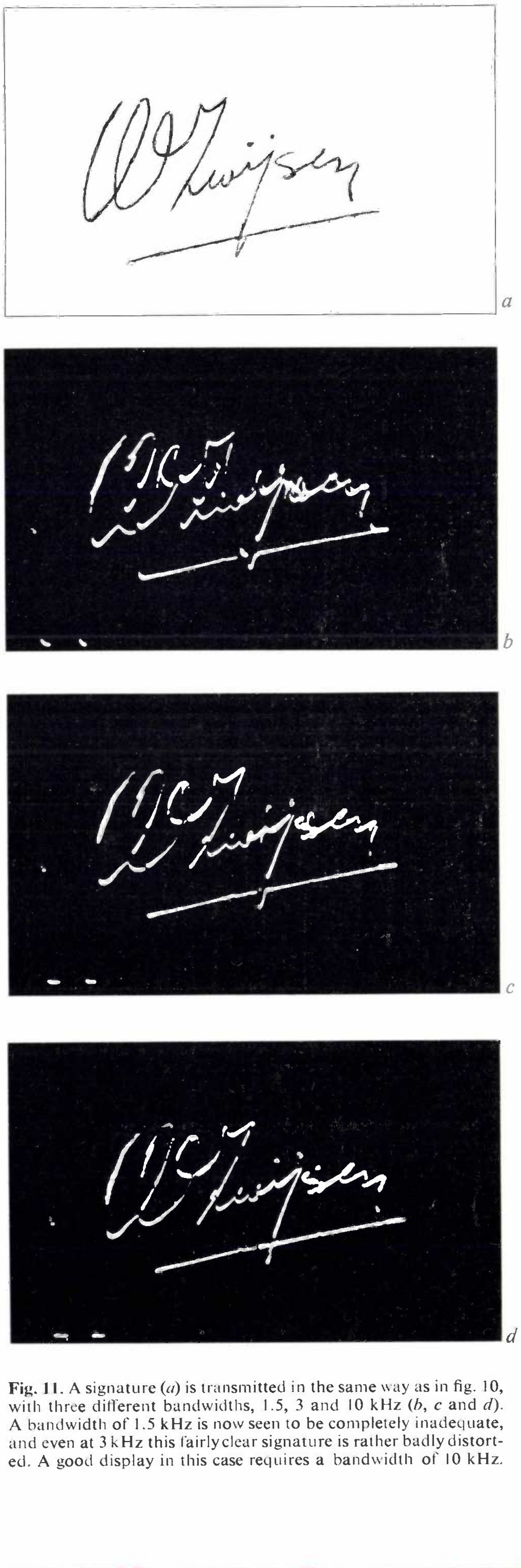


Fig. 11. A signature (*a*) is transmitted in the same way as in fig. 10, with three different bandwidths, 1.5, 3 and 10 kHz (*b*, *c* and *d*). A bandwidth of 1.5 kHz is now seen to be completely inadequate, and even at 3 kHz this fairly clear signature is rather badly distorted. A good display in this case requires a bandwidth of 10 kHz.

filter *LP*, 1.5, 3 and 10 kHz. (The double contours pointed out in fig. 5 can clearly be seen here.) The signature in fig. 11*a* is reproduced in the same way in fig. 11*b*, *c* and *d*. These examples show that simple pictures of this kind can be transmitted well by a signal with a bandwidth of 10 kHz, provided of course that the phase-response characteristic of the transmission channel is made sufficiently linear.

To compare this result with the bandwidth that would be needed to transmit figures of this kind with an ordinary television signal, we must make an estimate of the number of picture lines required to obtain an equally good transmission. Because of the thin narrow shape of most signatures we have not used the whole picture area in our experimental arrangement. With the vertical scan the raster would normally be greater in height than width, but we have only used a horizontal strip extending to one-third of the total height. The result is a raster with an aspect ratio of 2 : 1, scanned along 300 vertical lines, with a vertical line resolution of about 100. We should therefore compare our system with an ordinary television system using 100 picture lines, an aspect ratio of 2 : 1 and a transmission time per picture of 80 ms (this being the duration of our four scans). Such a system requires a bandwidth of about 125 kHz; compared with this the method described here, which uses a bandwidth of 10 kHz, gives a gain of at least a factor of 10.

The gain is nevertheless not sufficient to allow the signal merely to be transmitted via a telephone channel in a carrier system. A channel of this type handles a frequency band of no more than 300 to 3400 Hz, and moreover its phase-response characteristic is not good enough for our purpose. If it is desired to use existing telephone lines for long-distance transmission it will therefore be necessary to have a frequency band amounting to three telephone channels. The situation is more favourable if a permanent telephone line can be used, for example between a central bank and a branch office in the same town. Here again it may happen that the bandwidth in some part of this link is limited; but if a fixed coaxial pair is available, the complete signal can be transmitted along it, so that a single telephone line is sufficient for the purpose.

Summary. Simple pictures whose essential information consists of a limited number of light-to-dark transitions can be transmitted by a signal of much smaller bandwidth than an ordinary television signal. The method described here is of particular interest for handwriting, e.g. written numerals and signatures. These pictures are scanned by a television camera in the direction perpendicular to the line of writing. When a light-to-dark transition is passed by the scanning beam, its position is converted into a voltage level. A cycle of only four scans, each scan recording one particular transition, is shown to be sufficient for a satisfactory display. The transmission of the signal requires a bandwidth of only 10 kHz. Over short distances it will usually be possible to transmit the signal by a permanent telephone line; for longer distances existing carrier telephone links can be used, but in that case a frequency band equal to three telephone channels will be needed.

Miniature refrigerators for electronic devices

A. Daniels and F. K. du Pré

The miniature refrigerators presented here are the latest offshoots of the Philips family of Stirling-cycle machines. They are an alternative to the classical liquid gas cryostats, and are specially attractive for cooling purposes outside the laboratory. Their main application is at present in the cooling of infra-red detectors for field use.

Introduction

At very low temperatures the physical properties of materials are often significantly different from those encountered at normal ambient temperatures. In recent years this has opened the way to new or improved electronic devices. These devices sometimes require liquid nitrogen temperatures or lower, while the cooling power for the smaller devices where virtually no electric energy is dissipated is about 1 W. Typical applications are: infra-red detectors, masers, lasers, superconducting devices, etc. Numerical data and references are given in *Table I* for a number of applications.

Infra-red detectors are at present the most important of the cooled electronic devices and the rapid progress made in their use during the past decade has had a major impact on the development of miniature refrigerators. We shall start therefore by looking briefly at the necessity for cooling these detectors and then give an outline of the Stirling process and the specific requirements for miniature refrigerators. Next we shall look at the details of two types of miniature refrigerators (with working temperatures of 25 K and 77 K). Finally we shall deal with two experimental machines on which research has been done in our laboratory.

Infra-red detection

Let us consider the cooled infra-red detectors that are sensitive to the thermal radiation from our surroundings. This radiation has maximum intensity at a wavelength of about 10 μm. The detectors are used to obtain photographs, often from the air, without the need for visible light [4]. Such photographs can be taken in daytime or at night.

In infra-red quantum detectors, whether photoresistors or photodiodes, the effect of the radiation is to excite free charge carriers from shallow energy levels in the band structure of the detector material. At room temperature a considerable concentration of free carriers is produced by thermal excitation from these levels. For optimum detector sensitivity the noise in the thermal excitation should be small compared with the radiation-induced excitation. Therefore, the detector must be cooled until the thermal excitation and consequently the noise in this thermal excitation is reduced to a sufficiently low level.

Table I. Electronic devices that are either improved by cooling or cannot work without cooling. The refrigerating power needed is given as "small" when only 1 W is needed to compensate insulation and conduction losses for small electronic components. It is given as "appreciable" when the load is 10 W or more.

	Temperature needed	Refrigerating power needed	References
Solid-state maser	4 K	appreciable	[1]
Superconducting computer components	4 K	appreciable	[2]
Superconducting dynamo	4 K	appreciable	[3]
Infra-red detectors	25-77 K	small	[4]
GaAs junction laser	77 K	small	[5]
Superconducting Josephson junction	4 K	small	[2]

The temperature required to realize such background-noise-limited infra-red detectors depends on the excitation energy for free carriers in the material by long-wavelength radiation. As this excitation energy varies considerably from material to material, the temperature required can be quite different for different materials.

Refrigerating machines

Originally in the laboratory ice, dry ice and liquid air were used for cooling purposes. These coolants were

A. Daniels, M.S., and Dr. F. K. du Pré are with Philips Laboratories, Briarcliff Manor, N.Y., U.S.A.

[1] J. C. Walling and F. W. Smith, Philips tech. Rev. 25, 289, 1963/64.
[2] R. A. Kamper, Cryogenics 9, 20, 1969.
[3] J. Volger, Philips tech. Rev. 25, 16, 1963/64.
[4] F. Desvignes, J. Revuz and R. Zeida, Philips tech. Rev. 30, 264, 1969, and M. Jatteau, Philips tech. Rev. 30, 278, 1969.
[5] C. H. Gooch (editor), Gallium arsenide lasers, Wiley, London 1969, page 278.

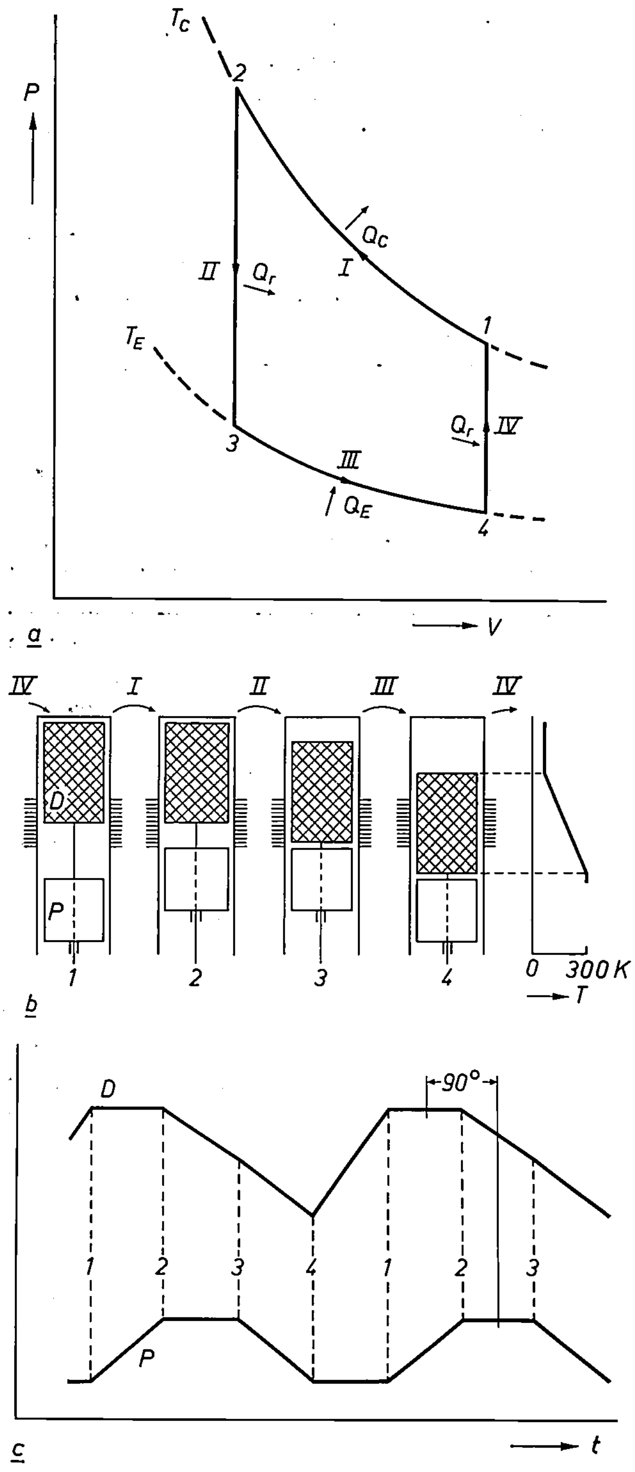


Fig. 1. *a*) Pressure and volume variations in the ideal Stirling cycle (p-V diagram). In the isothermal compression (phase *I*) a quantity of heat Q_C is removed from the working gas, and the amount Q_E is absorbed by the gas during the isothermal expansion in phase *III*, thus cooling the load. In the isochorous (equal-volume) cooling phase *II* the heat Q_r is stored in the regenerator; it is reabsorbed by the gas in the isochorous phase *IV*. *b*) Position of the compression piston *P* and the displacer *D*, containing the regenerator, in the cylinder at the points *1-4* in (*a*) on the ideal Stirling cycle. The phases *I-IV* are indicated. The temperature distribution along the cylinder is indicated on the right; the temperature-gradient part moves up and down with the regenerator inside the displacer. *c*) When the positions of piston and displacer are plotted as a function of time their movements are clearly seen to be 90° out of phase.

produced in large installations; transport and storage caused inconvenience and losses.

The development of relatively small cold-producing units based on the Stirling cycle was a major advance in cryogenics; liquified gases could now be produced on the spot and in the quantities needed. In the Stirling process [6] a fixed amount of gas (the working gas, as a rule He) is subjected to a thermodynamic cycle that can be divided into the four phases shown in *fig. 1a*: compression at room temperature (*I*), cooling to the working temperature (*II*), expansion at the working temperature (*III*), and finally reheating to room temperature (*IV*). Cold is produced at the expansion in phase *III*.

The working gas is taken through this cycle by the periodic reciprocating movements of a piston *P* (fig. 1*b*) and a displacer *D*. The piston alternately compresses the gas and allows it to expand again. The displacer, by forcing the gas through a porous high-heat-capacity structure (the regenerator), brings about the cooling and reheating. The movements of piston and displacer must approximately have a 90° phase difference (fig. 1*c*).

The first Stirling-cycle refrigerating machine developed and produced within the Philips group of companies some 20 years ago had a cold production of 1 kW at a temperature of about 80 K. From this design both larger [7] and smaller machines have been developed later on, as well as machines capable of attaining substantially lower temperatures [8].

Miniature refrigerators

The first miniaturized Stirling refrigerator prototype was designed and made by Dr. J. W. L. Köhler, Ir. G. Prast and their co-workers at Philips Research Laboratories in 1959. The unit produced about 1 watt of cold at 30 K. Its characteristics demonstrated the feasibility of a miniature Stirling refrigerator.

Concurrently with the demonstration of this prototype, the need arose in the United States for a small refrigerator capable of cooling infra-red detectors that were sensitive in the 8 to 14 μm spectral band to a temperature of 25 K. A refrigerator for this purpose had of course to be compatible with the system into which it is to be integrated. This meant that apart from the fact that it had to be small, it had to be light in weight, efficient and reliable. In most applications the refrigerator had to be able to work in any attitude. The machine should also be reasonably quiet and it should not cause interference in the associated electronic systems.

Several changes to the original prototype resulted in the refrigerator shown in *fig. 2*, sold as the "Cryogem" by the Cryogenic Division of the United States Philips Corporation.

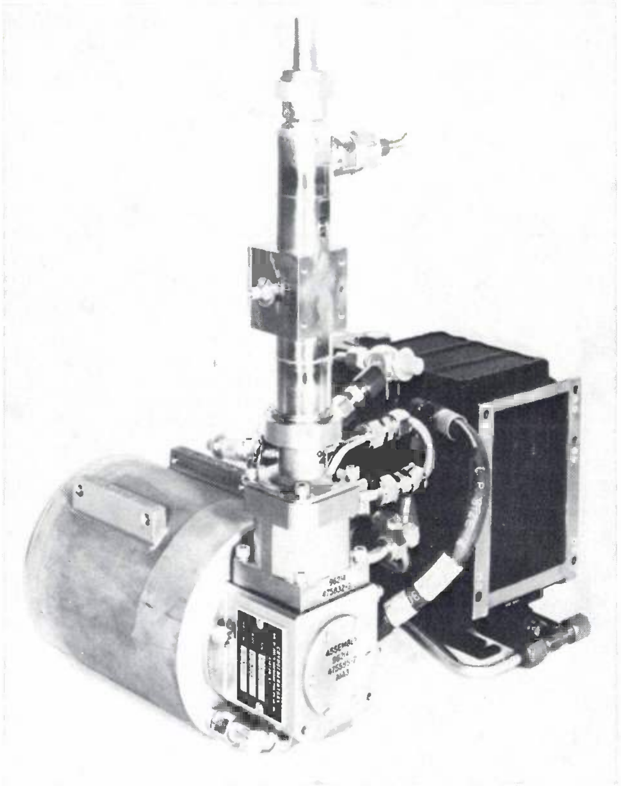


Fig. 2. The "Cryogem" 1 W, 25 K Stirling-cycle refrigerator. The motor and crankcase are shown on the left. The vertical structure is the working cylinder; the lower part is the heat dissipating region and the upper part, enclosed in a vacuum jacket for insulation, is the cooled end. The heat generated in the lower part of the cylinder is dissipated to the atmosphere by the heat exchanger on the right.

The "Cryogem" and "Micro Cryogem" miniature Stirling refrigerators

It would have been sufficient just to scale down the larger refrigerator for the applications mentioned above, but many of the other requirements were such that it was better to make considerable changes at a prototype stage.

There are two major differences between the 1 kW machine and the miniature version. The regenerator, which in the large machine is housed in channels arranged concentrically around the displacer cylinder, was put inside the displacer in the small ones. This makes the regenerator design somewhat less adaptable but results in a simpler and more compact machine configuration. The large machine is driven by a motor outside the unit through a shaft seal. The small refrigerators were designed to have the motors inside the crankcase, i.e. with both the stator and the rotor in the working gas. This avoids loss of the working gas by leakage through the shaft seal and, again, permits a more compact design.

The lubrication of cryogenic equipment presents a special problem: lubricating oil, or its vapour, has to be prevented from reaching the low-temperature areas, as it would freeze there and reduce the efficiency of the unit. The precautions taken in the large machines in order to avoid contamination cannot be realized in the miniature machine, because of the much smaller dimensions. In addition, the miniature machine is much more sensitive to disturbances than the 1 kW version is. Finally, since the small refrigerators have to work in any attitude, no oil could be used, otherwise lubricant could enter the critical working spaces.

To solve this problem we have made use of reinforced "Teflon" bearings and seals, which are self-lubricating and may, consequently, run essentially dry.

Two versions of the miniature Stirling refrigerator were designed both operating at 25 K: one was a single-expansion machine, the other a double-expansion machine. Essentially, in a double-expansion machine there is an extra expansion volume where cold is produced at an intermediate temperature between room temperature and the lowest temperature. This cold is then used to reduce most of the regenerator losses. In a single-expansion machine this is done entirely with cold generated at the lowest temperature. Since it is more economical to produce cold at an intermediate level [9], the process on the whole is more efficient [10].

The double-expansion miniature 25 K Stirling refrigerator is also manufactured under the trade name "Cryogem". Its salient characteristics are shown in *Table II*. The electric motor, for reasons mentioned above, is an integral part of the refrigerator; its shaft also serves as the crankshaft of the refrigerator drive. The refrigerator is equipped with a closed-cycle heat exchanger (see fig. 2), which operates in a manner

Table II. Numerical data for small refrigerators

	"Cryogem"	"Micro Cryogem"	
	25 K double-expansion machine	77 K single-expansion machine	77 K double-expansion machine
Refrigeration capacity (W)	1	1.2	0.75
at temperature (K)	25	77	77
Minimum temperature (K)	17	55	42
Cool-down time (min)	3	2.7	10
Power input (W)	475	40	27
Weight (kg)	9.0	2.5	2.3

[6] J. W. L. Köhler and C. O. Jonkers, Philips tech. Rev. **16**, 69 and 105, 1954/55, and J. W. L. Köhler, Progress in Cryogenics **2**, 41, 1960.

[7] A. A. Dros, Philips tech. Rev. **26**, 297, 1965.

[8] G. Prast, Philips tech. Rev. **26**, 1, 1965.

[9] J. D. Fast, Philips tech. Rev. **16**, 298, 1954/55, in particular equation (II, 33).

[10] Thermodynamical details of a double expansion Stirling refrigerator have been treated in the paper by G. Prast [8].

similar to an a automobile cooling system: liquid is pumped through the heat-extraction passages of the refrigerator, then transferred to a heat exchanger where the energy is dissipated into the atmosphere.

Recently infra-red detectors have been developed which, for background-noise-limited detection, only require cooling to temperatures of about 70-80 K. For cooling such detectors it is, of course, possible to use an open-loop cryogenic system, consisting of a liquid-nitrogen storage dewar and coolant transfer lines. However, the necessity to have liquid nitrogen available is sometimes undesirable and the use of a closed-cycle refrigerator instead of an open-loop system is then more appropriate.

When the Cryogem refrigerator, designed for operation at 25 K, is adjusted to operate at a temperature of 75-80 K, about 12 W of cold is produced. Since for cooling small electronic devices less than 1 W is needed, a special version had to be designed for operation at about 77 K. In fact, two designs evolved, one with single expansion, the other with double expansion. The salient parameters of these units both sold under the name "Micro Cryogem" are shown in Table II. This 77 K refrigerator is provided with cooling fins through which the heat of compression is transferred by free convection to the ambient atmosphere. This simplification is possible since these units have a relatively low power consumption (40 watts maximum).

Refrigerator with rhombic drive

The trend in recent infra-red imaging systems is to use a large number of small detectors arranged in a mosaic, instead of one single detector. In this case the mechanism used to scan the object can be much simpler than for a single detector. Obviously this puts stronger emphasis on the absence of mechanical vibrations in the refrigerator on which such a mosaic is mounted. The conventional crank-mechanism drive, which can never be perfectly balanced, does not fulfill these requirements [11].

Here we can use with advantage the rhombic drive developed some fifteen years ago to give fully balanced operation of a Stirling machine. This drive is powered by two identical and counter-rotating motors. In *fig. 3* it is demonstrated that the linkages in this drive if suitably dimensioned convert the rotations into two reciprocating movements with a 90° phase difference, as required for driving the piston and displacer in a Stirling machine. By placing suitable counterweights on shafts and linkages, the center of gravity of all the moving parts can be kept stationary [12]. Furthermore, because of the symmetry there is no resultant angular momentum. Since the refrigerator has both a stationary center of gravity and compensating angular momen-

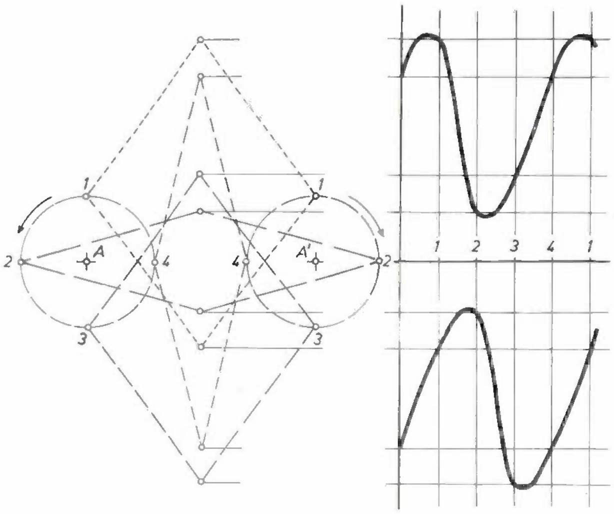


Fig. 3. The operation of the rhombic drive. Two counter-rotating cranks driven at *A*, *A'*, move two opposite vertices of a rhombus. As a result the rhombus is alternately squeezed together and pulled out while at the same time it moves up and down. This results in periodic motions of the other two opposing vertices showing a phase difference of 90° when the linkages are properly dimensioned, which can consequently be used for driving the piston and displacer in a Stirling-cycle refrigerator.

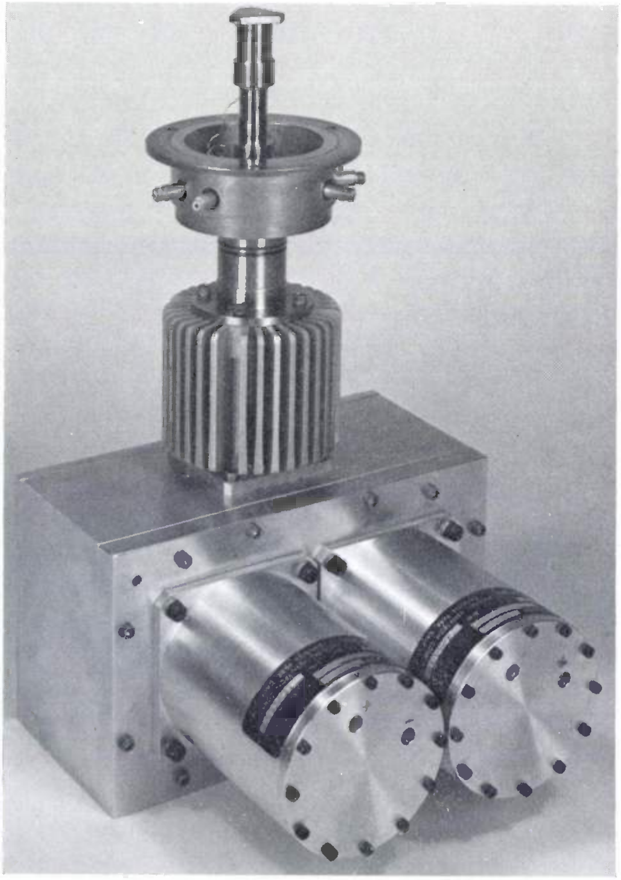


Fig. 4. Vibration-free Stirling-cycle refrigerator with a rhombic drive mechanism. The two identical motors powering the drive are inside the cylinders seen in front. The top of the machine is normally enclosed within a vacuum chamber for thermal insulation; part of this vacuum chamber has been removed for this photograph.

tums, it will introduce no vibration. If such a refrigerator were mounted in a satellite it would cause hardly any unwanted motion even during starting or stopping, which would not be the case for conventional motor-driven devices. In actual practice of course minute vibrations remain, due to imperfections in the materials and to manufacturing tolerances. In *fig. 4* is shown a double-expansion 77 K machine equipped with the rhombic drive mechanism. Thermodynamic specifications for this refrigerator are identical with those of the double-expansion "Micro Cryogem" (Table II, last column).

Thermal coupling of detector and refrigerator

Among the problems that had to be solved before miniature refrigerators could be integrated into electronic systems were the method of coupling the detectors, or other elements to be cooled, to the cold area of the refrigerator, and the thermal insulation of these elements. The main factors to be considered for optimum interfacing are: the heat conduction loss via the electrical leads connecting the detector(s) with the electronic system, the rigidity of these leads and the maintenance of the vacuum provided to give thermal insulation.

A typical detector-cooler interface is shown in *fig. 5*. For simplicity only one detector is shown, but units with more than 100 detectors have been made. In spite of the large temperature gradient along the leads connecting the cold detector to the outside of the vacuum chamber, the thermal load can be kept below 1 W at the operating temperature by proper choice of conductor material and geometry.

Microphony due to minute vibrations that remain even after careful balancing, can be reduced by making sure that the leads to and from the detector cannot move. When the refrigerator is running, the noise level is only 10 % higher than when it is switched off.

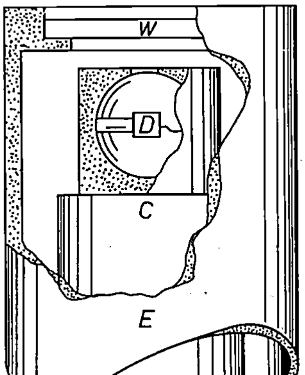


Fig. 5. Detector-cooler interface. The detector *D* is attached to the cold area *C* of the refrigerator by low thermal resistance components. The detector is surrounded by a reflector to concentrate the incident radiation. The space inside the envelope *E* is evacuated to give good thermal insulation. The envelope has an infra-red transparent window *W*.

To give good thermal insulation the detector and the cold surface of the refrigerator are surrounded by a vacuum that has to be better than 10^{-2} N/m^2 to be effective. It takes careful working to maintain this vacuum since the vacuum chamber contains several demountable joints and a large number of electrical feedthroughs that have to be sealed against either atmospheric pressure or the working pressure of the refrigerator. Another complication is that the detector material cannot usually be baked out at temperatures high enough for proper degassing.

It may happen that elements have to be cooled at a spot remotely located from the cold area of the refrigerator. It may also be necessary to allow for relative movement between the detector and the refrigerator (as for example with a detector mounted on gimbals): direct coupling is then impossible.

We have built a liquid-coolant transport system based on the Leidenfrost effect to solve these problems. Small drops of cryogen (liquid nitrogen) are introduced into a stream of gas flowing through uninsulated flexible tubes. The drops slowly evaporate during this transport and consequently they are surrounded by a relatively cold, isolating sheath of gas. Should a drop come too close to the wall, then the evaporation increases and a gas cushion is formed, preventing direct contact and the accompanying loss of cold. The cooling system is a closed circuit one: cryogen drops evaporate on cooling the load, the gas is returned to the refrigerator by a small compressor and reliquefied. The drops can only be transported over a few metres in this way, as the isolation obtained is never perfect.

The Leidenfrost transfer system, with a Stirling-cycle cooler to provide the liquid nitrogen, has about 10 % of the efficiency of one of the 77 K refrigerators with direct-contact cooling.

Triple-expansion Stirling-cycle refrigerator

As we indicated above, the presence of an extra expansion volume, at an intermediate temperature, reduces the heat leak into the low-temperature volume and thus it is possible to obtain lower temperatures. Small double-expansion refrigerators have reached temperatures of about 17 K (Table II) while the larger version reached 12 K [4].

It seems an obvious step to increase the number of expansion volumes to try to reach still lower temperatures with the Stirling cycle. This led us to the con-

[11] See for example page 280 of the paper by M. Jatteau [4].
[12] R. J. Meijer, Philips tech. Rev. 20, 245, 1958/59.

struction of an experimental refrigerator with two intermediate expansion volumes, i.e. a triple-expansion unit.

The construction is shown in *fig. 6*. As before, there is one compression piston and one displacer, but the diameter of the displacer is now reduced in two stages, in order to produce two extra expansion volumes. The three regenerators are inside the displacer. The two regenerators that operate at intermediate temperatures can be constructed in the usual way from fine copper gauze, but as a material for the regenerator operating at the lowest temperature this is unsuitable.

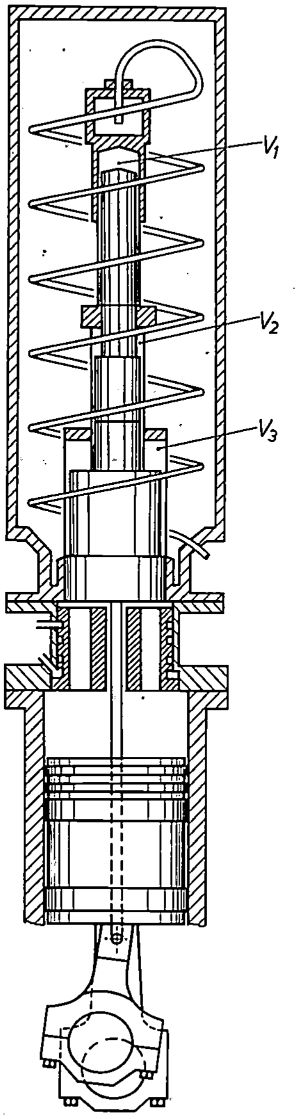


Fig. 6. Cross-sectional view (schematic) of a triple-expansion Stirling-cycle refrigerator. Apart from the main expansion volume at the top of the cylinder there are two extra expansion volumes V_2 and V_3 along the cylinder, where cold is generated at temperatures between room temperature and the operating temperature of the machine at the cylinder head.

The regenerator matrix must have a high heat capacity compared with that of the gas flowing through in each cycle. But, since the matrix is porous and contains a certain amount of gas in the interstices, the matrix heat capacity must also be high compared with that of the gas contained in it for the following reason. Some of this gas is being compressed in the regenerator while it flows towards the cold space, and it would be undesirable if its temperature were to rise because of this compression. More precisely, we require the heat capacity per cm³ of the regenerator material to be high compared with the heat capacity per cm³ of the helium gas.

The problem that this requirement causes at temperatures below 10-20 K is clearly seen from *fig. 7*, which shows the heat capacity per cm³ as a function of temperature of lead, copper and europium sulfide, three possible matrix materials, and also the heat capacity per cm³ of helium gas at 4×10^5 N/m² (4 atm), a reasonable working pressure for a triple-expansion Stirling machine. From these curves it is clear that copper is not a suitable material for a very-low-temperature regenerator. Lead can be used to temperatures in the neighborhood of 10 K; it is in fact the most suitable of the metals. Both copper and lead show the sharp decrease of the heat capacity with decreasing temperatures that is characteristic of a solid material. Europium sulfide also shows this general trend, but superimposed on it is a peak due to the transition to the ferromagnetic state at about 16 K [13]. Because of this peak, europium sulfide seems to be more suitable than lead at very low temperatures.

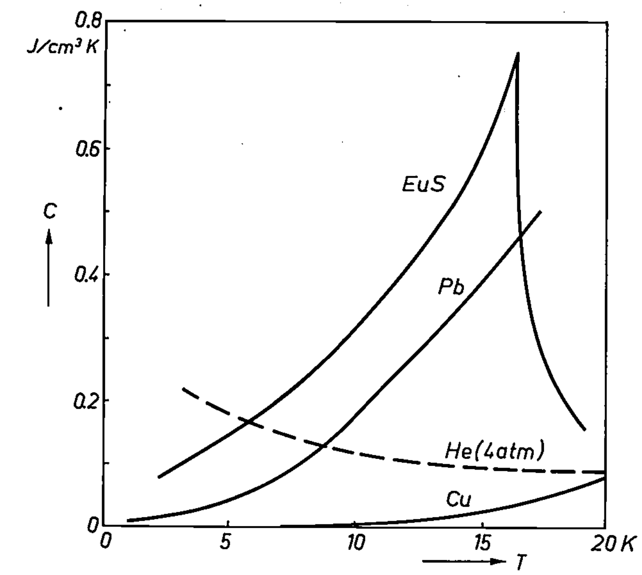


Fig. 7. Heat capacity per cm³ (C) as a function of temperature (T) for a number of substances that can be used as regenerator materials. For comparison the heat capacity of helium at the average working pressure of a refrigerator has also been plotted.

In view of these data, two types of low-temperature regenerator were tried. The first one contained only lead spheres of about 0.1 mm dia. The lowest temperature obtained was about 9 K. The second regenerator had its lower half filled with the same lead spheres, but the top half was filled with a fine powder of europium sulfide. Here the lowest temperature obtained was slightly less than 8 K. This is the lowest temperature reached so far with the Stirling cycle. The helium pressure was about 4×10^5 N/m^2 (4 atm), the operating speed 600 rev/min. Further progress may have to wait for the development of better regenerator materials. The use of ^{3}He might also help since it has a lower heat capacity than ^{4}He.

It might appear that better results could be obtained by using a lower helium pressure, thus reducing the heat capacity of the helium in the interstices. However, with lower helium pressure the cold production is also diminished and the working temperature, since it adjusts itself as the equilibrium between cold production and cold leak, does not go down.

The cylinder *A* is similar to the one in a Stirling refrigerator. We have the expansion volume at the top and the displacer containing a regenerator. In the Vuilleumier cycle the function of the compression piston is taken over by the cylinder *B* and its displacer. The top of this cylinder is kept at a high temperature (e.g. 1000 K). If the "hot" displacer is "down", much of the helium is in the hot area, the average helium temperature will be high, and the pressure will thus be high everywhere in the working space.

If the "hot" displacer is "up", very little of the helium is in the hot area, the average helium temperature will be low and the pressure will be low. Therefore, when the "hot" displacer moves up and down in the correct phase relationship to the motion of the "cold" one, the same pressure and volume variations are produced in the expansion space of cylinder *A* as in a Stirling-cycle machine and thus cold is produced there. The Vuilleumier refrigerator can also be considered to consist of two Stirling machines, one a motor and the other a refrigerator, placed back to back. The pressure variations in the motor part are not used to drive a piston but are transferred directly to the refrigerator part. The refrigerator thus has a thermal compressor instead of a mechanical one. As here the heat of both the motor and the refrigerator has to be removed the total heat output from a Vuilleumier refrigerator is considerably larger than from a comparable Stirling refrigerator.

The Vuilleumier refrigerator produces cold directly out of heat and should be useful whenever the energy supply is available in the form of fuel. The power needed to run the displacers can be kept small since, apart from frictional forces, the only forces on the

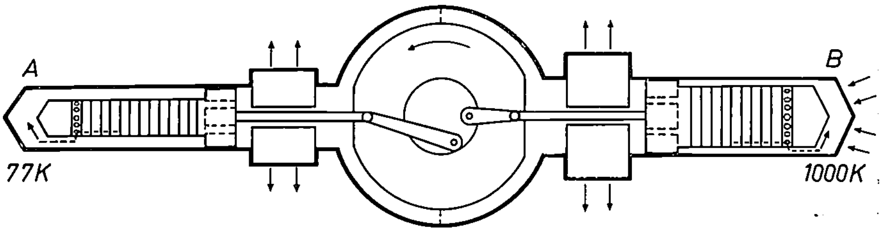


Fig. 8. Cross-sectional view (schematic) of a Vuilleumier refrigerator. The cylinder *A*, with its displacer containing a regenerator, is identical to the cylinder of a Stirling machine. The cylinder *B* is also equipped with displacer and regenerator; the top of this cylinder is kept at a high temperature (e.g. 1000 K). The reciprocating motion due to a small electric motor of the displacer in the cylinder *B* modulates the mean temperature of the working gas, thus providing a thermal compressor that replaces the piston compressor in a Stirling machine.

Vuilleumier refrigerator

Finally, we shall discuss the results obtained with a small experimental refrigerator whose principle was invented in 1918, the Vuilleumier refrigerator [14].

Here, unlike the situation in the Stirling machine the main energy needed for the production of cold is supplied in the form of heat and only a small amount of mechanical energy is needed.

Fig. 8 shows a schematic diagram of our version of the refrigerator proposed by Vuilleumier. The heavy outer line shows the working volume, filled with helium gas. It consists of two cylinders and a connecting space. Two displacers, driven by a small electric motor, reciprocate in the cylinders with a relative phase difference of about 90°. Both displacers have a regenerator inside, the construction being the same as in the Stirling-cycle case. There are no valves.

[13] V. L. Moruzzi and D. T. Teaney, Solid State Comm. **1**, 127, 1963.

[14] U.S. Patent 1,275,507 of Aug. 13th 1918; G. K. Pitcher and F. K. du Pré, Adv. cryog. Engng. **15**, 447, 1970.

displacers are due to the pressure drop in the helium flowing through them. As the forces on the moving parts are very small the wear should also be small, especially if the operating speed is low. Another consequence of the small mechanical forces is that operation is quiet and vibration-free.

The compression ratio is low in a Vuilleumier refrigerator. A typical value is 1.4, whereas in Stirling refrigerators a ratio of 2.5 is not unusual. The low compression ratio is due to the thermal compression method, which is hampered by the fact that materials limitations allow only moderately high temperatures to be used on the hot side. Moreover, the hot regenerator adds considerably to the dead space in the working volume. To reduce the effect of a small compression ratio, we have chosen the average operating pressure rather high, so that considerable variations in pressure could still be obtained. As cold production for a given expansion volume is determined by the pressure variations only, it is again similar to that of the Stirling cycle.

Fig. 9 shows a photograph of a small Vuilleumier refrigerator. The hot displacer is on the right, the cold one on the left. The small motor that drives the two displacers is inside the pressurized crankcase, and so cannot be seen in fig. 9. To facilitate measurements of the power input, an electrical heater was used to provide the heat input. This Vuilleumier machine produces 1.25 W of cold at 77 K.

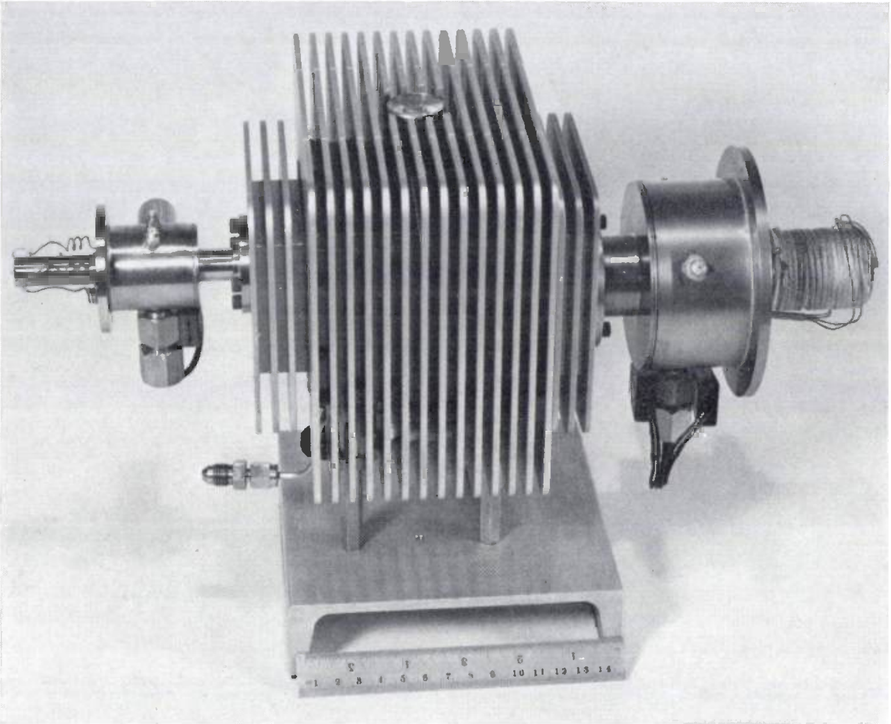


Fig. 9. An experimental Vuilleumier refrigerator. The cold cylinder is on the left. The hot cylinder at the right is heated electrically to simplify efficiency measurements. Any other method of heating can be used, provided it yields sufficiently high cylinder-head temperatures.

Summary. Miniature versions of the 1 kW Philips Stirling refrigerator have been derived, which have working temperatures of 25 K and 77 K. These machines, designed for cooling infra-red detectors and other components in electronic equipment, produce about 1 W of cold. One version, with a rhombic drive mechanism, is completely vibration-free and therefore specially suited for cooling infra-red detector arrays. For cooling of movable objects a cryogen, such as liquid nitrogen, is transported through uninsulated tubes. The Leidenfrost effect prevents direct contact of the cold fluid and the relatively hot tube wall. Finally two experimental machines are described: a three-stage machine (cold production 1 W at 9 K) and one with a thermal compressor (the Vuilleumier refrigerator) producing about 1 W of cold at 77 K by direct conversion of heat into cold.

Sorting of quartz tubing for high-pressure mercury-vapour lamps

One of the parameters that has to be kept constant in the manufacture of high-pressure mercury-vapour lamps to obtain a reproducible product is the mercury-vapour pressure in the lamp. In an ideal situation the arc tubes of the lamps would be made identical in volume, so that it would be sufficient to fill them all with the same quantity of mercury. However, with certain types of high-pressure mercury-vapour lamps this cannot be done, since the arc tubes are made from quartz tubing (*fig. 1*), and it is almost impossible to produce quartz tubing in quantity with an accurately constant inside diameter. A way out of the difficulty is to sort the tubing into a number of batches that have a sufficiently low spread in inside diameter, and then to adapt the mercury filling to the diameter per batch.

Besides the inside diameter, the wall thickness of the quartz tubing is also required to have a lower spread than is feasible in quantity production. This is because the softening range of quartz is very limited, so that when the pinch is made the heating of the material has to be critically controlled. In fully mechanized production the wall thickness of the starting material should therefore not vary too greatly.

There are difficulties in the mechanical measurement of wall thickness and inside diameter; it takes a long time and the tubing may get broken. Moreover it is not possible to determine the inside diameter with a plug gauge because quartz tubing is not usually accurately circular. We have therefore developed a method of measuring both wall thickness and inside diameter quickly and without mechanical contact by using a television camera.

In the measurement a piece of tubing is obliquely illuminated at one end. The other end is then viewed by a television camera, which sees it as a bright ring against a dark background. The video signal delivered by the camera now contains information that can be used for determining the cross-sectional area of the tube wall and of the inside of the tube. If the picture lines cut the picture of the tube cross-section, the video signal shows one or two positive rectangular pulses (*fig. 2*). The total duration of all these pulses can be measured and is a measure of the cross-section of the tube wall. The times between two successive pulses on each picture line, summed over all picture lines, give a measure of the inside diameter of the tube.

We have built a sorting machine for production purposes which is based on this principle (*fig. 3*). A description of this machine now follows.

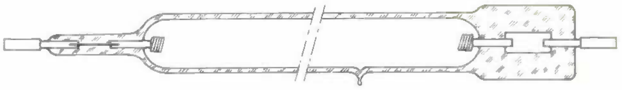


Fig. 1. The arc tube of a high-pressure mercury-vapour lamp, consisting of a quartz tube with an electrode at each end. The electrical connections to the electrodes are thin molybdenum strips, sealed vacuum-tight in a pinch. To show the construction of the pinch, the left-hand end of the arc tube is shown rotated through 90° with respect to the other end. The sealed-off exhaust tube can be seen in the middle of the tube. The auxiliary ignition electrode fitted at one end has been omitted. The arc tube is mounted in an outer bulb which is filled with nitrogen to a pressure of about 25×10^3 Pa. The inside of this bulb is coated with a phosphor, which converts the emitted UV radiation into visible light.

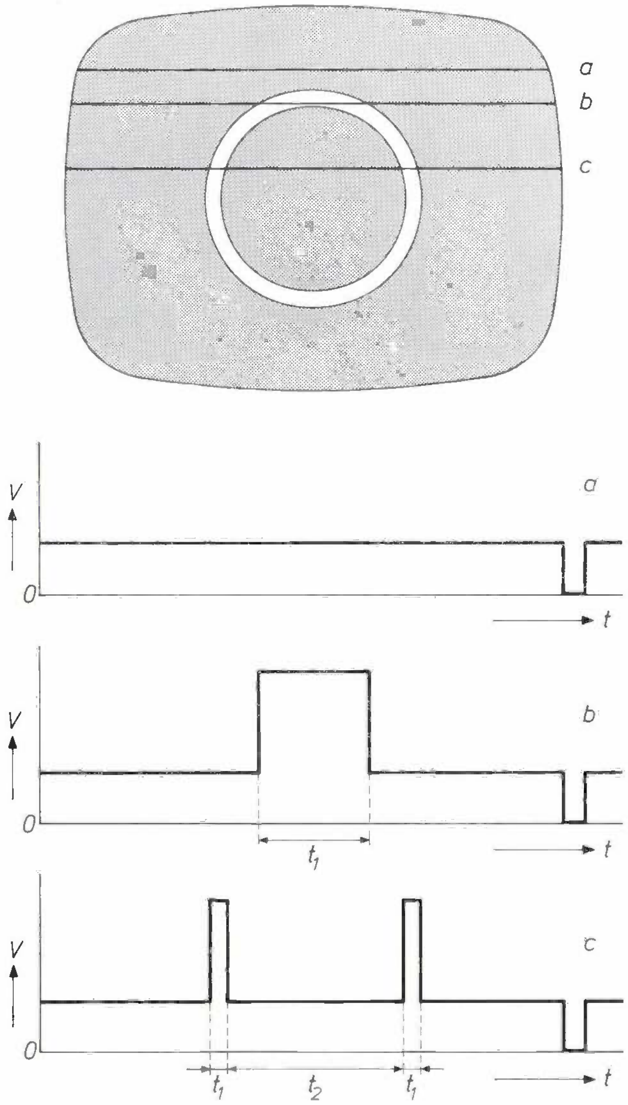


Fig. 2. Principle of the measurements. *Above*: the picture displayed on a monitor and three picture lines, *a*, *b* and *c*. *Below*: the video signals for these three picture lines. Summation of all the times t_1 provides a measure of the cross-sectional area of the tube wall. The total of the times t_2 is a measure of the surface area of the inside of the tubing.

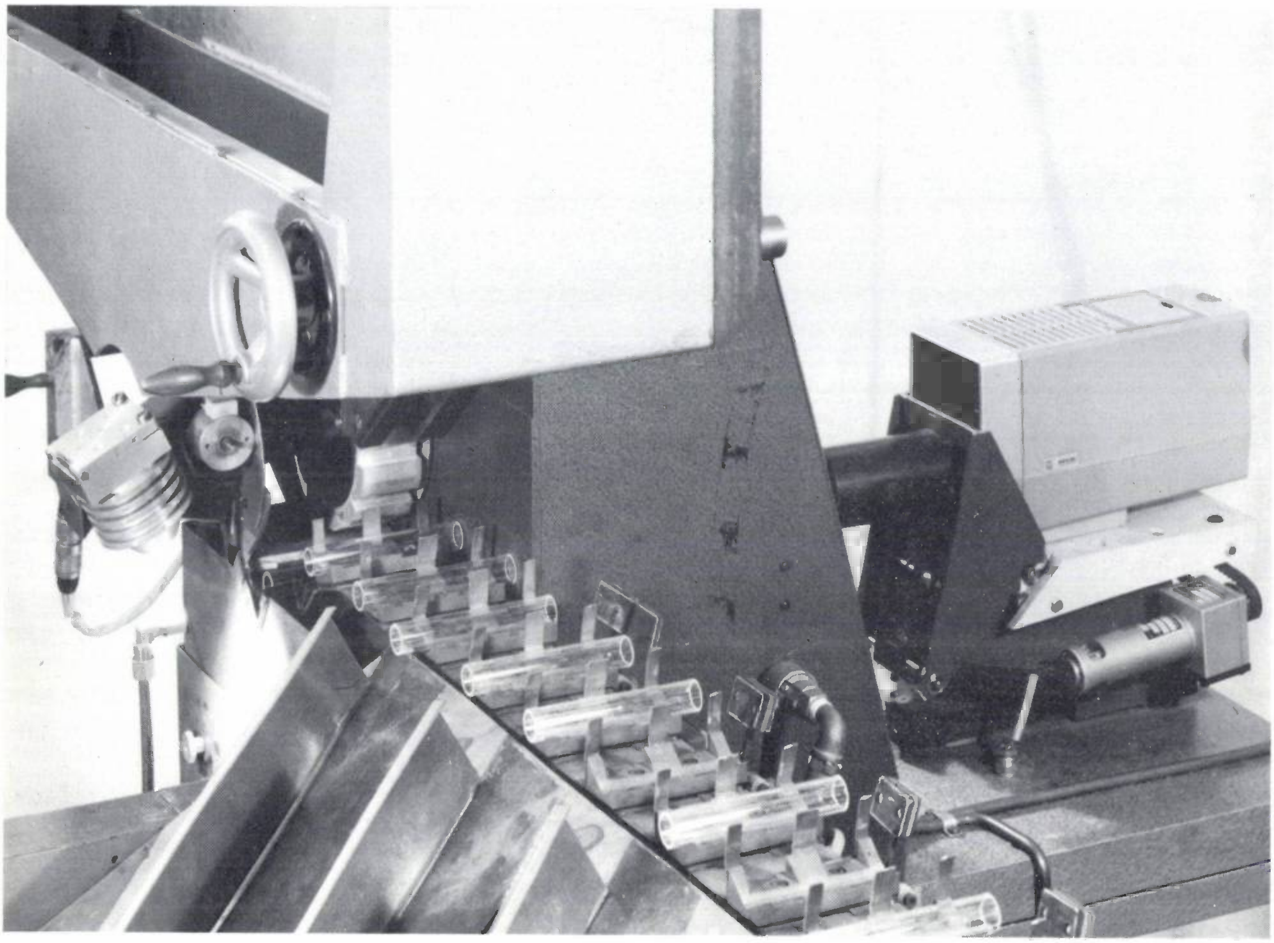


Fig. 3. The sorting machine. The storage unit can be seen at the top of the picture, above the grooved roller that puts the pieces of tubing one by one into the position for measurement. During the measurement the tubing is illuminated from the left and viewed from the right by a television camera. After measurement the piece of tubing is picked up by a conveyor belt. The tappers behind the belt are operated by ejector magnets to push the piece of tubing into the appropriate tray (foreground) corresponding to its dimensions.

Pieces of quartz tubing, cut to length for further processing, are taken one by one out of a storage unit by means of a rotating grooved roller, which lays them in the position for measurement. A lining of damping material ensures that the piece of tubing quickly comes to rest. In the measuring position the tube is illuminated at one end and the other end is viewed with a television camera, as described above. The actual measurement takes only 0.02 seconds, since only one field of the picture is required. After the measurement the tubing is picked up by a conveyor belt, which carries it past twenty ejector positions.

The dimensional limits of the cross-section of the pieces of tubing to be collected in the trays placed beside the ejector positions have previously been stored in a memory. The values measured for each piece are now stored in the memory as well, and then passed through it synchronously with the movement of the conveyor belt. Every time the measured values for a piece of tubing in a particular position come within the limits set for that position, an ejector magnet is energized which causes the tubing to be ejected into the corresponding tray. The machine shown in the photograph is capable of sorting tubing into twenty classes with preset limits. The sorting rate is 6000 pieces an hour.

P. G. Havas

Ir. P. G. Havas is with the Philips Lighting Division, Eindhoven.

Recent scientific publications

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

Philips Research Laboratories, Eindhoven, Netherlands	*E*
Mullard Research Laboratories, Redhill (Surrey), England	*M*
Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France	*L*
Philips Forschungslaboratorium Aachen GmbH, Weißhausstraße, 51 Aachen, Germany	*A*
Philips Forschungslaboratorium Hamburg GmbH, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium.	*B*

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

G. A. Acket & J. J. Scheer: Relaxation oscillations and recombination in epitaxial *n*-type gallium arsenide.
Solid-State Electronics **14**, 167-174, 1971 (No. 2). *E*

C. S. Aitchison: Parametric amplifiers designed for receiving systems.
Electronic Engng. **43**, No. 520, 56-59, June 1971. *M*

G. Arlt, H. Oepen & K. Reiber: Ein piezoelektrischer Synchronmotor für den Antrieb einer Uhr.
Feinwerktechnik **75**, 149-154, 1971 (No. 4). *A*

H. Bex: Über konzentrierte Zirkulatoren.
Nachrichtentechn. Z. **24**, 249-254, 1971 (No. 5). *A*

R. Bleekrode: Some recent spectroscopic investigations of low-pressure oxyacetylene flames.
Analytical Flame Spectroscopy, editor R. Mavrodineanu, Philips Technical Library, Eindhoven 1970, pp. 411-430. *E*

P. M. Boers: Measurements on dipole domains in indium phosphide.
Physics Letters **34A**, 329-330, 1971 (No. 6). *E*

G. A. Bootsma, W. F. Knippenberg & G. Verspui: Phase transformations, habit changes and crystal growth in SiC.
J. Crystal Growth **8**, 341-353, 1971 (No. 4). *E*

H. Bouma (Institute for Perception Research, Eindhoven): Visual recognition of isolated lower-case letters.
Vision Res. **11**, 459-474, 1971 (No. 5).

C. J. Bouwkamp, P. Janssen & A. Koene: Note on pantactic squares.
Math. Gazette **54**, 348-351, 1970. *E*

K. H. J. Buschow & A. S. van der Goot: The crystal structure of the two Th_2Fe_7 phases.
J. less-common Met. **23**, 399-402, 1971 (No. 4). *E*

H. B. G. Casimir: Computers in medicine — Why?
Hart Bull. **1**, 33-35, 1970 (No. 2). *E*

M. Cathelin, J. Magarshack & J. K. Vogel (Valvo, Hamburg): Experimental investigation of noise sources in Gunn diodes.
Proc. MOGA Conf., Amsterdam 1970, pp. 9.9-9.13. *L*

P. Dewilde (Stanford University, Calif., USA), **V. Belevitch & R. W. Newcomb** (Stanford Univ.): On the problem of degree reduction of a scattering matrix by factorization.
Proc. Kyoto Int. Conf. Circuit and System Theory 1970, p. 79. *B*

R. A. van Doorn: Nieuwe displaytechnieken, 1) Op zoek naar goede beelden, 2) Van telwerk tot platte televisiebuis.
Natuur en Techniek **39**, 123-129 & 196-206, 1971 (Nos. 4 & 5). *E*

Y. Genin: Minimum-fuel orbital transfers with low-thrust rockets.
Philips Res. Repts. Suppl. 1971, No. 2. (Part of thesis, Liège 1969.) *B*

J. A. Geurst: Continuum theory and focal conic texture for liquid crystals of the smectic mesophase.
Physics Letters **34A**, 283-284, 1971 (No. 5). *E*

J. M. Goethals & J. J. Seidel (Eindhoven University of Technology): A skew Hadamard matrix of order 36.
J. Austral. Math. Soc. **11**, 343-344, 1970 (No. 3). *B*

W. J. A. Goossens: A molecular theory of the cholesteric phase and of the twisting power of optically active molecules in a nematic liquid crystal.
Mol. Cryst. liq. Cryst. **12**, 237-244, 1971 (No. 3). *E*

R. G. Gossink: Properties of vitreous and molten alkali molybdates and tungstates.
Thesis, Eindhoven 1971. *E*

P. Hansen & W. Tolksdorf: Anisotropy of ruthenium-substituted yttrium-iron-garnet.
J. Physique **32**, C1/200-201, 1971 (Colloque No. 1, Vol. I). *H*

J.-P. Hazan & J. Haisma: Higher order non linear effects in some organic compounds.
Optics Comm. **2**, 343-348, 1970 (No. 7). *L*

J. C. M. Henning & J. P. M. Damen: Exchange interactions within nearest-neighbor Cr^{3+} pairs in chromium-doped spinel $ZnGa_2O_4$.
Phys. Rev. B **3**, 3852-3854, 1971 (No. 11). *E*

T. Holtwijk, W. Lems, A. G. H. Verhulst & U. Enz: Light modified switching properties of garnets and ferrites.
IEEE Trans. **MAG-6**, 853-857, 1970 (No. 4). *E*

R. N. Jackson & K. E. Johnson: Address methods for dc gas discharge display panels.
IEEE Trans. **ED-18**, 316-322, 1971 (No. 5). *M*

M. A. Karsmakers & A. J. Sanders: Het insmelten van gesinterde siliciumcarbide filterplaten.
Glastechn. Meded. **8**, 162-164, 1970 (No. 5). *E*

J. T. Klomp: Adhesie van metaal aan keramiek.
Klei en Keramiek **21**, 58-64, 1971. *E*

H. Koelmans & A. M. van Boxtel: Electrohydrodynamic flow in nematic liquid crystals.
Mol. Cryst. liq. Cryst. **12**, 185-191, 1971 (No. 2). *E*

F. A. Kuijpers & H. H. van Mal: Sorption hysteresis in the $LaNi_5$-H and $SmCo_5$-H systems.
J. less-common Met. **23**, 395-398, 1971 (No. 4). *E*

H. de Lang: Interferometric methods for the study of non-linear anisotropy in a lasing gas.
Optical Instruments and Techniques, Proc. Conf. Univ. of Reading 1969, pp. 125-130; 1970. *E*

F. K. Lotgering & G. H. A. M. van der Steen: Metal-deficient sulphospinels in the system Cr_2S_3-In_2S_3.
J. inorg. nucl. Chem. **33**, 673-678, 1971 (No. 3). *E*

J. Magarshack & A. Mircea: Stabilization and wide-band amplification using over-critically doped transferred-electron diodes.
Proc. MOGA Conf., Amsterdam 1970, pp. 16.19-16.23. *L*

J. M. Noothoven van Goor: Donors and acceptors in bismuth.
Thesis, Leiden 1971. *E*

G. Salmer (Faculté des Sciences de Lille), **E. Allamando** (Fac. Sci. Lille), **E. Constant** (Fac. Sci. Lille) **& A. Semichon:** Frequency multiplication using an avalanche diode.
Proc. MOGA Conf., Amsterdam 1970, pp. 12.13-12.18. *L*

E. Schröder: Elimination of granulation in laser beam projections by means of moving diffusers.
Optics Comm. **3**, 68-72, 1971 (No. 1). *H*

R. A. M. Scott: The temperature distribution in $ZnWO_4$ crystals during growth from the melt.
J. Crystal Growth **10**, 39-44, 1971 (No. 1). *M*

A. Semichon, J. Michel, E. Constant (Faculté des Sciences de Lille) **& A. Vanoverschelde** (Fac. Sci. Lille): Microwave oscillation of a tunnel transit-time diode.
Proc. MOGA Conf., Amsterdam 1970, pp. 7.15-7.20. *L*

P. J. Severin: Measurement of resistivity of silicon by the spreading resistance method.
Solid-State Electronics **14**, 247-255, 1971 (No. 3). *E*

R. Spitalnik: Analytic approach to the L.S.A. relaxation mode.
Proc. MOGA Conf., Amsterdam 1970, pp. 20.13-20.17. *L*

P. J. Strijkert, R. Loppes & J. S. Sussenbach: Arginine metabolism in Chlamydomonas reinhardi. Regulation of uptake and breakdown.
FEBS Letters **14**, 329-332, 1971 (No. 5). *E*

C. H. F. Velzel: Small phase differences in holographic interferometry.
Optics Comm. **2**, 289-291, 1970 (No. 6). *E*

C. H. F. Velzel: Influence of non-linear recording on image formation in holography.
Optics Comm. **3**, 133-136, 1971 (No. 3). *E*

J. Vlietstra & D. K. Wielenga: PHILCON.
Numerical Control Programming Languages, Proc. 1st Int. IFIP/IFAC Proclamat Conf., Rome 1969, pp. 53-70; 1970. *E*

J. J. Vrakking & F. Meyer: Auger electron spectroscopy made quantitative by ellipsometric calibration.
Appl. Phys. Letters **18**, 226-228, 1971 (No. 6). *E*

Q. H. F. Vrehen: Spectral distribution of the stimulated emission of a rhodamine B dye laser.
Optics Comm. **3**, 144-146, 1971 (No. 3). *E*

H. Weinerth: Integrierte lineare Breitbandverstärker.
Tagungsbroschüre der VDE-Fachtagung Elektronik 1970, Hannover, pp. 104-113. *E*

P. A. C. Whiffin & J. C. Brice: The suppression of thermal oscillations in Czochralski growth.
J. Crystal Growth **10**, 91-96, 1971 (No. 1). *M*

S. Wittekoek & G. Rinzema: The magneto-optic Kerr effect and Faraday rotation of $CdCr_2S_4$ for radiation between 0.1 and 4 eV.
Phys. Stat. sol. (b) **44**, 849-860, 1971 (No. 2). *E*

Volume 32, 1971, No. 2 pages 33-60 Published 19th November 1971

PHILIPS TECHNICAL REVIEW

VOLUME 32, 1971, No. 3/4

Magnetic deflection in television picture tubes

R. Vonk

To make television sets more compact it is necessary to design shorter picture tubes. This implies that for the same size of screen the electrons have to be deflected through larger angles. In electron optics, as in geometrical optics, the errors in a system increase with the angle between the rays and the axis. In the study of an electron-optical system it is important to be able to measure with sufficient accuracy the quantities that characterize the imaging errors. The article below explains how this can be done by a carefully designed system of measurements of magnetic field strength with automatic processing for the results of the measurements.

Introduction

When we bear in mind that some 400 000 picture elements can be separately reproduced in a television picture, it will be obvious that all the information which this requires cannot readily be transmitted simultaneously. Because of the slow response of the human eye, however, it is possible to transmit and reproduce the information element by element, provided it is done fast enough.

In the systems currently in use the picture is scanned in a line raster. At the display the same raster is described on the screen of a cathode-ray tube by an electron beam, whose intensity is at every instant proportional to the brightness of the picture element to be reproduced. The beam, in which the electrons have a velocity corresponding to an energy of some tens of keV, is focused by means of electron lenses to form a small spot on the screen.

The desired line raster is produced by deflecting the beam with two time-dependent transverse magnetic fields. These fields are generated by means of two pairs of deflection coils fitted to the neck of the cathode-ray tube (*fig. 1*). In an ideal deflection system the focusing of the beam on the screen would remain unaltered in every deflected position, and the displacement on the screen would always be linearly proportional to the strength of the deflecting field.

In practice this ideal situation can only be approximated; electron-optical imaging has the same kinds of error as light-optical imaging: distortion, astigmatism, image curvature, coma and spherical aberration. The relation between these errors and the spatial distribution of the magnetic field strength can be calculated by solving the equation of motion for an electron in the given field. This is done by expressing the magnetic field strength as a series expansion, whose coefficients for any particular case are found from accurate measurements of the field-strength distribution. The equation of motion for the electron can then be solved numerically, and the accuracy with which the solution describes the actual physical situation will increase with the number of terms taken in the series.

The results of these calculations enable the designer of deflection coils to find an optimum compromise between the desired deflection characteristics of the coil for a particular application and the imaging errors.

The usual procedure in designing deflection coils is first to build an experimental set of coils on the basis of experience and theoretical knowledge. Measurement of the field-strength distribution of these coils gives information about their deflection characteristics, and the results provide an indication of the way in which the design should be modified. The final result is arrived at after repeated measurements and corrections. It is only in the last stage of the design that the measure-

Ir. R. Vonk is with Philips Research Laboratories, Eindhoven.

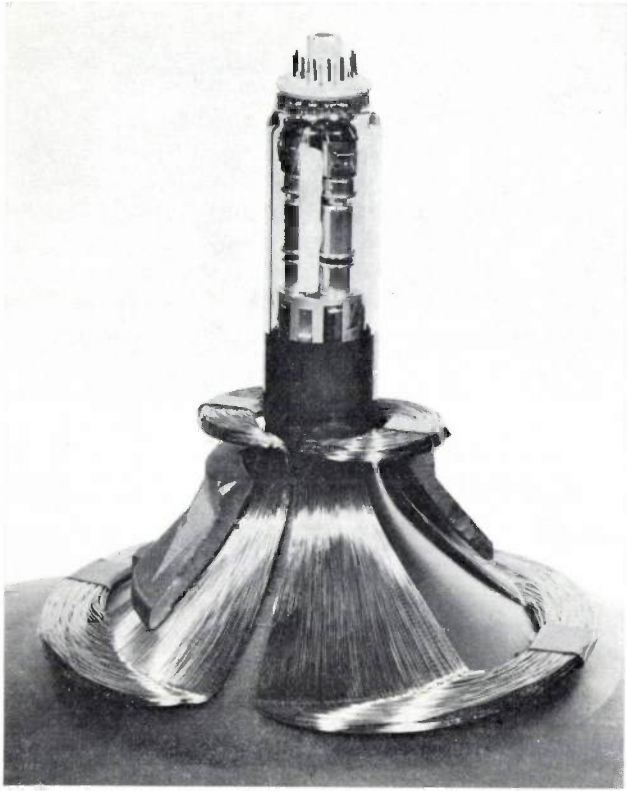


Fig. 1. A pair of deflection coils on the neck of a picture tube. To show the coils clearly, half of the ferrite core has been removed. Normally the two pairs of coils and a ferrite core are secured in a plastic encapsulation.

ments of the field-strength distribution are followed by detailed calculations of the behaviour of the electron beam in the field.

To a large extent, the continuing refinement in methods for measuring magnetic field strength and the development of computer programs for processing the results have made possible the design of modern television picture tubes and their deflection coils. These methods enable us to make optimum use of long-established theoretical knowledge of the magnetic deflection of an electron beam and the associated errors.

In the following we shall first discuss the mathematical description and measurement of the magnetic field, and then we shall deal with the deflection of an electron beam in this field and the deflection errors. Finally we shall present two examples of coil design for particular applications.

The field of the deflection coils

Mathematical description of the field

To describe the field of a pair of deflection coils we use a rectangular coordinate system as shown in *fig. 2*. The z-axis coincides with the axis of the tube and with the direction of the undeflected electron beam; the x-axis is horizontal, the y-axis vertical. The situation represented applies to the vertical-deflection coils that will now be discussed. The relations for the horizontal deflection coils are found by interchanging x and y in the final results.

Apart from the beam current, which we may neglect here, there are no currents in the region where the deflection takes place. This means that the line integral of the magnetic field strength **H** is zero along every closed contour in the region, or curl **H** = 0, so that we can write **H** as the gradient of a scalar magnetic potential Φ: **H** = grad Φ.

The potential Φ can be expanded as a series in the coordinates x and y, in which the coefficients of the terms are functions of z. The most general form of this series is:

$$\Phi(x,y,z) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + \\ + a_{11}xy + a_{02}y^2 + a_{30}x^3 + \dots \quad (1)$$

The derivatives of this series with respect to x, y and z now yield the series for the components H_x, H_y and H_z of the field strength. Since the field is symmetrical, because of the symmetry of the pair of coils, many of the coefficients in these series will be zero. The series for Φ is thus reduced to:

$$\Phi(x,y,z) = a_{10}x + a_{30}x^3 + a_{12}xy^2 + \dots \quad (2)$$

Here all coefficients are functions of z.

The influence of the symmetry of a pair of coils on the series for Φ and hence on the series for the components of **H** can be calculated as follows.

Differentiation of the series (1) with respect to x yields the x-component of the field strength:

$$H_x = \frac{\partial\Phi}{\partial x} = a_{10} + 2\,a_{20}x + a_{11}y + 3\,a_{30}x^2 + \dots .$$

It can be seen from the schematic representation of the lines of force in *fig. 3* that H_x has the same sign everywhere, and hence we must have $H_x(x,y,z) = H_x(-x,y,z) = H_x(x,-y,z)$. If the series for H_x is to satisfy this, it must only contain even powers of x and y, so that a_{20}, a_{11}, a_{21}, etc. must all be zero. The sign of H_y changes for a change in the sign of either x or y: $H_y(x,y,z) = -H_y(-x,y,z) = -H_y(x,-y,z)$ so that in the series $H_y = a_{01} + a_{11}x + 2\,a_{02}y + a_{21}x^2 + 2\,a_{12}xy + \dots$, the coefficients a_{01}, a_{11}, a_{02}, a_{21}, a_{03}, ... must be zero. If, taking into account the results already found, we now write the series for H_z, we obtain:

$$H_z = \frac{\partial\Phi}{\partial z} = a'_{00} + a'_{10}x + a'_{30}x^3 + a'_{12}xy^2 + \dots ,$$

in which the dashes indicate differentiation with respect to z. It can be seen from fig. 3 that we must have $H_z(x,y,z) = -H_z(-x,y,z) = H_z(x,-y,z)$, so that a'_{00} will have to be zero and the coefficient a_{00} is therefore a constant. Without loss of generality this constant can be taken to be zero.

Since there are no isolated magnetic poles, all lines of force are closed contours, or div **H** = 0. This means that the potential Φ must satisfy Laplace's equation, for:

$$\operatorname{div}\mathbf{H} = \frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2} = 0.$$

Substituting the series (2) for Φ gives:

$$\operatorname{div}\mathbf{H} = (6\,a_{30}+2\,a_{12}+a_{10}'')x+(20\,a_{50}+2\,a_{32}+a_{30}'')x^3+ \\ +(6\,a_{32}+12\,a_{14}+a_{12}'')xy^2+\ldots = 0,$$

in which the dashes indicate differentiation with respect to z. This relation must be valid for every value of x and y, and therefore the coefficient of each term must be zero. This yields equations for the coefficients a of the series (2) for Φ. The first three equations are:

$$6\,a_{30} + \;\; 2\,a_{12} + a_{10}'' = 0, \qquad (3a)$$

$$20\,a_{50} + \;\; 2\,a_{32} + a_{30}'' = 0, \qquad (3b)$$

$$6\,a_{32} + 12\,a_{14} + a_{12}'' = 0. \qquad (3c)$$

Adding to these the equation obtained by differentiating (3a) twice with respect to z:

$$6\,a_{30}'' + 2\,a_{12}'' + a_{10}'''' = 0,$$

we can then express a_{30}, a_{50}, a_{32} and a_{30}'' in terms of a_{10}, a_{12}, a_{14} and their derivatives. With the notation $a_{10} = H_0$, $a_{12} = H_2$ and $a_{14} = H_4$, this finally yields the following expressions for the components of $\mathbf{H}$:

$$\left.\begin{aligned} H_x(x,y,z) &= H_0 - (H_2 + \tfrac{1}{2}H_0'')x^2 + H_2y^2 + \\ &\quad + (H_4 + \tfrac{1}{6}H_2'' + \tfrac{1}{24}H_0'''')x^4 - \\ &\quad - (6\,H_4 + \tfrac{1}{2}H_2'')x^2y^2 + H_4y^4 + \ldots \\ H_y(x,y,z) &= 2\,H_2xy - (4\,H_4 + \tfrac{1}{3}H_2'')x^3y + \\ &\quad + 4\,H_4xy^3 + \ldots \\ H_z(x,y,z) &= H_0'x - (\tfrac{1}{3}H_2' + \tfrac{1}{6}H_0''')x^3 + \\ &\quad + H_2'xy^2 + \ldots . \end{aligned}\right\} \qquad (4)$$

If we know the numerical values of H_0, H_2, H_4 and the appropriate derivatives, we find that for values of x and y that are not too large the field is given with good accuracy by the initial terms of the series expansion given in (4) (see *fig. 4*). The larger the deflection angles for which we wish to know the behaviour of the beam, the more terms we need of the series expansion, since the beam then goes through regions of the field which are farther away from the axis. For small deflection angles, on the other hand, even the terms H_2 and H_4 are unimportant.

For a coil of not unduly complicated shape and a limited number of turns, we can use Ampère's law to express the coefficients H_0, H_2 and H_4 in terms of the parameters that characterize the geometry of the coil [1]. In this case there should be no ferromagnetic components in the vicinity, such as the cores of coils or magnetic screening. However, these usually are present and moreover the geometry of the windings is so complicated that it is impossible to calculate the coefficients. The distribution of the field strength therefore has to be determined by measurements.

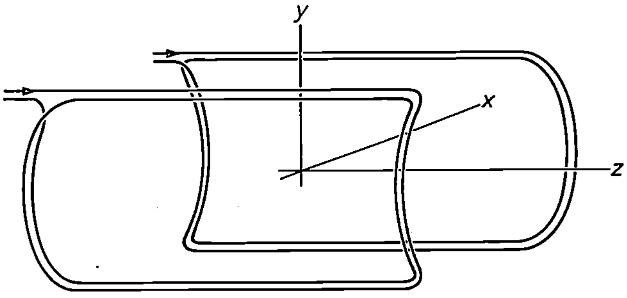


Fig. 2. Coordinate system for the mathematical description of the magnetic field of the deflection coils. The undeflected electron beam coincides with the positive z-axis. The schematically represented coils are those for the vertical deflection (y-direction).

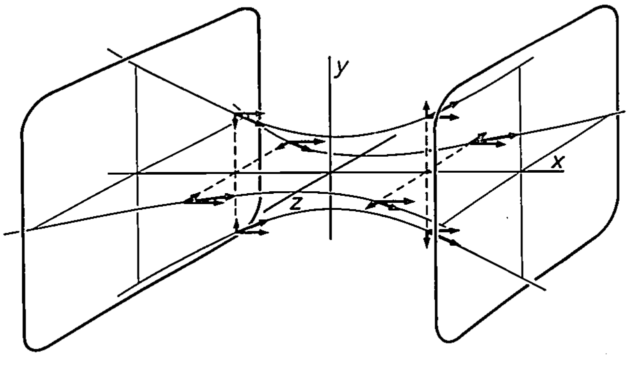


Fig. 3. Lines of force for the vertical-deflection coil. The field-strength vectors have been resolved into components to show the symmetry of the field.

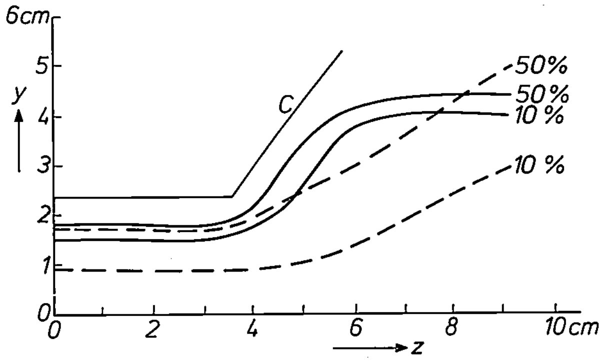


Fig. 4. Contours of the regions in the y-z-plane within which the actual and the approximated values of H_x, the x-component of the field strength, differ from one another no more than by the indicated percentages. The dashed lines apply to the approximation $H_x = H_0 + H_2y^2$, the solid lines to the approximation $H_x = H_0 + H_2y^2 + H_4y^4$. The line C gives the cross-section of the coil.

[1] J. Kaashoek, Thesis, Eindhoven 1968 (also pnblished as Philips Res. Repts. Suppl. 1968, No. 11), pages 44-48.

Principle of measurement of coefficients H_0, H_2 and H_4

For $x = 0$ the series (4) for the components of H become:

$$\left.\begin{aligned} H_x(0,y,z) &= H_0 + H_2y^2 + H_4y^4 + \dots,\\ H_y(0,y,z) &= 0,\\ H_z(0,y,z) &= 0. \end{aligned}\right\} \qquad (5)$$

This means that in order to calculate H_0, H_2 and H_4 it is sufficient to measure H_x alone, and only at points in the y-z-plane. In theory it would be sufficient to measure H_x at only three different values of y for each value of z. A measurement for $y = 0$ yields H_0, and two measurements with y equal to say 5 mm and 10 mm would then be sufficient for determining H_2 and H_4.

However the accuracy needed for these measurements is very high. In addition to the coefficients H_0, H_2 and H_4 we also want to be able to calculate from the measurements a number of derivatives of these coefficients with respect to z, i.e. derivatives of H_0 up to the fourth order and derivatives of H_2 up to the second order. Moreover in the vicinity of the axis the contribution of the terms with coefficients H_2 and H_4 is found to be very small. In the region considered the following applies in a given case: $H_2y^2/H_0 \leqslant 0.025$ and $H_4y^4/H_0 \leqslant 0 .006$ where $y = 5$ mm. It is not possible to make the situation more favourable by performing the measurements for substantially higher values of y, because the approximation $H_x = H_0 + + H_2y^2 + H_4y^4$, used in processing the measurements, would then no longer be sufficiently accurate (see fig. 4). Nor does the geometry of the coils usually allow us to carry out measurements taking y greater than 5 to 10 mm. To achieve the required accuracy we measure H_x at every z-value for at least five different values of y. From the results we can then calculate not only H_0, H_2 and H_4 and their derivatives but also any asymmetry that may be present in the coils. Carrying out more measurements than are strictly necessary also has the advantage that a number of systematic errors of measurement can then be determined. If necessary the measuring results can first be corrected for these errors before further processing. We shall return to these corrections after we have discussed the actual measurements.

The measurement procedure

For the field-strength measurement the coil is energized by an a.c. current of 1 A at a frequency of 4 kHz. The voltage induced in a search coil placed in the field is a measure of the field-strength component parallel to the axis of the search coil. If the shape of the search coil is suitably chosen the field strength can be measured for a point at the centre of the coil instead of a value averaged over its entire volume [2], which would be the case with other types of probe such as the Hall probe.

For part of our measurements we use two identical search coils. One of these can be placed on the z-axis in the field to be measured and the other parallel to it, at a distance of 5 mm in the y-direction. Connected in opposite polarity these two coils deliver a signal which is a measure of the difference in field strength between these two points. Equation (4) shows that this difference is equal to $H_2y^2 + H_4y^4$, and it can thus be used directly for calculating H_2 and H_4. An advantage is that we have exactly the same difference in position between the two coils in all measurements.

Fig. 5 shows a diagram of the equipment used for these measurements. The search coils are moved through the field along a straight line parallel to the z-axis by means of a worm shaft driven by a stepping motor. The motor makes a large number of steps for one revolution of the shaft, so that the distance between two successive measuring points can be adjusted with the required accuracy. After the whole field has been traversed in the z-direction for one value of y, the deflection coil and the search coils are moved apart by an accurately known distance in the y-direction, and a new series of measurements is carried out. *Fig. 6* shows a complete deflection unit on our measuring equipment.

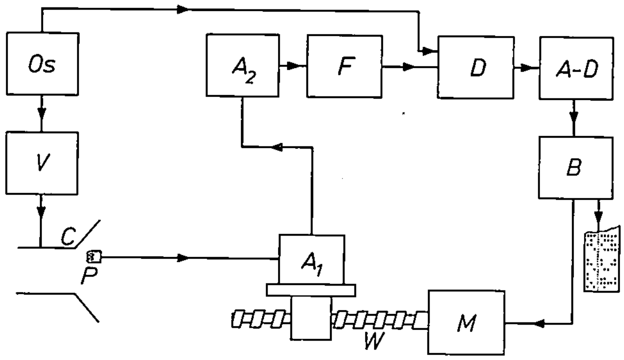


Fig. 5. Block diagram of the measuring equipment. *Os* oscillator (4 kHz). *V* power amplifier, supplying the current to the coil *C* being measured. *P* search coil, which is moved through the field by means of a worm shaft *W* driven by a stepping motor *M*. *F* 4 kHz filter. *D* synchronous detector. *A-D* analog-to-digital converter. *B* tape punch and digital control. A_1 and A_2 amplifiers.

The signal from a search coil, or the difference signal from two search coils, is amplified, passed through a 4 kHz filter and then measured with a synchronous detector. The reference signal for this detector is derived from the same 4 kHz oscillator that provides the current that, after power amplification, energizes the

[2] R. F. K. Herzog and O. Tischler, Rev. sci. Instr. 24, 1000, 1953..

deflection coils. After analog-to-digital conversion the results of the measurements are transferred to punched tape as a direct input for computer processing.

Before the measured data receive their final processing they are first corrected for inequalities between the two search coils. Such inequalities are determined beforehand in a separate measurement. After this correction H_0, H_2 and H_4 are calculated from the measurements made for a single value of z. It is also possible to determine from the measurements whether the line along which the search coils have moved was at the right distance parallel to the z-axis. If not, the data can be corrected both for the incorrect distance and for deviations from parallel; the final values of H_0, H_2 and H_4 are then calculated as a function of z, and the derivatives of these coefficients with respect to z are computed.

The measured values of H_0, H_2 and H_4 and the values calculated using Ampère's law are shown in *fig. 7* as a function of z for a coil of simple geometry. The difference between measured and calculated values can be completely accounted for by a single slight difference between the actual geometry of the coil and the geometry assumed for the calculations: this demonstrates the reliability of our measurements.

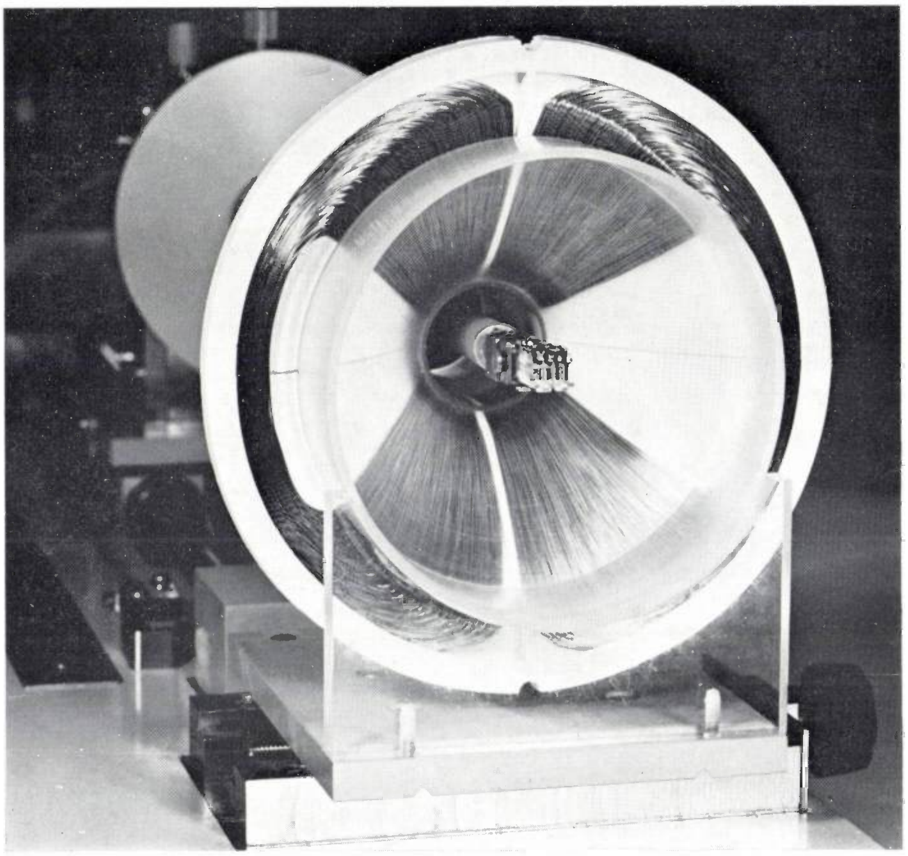


Fig. 6. A deflection unit on the measuring equipment for measuring the field-strength distribution near the axis. From this field-strength distribution the coefficients H_0, H_2 and H_4 and their derivatives can be calculated. Only one of the two pairs of coils can be seen. The two search coils protrude through the central hole. In the position shown here the search coils measure the field of the pair of deflection coils that cannot be seen. The deflection unit is mounted on a carriage that permits the unit to be translated perpendicular to the line the search coils follow during a series of measurements.

The movement of an electron in a magnetic field

The equation of motion

There is a close analogy between geometrical optics and electron optics. To derive the equation for the path of an electron in a given electric or magnetic field we can therefore proceed from Fermat's principle, used in geometrical optics for calculating the path of light rays. Fermat's principle states that the path of a light ray between two points P and Q is such that the actual optical path, measured in wavelengths of the light, is no longer than any possible neighbouring paths. In other words, the variation of the optical path length

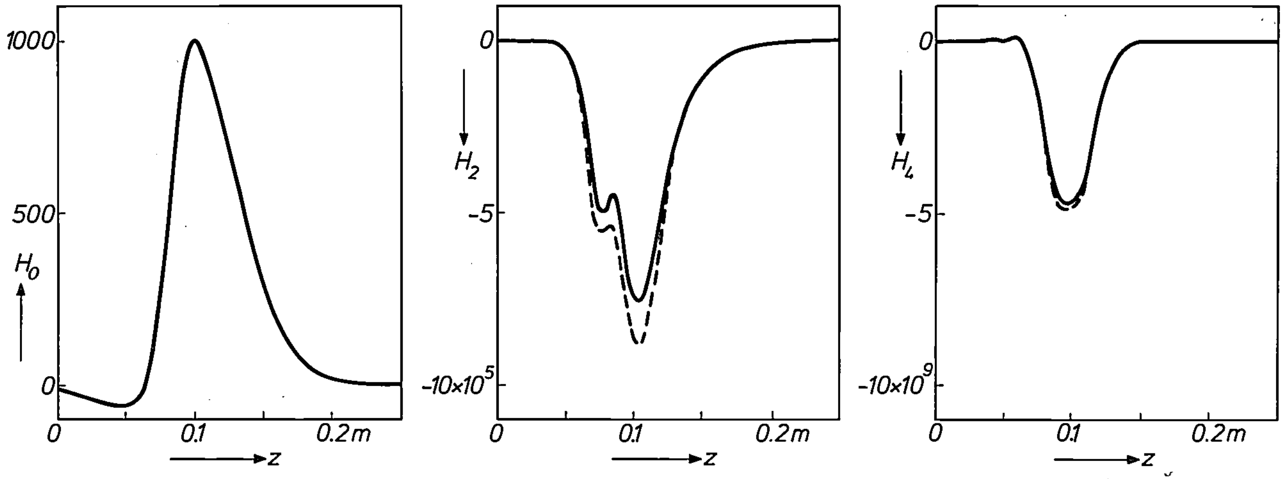


Fig. 7. Measured values (solid lines) and calculated values (dashed) of the coefficients H_0, H_2 and H_4 for a coil of simple geometry. The slight difference between the measured and the calculated curves can completely be accounted for by assuming a single slight difference in geometry between the actual coil and the configuration taken for the calculations.

must be zero; this is expressed by the equation:

$$\delta \int_P^Q \frac{\mathrm{d}l}{\lambda_0/n} = 0,$$

where δ is the variation sign, dl is an element of the light path between P and Q, n is the refractive index at the location of dl, and λ_0 is the wavelength of the light *in vacuo*. Since λ_0 is a constant, the above equation is equivalent to:

$$\delta \int_P^Q n \, \mathrm{d}l = 0. \qquad (6)$$

This equation can also be used for calculating electron paths in electric and magnetic fields, provided that we substitute:

$$n = \sqrt{V} - \sqrt{e/2m}\, \mathbf{A} \cdot \mathrm{s}. \qquad (7)$$

In this expression V is the electrostatic potential with respect to the cathode, where the electrons are emitted, e and m are the charge and mass of the electron, **A** is the magnetic vector potential and s is the unit vector in the direction of the electron path. (The vector potential **A** is obtained by integration of the electric currents contributing to the field, in the same way as the electrostatic potential is obtained by integration of a charge distribution; the vector potential **A** and the magnetic induction **B** are related by **B** = curl **A**.) Since the scalar product of **A** and s appears in the expression (7) for the refractive index, this index is direction-dependent: the electrons move in an anisotropic medium.

To calculate the electron paths in a picture tube we adopt the same rectangular system of coordinates as we used for describing the magnetic field. Again taking z as an independent variable we find, after substituting (7) in (6):

$$\delta \int \left\{ \sqrt{V(1+x'^2+y'^2)} - \right.$$
$$\left. - \sqrt{\frac{e}{2\,m}}\,(A_x x' + A_y y' + A_z) \right\} \mathrm{d}z = 0. \qquad (8)$$

If a function F is to satisfy $\delta \int F \mathrm{d}z = 0$, it must also satisfy the Euler-Lagrange differential equations [3]:

$$\frac{\partial F}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}z}\frac{\partial F}{\partial x'} = 0,$$

$$\frac{\partial F}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}z}\frac{\partial F}{\partial y'} = 0.$$

Substituting for the function F the expression between curly brackets in equation (8), and using $\mathbf{B} = \mu_0 \mathbf{H} =$ = curl **A**, we find after some manipulation the following differential equations for the path of an electron [4]:

$$\left.\begin{aligned} x''(z) &= -\mu_0 \sqrt{\frac{e}{2mV}}\, \sqrt{1+x'^2+y'^2} \times \\ &\qquad \times \{x'y'H_x - (1+x'^2)H_y + y'H_z\} \\ y''(z) &= \mu_0 \sqrt{\frac{e}{2mV}}\, \sqrt{1+x'^2+y'^2} \times \\ &\qquad \times \{x'y'H_y - (1+y'^2)H_x + x'H_z\} \end{aligned}\right\} \qquad (9)$$

If we now substitute the series (4) derived previously for the components of **H**, and in addition expand the expression $\sqrt{(1+x'^2+y'^2)}$ as a series, we find series in x, y, x' and y' for x'' and y''. By terminating these series after a certain number of terms and solving the resultant differential equations, we obtain approximate expressions for the electron paths. The boundary conditions for the solution of the differential equations are determined by the position at which an electron enters the field and the slope of the path at that position. Since the familiar McLaurin series for $\surd(1 + x'^2 + y'^2)$ is valid only for $\surd(x'^2 + y'^2) \leqslant 1$, all further results obtained with this will be subject to the same limitation. This implies that they will therefore basically not be valid for deflection angles greater than 45°. The fact that the approximate expressions obtained by terminating the series (4) fail earlier is understandable; the extent to which the various approximations are still valid appeared from fig. 4. For deflection angles greater than 40° we must use measurements of the entire field, which we still have to discuss, and no longer make do with measurements only in the vicinity of the z-axis.

First-order approximation

We obtain the first-order approximation of the electron paths by neglecting all terms of first and higher degree in x, y, x' and y' in the series expansion for equations (9). We then find:

$$\left.\begin{aligned} x'' &= 0, \\ y'' &= -\mu_0 \surd e/2mV\, H_0 = -k\, H_0. \end{aligned}\right\} \qquad (10)$$

If the electron enters the magnetic field at the point (x_0, y_0, z_0) at an angle given by x_0' and y_0', and if we assume that V and therefore k are constant, then by integrating equations (10) twice we find the first-order approximation of the equations for the electron paths:

$$\begin{aligned} x(z) &= x_0 + x_0'(z - z_0), \\ y(z) &= y_0 + y_0'(z - z_0) - k \int\limits_{z_0}^{z} \int\limits_{z_0}^{\zeta} H_0(\xi)\,\mathrm{d}\xi\,\mathrm{d}\zeta, \end{aligned}$$

where ξ and ζ are integration variables. These equations are also known as Gaussian or paraxial-ray equations. The name "paraxial" signifies that the equations only give an accurate description of paths in the immediate vicinity of the z-axis and lying at small angles to it.

Since x_0, y_0, x_0' and y_0' are usually difficult to determine, we prefer to use the values of these parameters recalculated for the position z_s of the fluorescent screen. Substituting

$$\begin{aligned} x_s' &= x_0', \quad x_s = x_0 + x_0'(z_s - z_0), \\ y_s' &= y_0', \quad y_s = y_0 + y_0'(z_s - z_0), \end{aligned}$$

then yields the paraxial-ray equations in the form:

$$\begin{aligned} x(z) &= x_s + x_s'(z - z_s), \\ y(z) &= y_s + y_s'(z - z_s) - k \int\limits_{z_0}^{z} \int\limits_{z_0}^{\zeta} H_0(\xi)\mathrm{d}\xi\mathrm{d}\zeta = \\ &= y_s + y_s'(z - z_s) + Y. \end{aligned}$$

The beam parameters x_s, y_s, x_s' and y_s' give the position and slope of the non-deflected beam at the screen, and Y gives the displacement of the beam as a result of the magnetic field. This displacement is directly proportional to the strength of the field and is independent of the beam parameters, thus remaining the same for every position of the non-deflected beam. The deflection is therefore ideal, and in this first-order approximation there is no question of any imaging errors.

The series expansions for the equations (9) contain no terms that are linear in x, y, x' or y', and therefore the second-order approximation is essentially the same as the first-order approximation.

Third-order approximation

If we terminate the series expansions for equations (9) after the terms of the second degree in x, y, x' and y', we obtain the differential equations:

$$\begin{aligned} x''(z) &= -k(x'y'H_0 + xy'H_0' - 2\,xy\,H_2), \\ y''(z) &= -k\{H_0 + \tfrac{1}{2}(x'^2 + y'^2)H_0 - x'x\,H_0' - \\ &\quad - x^2(H_2 + \tfrac{1}{2}H_0'') + y^2 H_2 + y'^2 H_0\}. \end{aligned}$$

In addition to the coefficient H_0, which was also present in the first-order approximation, it now appears that we must also know H_2, H_0' and H_0'' to be able to carry out numerical calculations.

The solution of the above two simultaneous differential equations can be approximated by substituting for x, y, x' and y' the solutions of the first-order equations and then integrating twice the expressions thus obtained for x'' and y''. These calculations result in the following expressions for the landing point where the electron path meets the screen:

$$\begin{aligned} x(z_s) &= x_s + (B_5\, x_s' + B_{10}\, x_s)\, Y_s^2 + \\ &\quad + (2\, B_8 x_s' y_s' + B_{15} x_s y_s + B_{17} x_s' y_s + B_{18} x_s y_s')\, Y_s \end{aligned}$$

$$\begin{aligned} y(z_s) &= y_s + Y_s + B_1 Y_s^3 + (B_4 y_s' + B_9 y_s)\, Y_s^2 + \\ &\quad + (B_7 y_s'^2 + B_8 x_s'^2 + B_{13} y_s^2 + B_{14} x_s^2 + B_{16} y_s y_s' + \\ &\quad + B_{18} x_s x_s')\, Y_s. \end{aligned} \qquad (11)$$

Here $Y_s = -k \int\limits_{z_0}^{z_s} \int\limits_{z_0}^{\zeta} H_0(\xi)\,\mathrm{d}\xi\,\mathrm{d}\zeta$, and the coefficients B are integrals of the same form which contain the field coefficients H_0, H_2, H_0' and H_0''.

[3] M. Born and E. Wolf, Principles of optics, 3rd edition, Pergamon Press, London 1965, Appendix I; H. Margenau and G. M. Murphy, The mathematics of physics and chemistry, 2nd edition, Van Nostrand, Princeton, N.J., 1956, chapter 6; J. Mathews and R. L. Walker, Mathematical methods of physics, 2nd impression, Benjamin, New York 1965, chapter 12.

[4] Details of the calculation will be found on page 79 of J. Kaashoek's thesis [1].

These expressions for the deflection are valid up to greater distances from the z-axis than the results of the first-order approximation. We see now, however, that imaging errors may occur, which increase with the distance of the undeflected beam from the z-axis and the angle between beam and axis. The presence of terms which are dependent on Y_s in the expression for $x(z_s)$ indicates that when the beam is deflected in the y-direction there is usually a deviation in the x-direction in cases where not all the beam parameters are zero.

The error that does not depend on the beam parameters x_s, y_s, x_s' and y_s', and which is found to be proportional to Y_s^3, is called distortion. The terms in (11) that are linearly proportional to the beam parameters and to the square of Y_s describe the image-field curvature and the astigmatism. Finally there is the error known as coma, which depends on the square of the beam parameters and is linearly proportional to Y_s.

For a parallel beam that enters the magnetic field along the z-axis ($x_s = y_s = x_s' = y_s' = 0$) the above equations (11) become much simpler:

$$x(z_s) = 0\,,$$

$$y(z_s) = Y_s + B_1 Y_s^3\,.$$

Here the coefficient B_1 characterizes a distortion, the non-linearity of the displacement (*fig. 8a*). In practice this error is corrected by causing the deflection current to increase with time in such a way (fig. 8*b*) that the spot on the screen of the picture tube moves linearly with time (S correction).

A beam of finite cross-section, and focused on the centre of the screen in the undeflected state ($x_s = y_s = 0$), is found to give an elliptical spot when it is deflected. Further study of the beam shows (*fig. 9*) that there are two locations, A and B, where the beam is focused to a thin line (the phenomenon known in optics as astigmatism). At a point C, which lies somewhere between A and B, the beam will have a circular cross-section. If points A and B coincide with C, there is no astigmatism, and the deflection coils are then said to be anastigmatic. The surface including all the points A that are obtained on deflection through different angles is the sagittal image surface, the points B give the meridional image surface, and the points C between these two surfaces give the mean image surface. These surfaces are not planar, even when they coincide. The coefficients B_4 and B_5 in equations (11) give the curvature of the meridional and sagittal image surface.

The relation between B_4 and the curvature of the meridional image surface is found by considering two sub-beams (*1* and *2* in *fig. 10*) of an electron beam which in the non-deflected state is focused on the centre of the screen. For the beam *1* we have $x_s = y_s = x_s' = y_s' = 0$, for *2* we have $x_s = y_s = x_s' = 0$, while in this case y_s' differs from zero, but it is so small that $y_s'^2 \approx 0$. To describe the behaviour of these two beams we can thus simplify equations (11) to:

$$x(z_s) = 0\,,$$

$$y(z_s) = Y_s + B_1 Y_s^3 + B_4 y_s' Y_s^2\,.$$

If they are deflected through a distance Y_s the difference in the

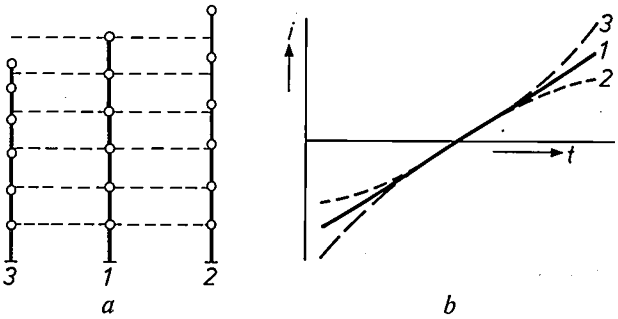


Fig. 8. *a*) Vertical deflection of an electron beam which enters the magnetic field along the z-axis. *1* Ideal deflection, linearly proportional to the strength of the magnetic field (the coefficient B_1 in $y(z_s) = Y_s + B_1 Y_s^3$ is zero). *2* Distortion in the case where $B_1 < 0$. *3* Distortion in the case where $B_1 > 0$. *b*) Deflection current i as a function of time t. *1* Current varying linearly with time (for ideal deflection). *2* and *3* Corrected currents for $B_1 < 0$ and $B_1 > 0$ (S correction).

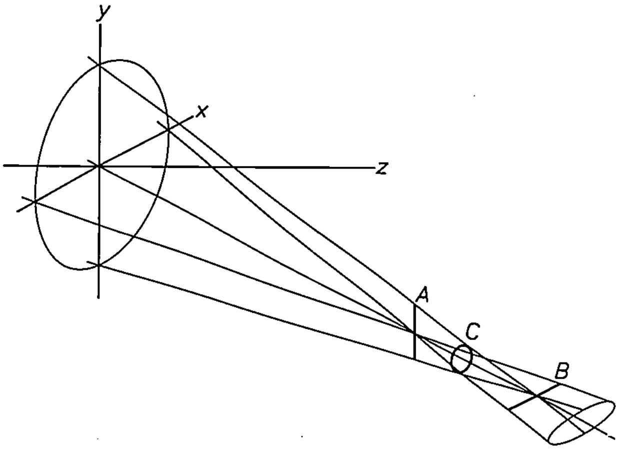


Fig. 9. Astigmatic beam. The beam is not focused at one point but gives two thin lines at A and B. At C the beam has a circular cross-section.

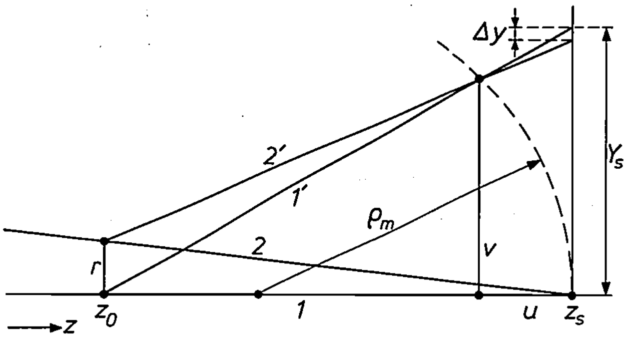


Fig. 10. Diagram used in calculating the relation between the coefficient B_4 and the radius of curvature ρ_m of the meridional image surface.

position of the two sub-beams on the screen is $\Delta y(z_s) = B_4 y_s' Y_s^2$. There is a point where the deflected beams *1'* and *2'* intersect ($\Delta y(z) = 0$). The locus of these points of intersection is the meridional image surface. This surface is spherical and has a radius of curvature ϱ_m.

Beam *2* enters the magnetic field at a distance r from the axis, so that $(y_s')_2 = -r/(z_s - z_0)$. Using the notation for the points and line-sections given in fig. 10 we can proceed with the calculation as follows. From fig. 10 it can be seen that $\Delta y/r = u/(z_s - z_0 - u)$, and $\varrho_m^2 = v^2 + (\varrho_m - u)^2$. This latter equation can be written as $v^2 = 2\,\varrho_m u - u^2$ from which, if u is so small that $u^2 \ll 2\,\varrho_m u$ and $Y_s \approx v$, it follows that $u \approx Y_s^2/2\,\varrho_m$. Substituting this in the relation for $\Delta y/r$ and replacing $z_s - z_0 - u$ by $z_s - z_0$, we find:

$$\Delta y \approx \frac{Y_s^2 r}{2\,\varrho_m(z_s - z_0)}$$

or, since $y_s' = -r/(z_s - z_0)$,

$$|\Delta y| \approx \frac{y_s' Y_s^2}{2\,\varrho_m}.$$

This expression must be equal to $B_4 y_s' Y_s$, and therefore $B_4 = 1/(2\,\varrho_m)$. Similarly, by proceeding from two sub-beams which enter the magnetic field with a slight difference in position in the x-direction, we find that $B_5 = 1/(2\,\varrho_s)$, where ϱ_s is the radius of curvature of the sagittal image surface.

Fifth-order approximation

If we continue the series expansion of equations (9) up to and including the terms of the fourth degree in x, y, x' and y', we obtain the fifth-order approximation. There is no fourth-order approximation because the series contains no terms of third degree in these variables.

The coefficients of the fifth-order equations are integrals which contain the coefficients H_0, H_2 and H_4 of the series expansion for the magnetic field, and also the derivatives H_0', H_0'', H_0''', H_0'''', H_2' and H_2''. All these quantities can be determined, as described earlier, from the measurements of the magnetic field strength.

The results of the fifth-order approximation are rather involved and will not be dealt with here [5]. It will be sufficient to give a graph (*fig. 11*) showing the results of measurements for two particular image errors, together with the curves that give the errors as calculated with the aid of the third-order and fifth-order approximations. The gain in accuracy obtained with the fifth-order approximation for the parts of the picture screen near the edges can clearly be seen.

So far we have deliberately restricted the whole treatment to deflection in one direction by the magnetic field of one pair of coils. By combining the two series for the mutually perpendicular magnetic fields of two pairs of coils series expansions for equations (9) can be obtained, which describe the motion of an electron in the combined field of the two pairs of coils [6]. We shall not go into this here.

Calculation for wide deflection angles

As we noted earlier, the approximations given above only apply to deflection angles smaller than 45°. In practice the approximations will be useful up to deflection angles of not more than 40°.

In designing modern 110° picture tubes, in which the largest deflection angle is 55°, we were therefore compelled to solve the electron-path equations (9) for the edges of the screen by direct integration. To do this we have to know the field components H_x, H_y and H_z as functions of x, y and z for the whole of the region in question. We cannot use the series expansions (4) for the field strength for this. If we use a limited number of terms, these series expansions give unacceptable deviations between the calculated and actual field strength for large values of x and y. The only thing left to do is to measure the components of the field strength in this region at a sufficiently large number of points. The measuring equipment constructed to do this is shown in *fig. 12*. Experience gained with the automation of the measurements described above has made it possible

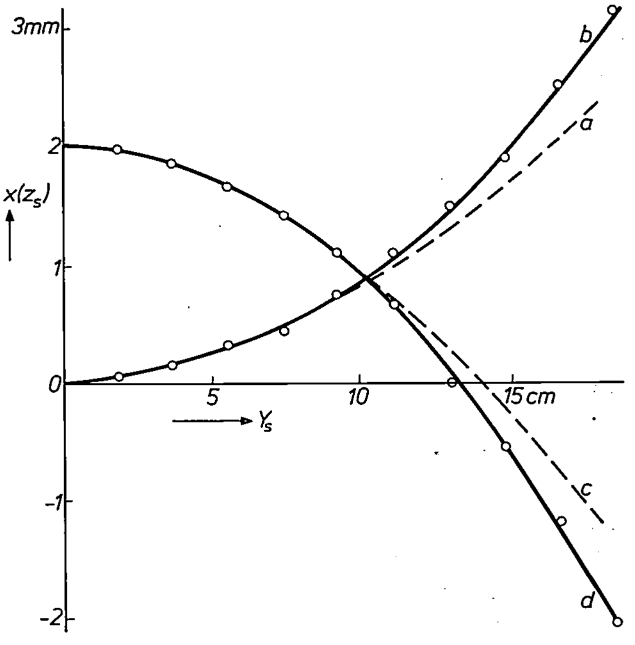


Fig. 11. When deflected in the y-direction the beam is also deflected in the x-direction. The graph gives the measured and calculated values of deflection errors $x(z_s)$ for deflection through a distance Y_s in the y-direction. (Y_s is proportional to the current through the deflection coils.) *a*) Third-order approximation of a beam for which only $x_s' \neq 0$. *b*) Fifth-order approximation for the same case. *c*) Third-order approximation for a beam where only $x_s \neq 0$. *d*) Fifth-order approximation for the same case. The circles indicate the measured deflection errors.

[5] A full treatment of the fifth-order approximation is given in chapter 5 of J. Kaashoek's thesis [1].

[6] J. Haantjes and G. J. Lubben, Philips Res. Repts. 12, 46, 1957 and 14, 65, 1959.

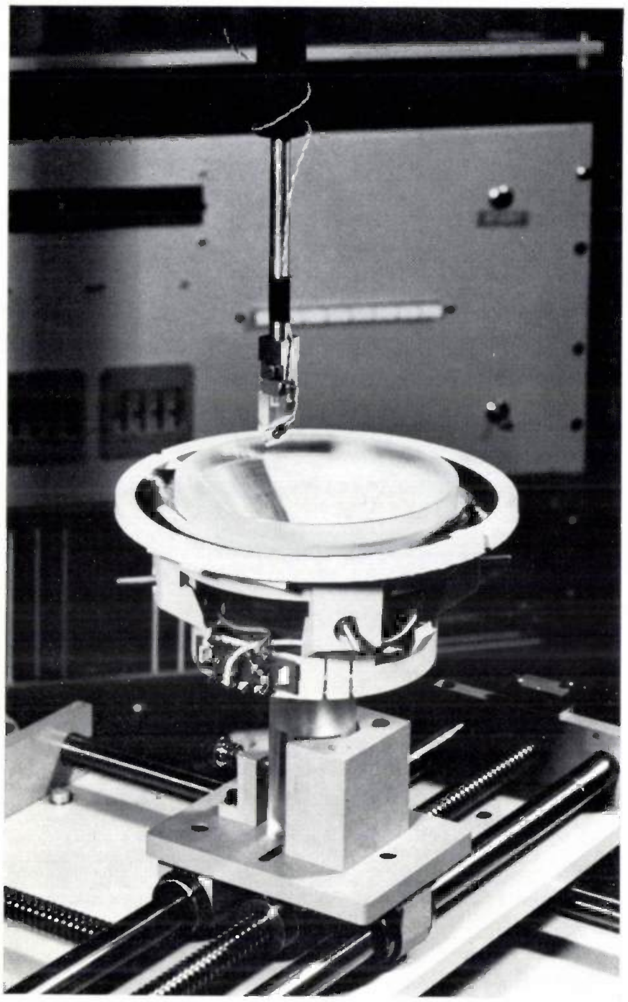


Fig. 12. Equipment for measuring the field strength in the region swept by the beam. The probe with a search coil can be seen above the deflection unit. The search coil is set at an angle of 45° to the z-axis of the deflection unit (the position of the deflection unit is such here that the z-axis is vertical). At every point four measurements are carried out, and after each measurement the probe is rotated through 90° about a vertical axis. The probe can reach any point desired in the magnetic field; it can be moved up and down while the deflection unit can be shifted in both horizontal directions on a slide system. All movements are made automatically in accordance with a preconceived plan. The results of the measurements are delivered on punched tape for computer processing.

to automate the considerable amount of measurement work involved and to process the results with the aid of a computer. With these data we are now in a position to make a complete calculation of the performance of a given coil.

Two applications

The shadow-mask tube

The shadow-mask tube is the type of colour picture tube now in general use. The colours are obtained by additively mixing three basic colours: red, green and blue. The tube contains three electron guns, usually arranged in an equilateral triangle around the axis of the tube, and a regular pattern of red, green and blue phosphor dots is applied to the screen. At a short distance from the screen a perforated plate, the shadow mask, ensures that the electron beam from each gun only strikes dots of one particular colour. This is achieved by situating the guns at a distance from the tube axis so that the beams arrive at the mask from different angles. In passing through the same hole in the mask the beams strike three different dots (*fig. 13*) and thus produce three separate pictures in the basic colours.

A prerequisite for good colour reproduction is that the three separate pictures should be in exact register with one another everywhere on the screen. This means that the three electron beams must be directed on to the same place on the screen in every deflected position. A correction for convergence is necessary to achieve this [7].

If the tube is to operate correctly with a regular pattern of phosphor dots on the screen, the three beams passing through the same hole in the mask must always strike the screen at points that form the corners of an equilateral triangle (*fig. 14*). Consequently the three beams must not only converge in each deflected position, but they must also at all times include the same angle when they arrive at the shadow mask. If the landing pattern does not form exactly the same equilateral triangle everywhere on the screen, this can be corrected by locally adjusting the distance between mask and screen.

Astigmatism of the deflection system causes non-similar deformation of the landing pattern. The deflection coils are therefore required to be anastigmatic. This implies that the coefficients B_4 and B_5 must be identical. Some astigmatism is permissible, however, since the phosphor dots on the screen have a greater cross-section than the beams that pass through the holes in the shadow mask.

The best practical compromise is achieved by making the horizontal and vertical deflection coils have the opposite astigmatism, so that the errors of the two pairs of coils cancel out in the corners of the screen. This does, however, give some astigmatism in the middle of the sides of the screen.

Slight corrections to the shadow mask and to the phosphor-dot pattern have traditionally been determined experimentally in one of the last phases of the design. However, for the 110° tube, with its greater deflection angles, it is essential to make a complete calculation of the electron paths for the outer edge of the screen, and the corrections required can therefore be established beforehand.

Another advantage of calculating the outermost electron paths is that we know exactly the outer boundary of the region swept by the beam. This enables us to choose the shape of the glass envelope in such a way that the deflection coils are situated as close as possible to the beam. The sensitivity of these coils can thus be made as high as possible, i.e. the currents required for a given deflection can be minimized.

The beam-indexing tube

This type of colour television tube (*fig. 15*) has only one electron gun, and the beam again produces a raster of horizontal lines on the screen [8]. The three colour phosphors are now not applied to the screen in a pattern of dots but in vertical strips separated by equally wide black strips.

The beam current is modulated in turn by the red, the green and the blue colour signal. After every two colour phosphor lines an ultra-violet fluorescent strip is applied, on the inside, to a black strip, and the fluorescence is detected outside the picture tube with a photomultiplier. The signal from this multiplier tube provides information on the position of the beam. This signal is used for synchronizing the landing point of the beam and the video signal, containing the information for the reproduction of the three colours, which modulates the beam.

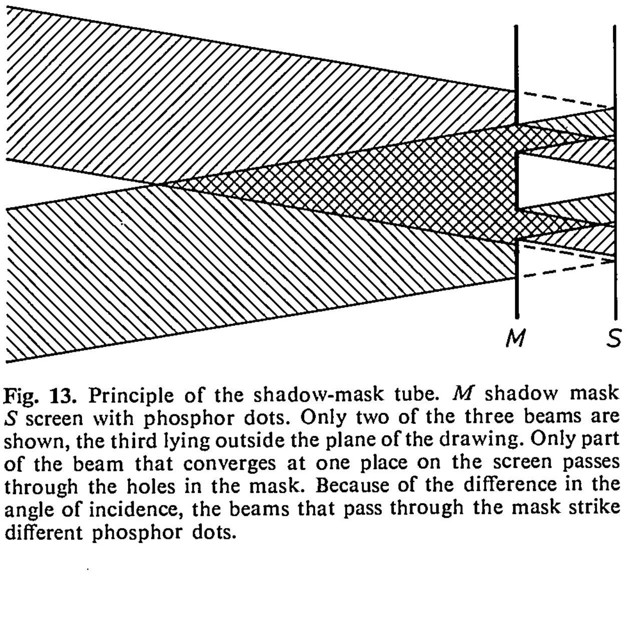


Fig. 13. Principle of the shadow-mask tube. *M* shadow mask *S* screen with phosphor dots. Only two of the three beams are shown, the third lying outside the plane of the drawing. Only part of the beam that converges at one place on the screen passes through the holes in the mask. Because of the difference in the angle of incidence, the beams that pass through the mask strike different phosphor dots.

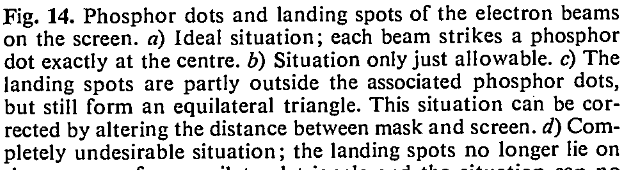


Fig. 14. Phosphor dots and landing spots of the electron beams on the screen. *a*) Ideal situation; each beam strikes a phosphor dot exactly at the centre. *b*) Situation only just allowable. *c*) The landing spots are partly outside the associated phosphor dots, but still form an equilateral triangle. This situation can be corrected by altering the distance between mask and screen. *d*) Completely undesirable situation; the landing spots no longer lie on the corners of an equilateral triangle and the situation can no longer be corrected by changing the distance between mask and screen.

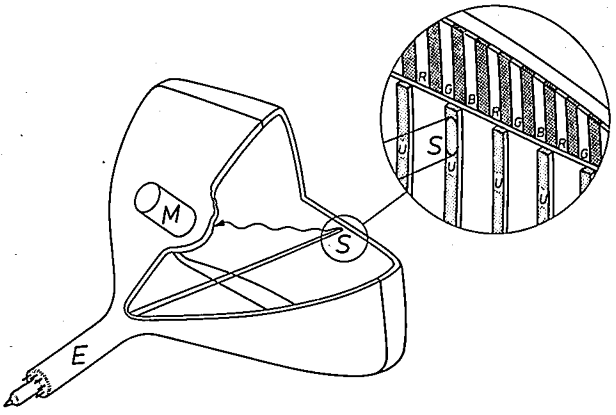


Fig. 15. Beam-indexing tube for colour television. The electron gun *E* produces a beam that gives an elliptical spot *S* on the screen. The deflected beam strikes consecutively the red, green and blue phosphor strips, *R*, *G* and *B*. The strips *U* are ultra-violet phosphor strips at the back of the screen. When the beam strikes these strips ultra-violet flashes are produced, which are detected outside the picture tube by the photomultiplier *M*. The pulses from the multiplier serve to synchronize the scanning motion of the beam with the intensity information for the three colours, which modulates the electron beam.

Since only one phosphor line at a time should be struck by the beam, the spot must be narrow in the horizontal direction. The vertical dimension of the spot, on the other hand, must not be too small if sufficient light intensity is to be obtained without saturating the phosphors. This is achieved by making the beam cross-section elliptical instead of circular at the point where the beam meets the screen.

The above requirement means that we must use an astigmatic beam and then ensure that the meridional image surface of the horizontal-deflection coil and the sagittal image surface of the vertical-deflection coil coincide with the screen.

Here, as previously, we shall only consider the consequences that this has for the vertical deflection. The situation for the horizontal deflection is entirely analogous.

If the screen of a picture tube has a radius of curvature of say 1 m, this will also be the radius of curvature ϱ_s of the sagittal image surface of the coil under consideration. We have seen that $1/(2\varrho_s) = B_5$, where B_5

[7] See Electronic Appl. **30**, 33, 1970 (No. 1).

[8] More details of this type of picture tube will be found in: E. F. de Haan and K. R. U. Weimer, Roy. Telev. Soc. J. **11**, 278, 1967; G. J. Lubben, Onde électr. **48**, 918, 1968; P.M. van den Avoort, Onde électr. **48**, 921, 1968.

is one of the coefficients of equation (11) for the landing point of a deflected electron path. We thus require our coil to give a field-strength distribution such that $B_5 = \frac{1}{2}$ m^{-1}. Since there is a relationship between the coefficients B, the result is that B_4, which as we saw earlier gives the curvature of the meridional image surface, becomes much greater than 1 m^{-1}. A situation like that shown in *fig. 16a* is undesirable because here the two image surfaces intersect somewhere on the screen. In the vicinity of this intersection the ellipse gradually changes into a circle, causing saturation of the phosphors at this location and thus giving rise to colour errors.

It is found, however, that the shape and position of the image surfaces are independent of one another. We can therefore allow the meridional image surface to lie completely inside the tube without it changing the curvature of this surface (fig. 16*b*). Intersections of the image surface cannot now occur and the cross-section of the beam remains elliptical everywhere on the screen.

It will be evident from these examples that each type of tube sets different requirements for its deflection coils. For a rapid understanding of the characteristics of particular deflection coils the third-order approximation is quite adequate. For a more accurate description the fifth-order approximation must be used, and for large deflection angles numerical calculations of the electron paths cannot be avoided. With automation of the measurements and computer processing of the results numerical calculations no longer present an insuperable difficulty.

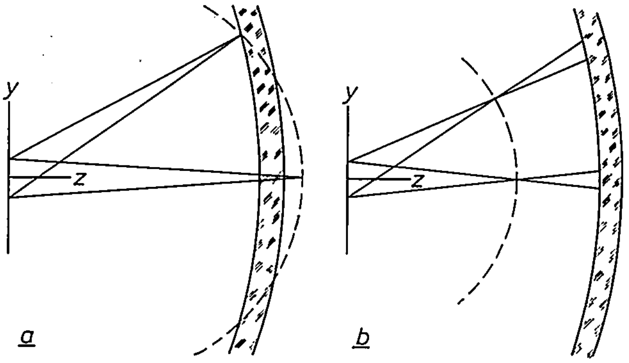


Fig. 16. The screen of a beam-indexing tube coincides with the sagittal image surface of the vertical deflection coils. If the meridional image surface lies in the centre of the screen outside the tube, then since the curvature of the meridional image surface is greater than that of the sagittal image surface, the landing point along the upper and lower edges of the screen will be circular (*a*). This results in local saturation of the phosphors. This undesirable situation is avoided if the meridional image surface lies completely inside the tube (*b*).

Summary. The components of the magnetic deflection field in a television picture tube can be expanded in series in terms of the coordinate positions relating to the axis of the tube. The coefficients of these series are in principle easy to measure. However, because of the relative magnitudes of the coefficients highly accurate measurements are required, and the results are most conveniently processed by computer. The first-order and third-order approximations of the solution of the equation of motion for an electron in the deflection field are given; the relations between the coefficients in the third-order approximation and the various image errors are discussed. The fifth-order approximation for the solution is briefly mentioned, but because of its complexity it is not discussed in detail. The approximations are not valid for deflections of the electron beam through angles greater than 40°, as in the 110° colour television tube. The field strength then has to be measured point by point to enable a numerical solution of the equation of motion for the electrons to be found. To illustrate the applications of the methods of measurement and calculation a description is given of problems arising in the design of two types of colour television tube, the shadow-mask tube and the beam-indexing tube.

Analysis and synthesis of handwriting

J. Vredenbregt and W. G. Koster

Although the work described here will very likely be of interest to designers of equipment for reading handwriting automatically, the authors did not set out with this aim in mind. Their work is in fact part of a more general study being made at the Institute for Perception Research (IPO) in Eindhoven on the motor system of the human body. The handwriting process is used for studying the programming of muscle activity in relation to the movement produced.

Introduction

Every movement made by a human being is the result of the precise coordination of a number of muscles. A movement carried out using only one joint is referred to as a single movement, one using more than one joint as a composite movement. The writing process is an example of a composite movement. The process is controlled by programmed activity of several groups of muscles, which produce the movements of forearm, hand and fingers. The process comprises virtually all aspects of the motor system of the human body. As a logical continuation of the studies on muscle mechanics [1] started some years ago at the Institute for Perception Research (IPO), and in pursuance of the work of J. J. Denier van der Gon [2] and of J. S. MacDonald [3], it was therefore decided to embark on a study of the production of handwriting [4]. It was reasonable to expect that a study of the static and dynamic behaviour of the muscles during writing would provide a better understanding not only of the writing process itself but also of the rules underlying the programming of composite movement processes. This could be useful in developing power-controlled artificial limbs and muscle stimulators [5].

The methodology underlying this investigation of muscle activity is comparable with that used in the phonetic research at IPO. An instrument was constructed for analysing the phenomenon — in this case the writing process — and a second instrument was built which simulates the writing movements of the hand and can be used to synthesize letter characters. Cursive handwriting can be simulated and slight changes made in the shape of characters by simply changing the time parameters. The success achieved in synthesizing characters is an indication that the hypothesis on which the instrument is based is not incorrect.

Analysis of the writing movement

The handwriting analyser we have constructed is an instrument that records a displacement-time diagram, i.e. a curve that gives the position of the pen as a function of time. In designing the analyser we regarded the writing movement as being composed of two separate movements: one in the direction of writing — which we shall refer to as the x-direction — and one in the direction perpendicular to it — the y-direction. The x-movement is the result of a rotation of the hand from the wrist or of the forearm from the elbow joint, or of a translational movement of the forearm. The movement perpendicular (or slightly oblique) to the writing direction is produced by the thumb, the index finger and the middle finger; sometimes it may be produced by the whole forearm. In addition to these two movements there is, of course, a third one, to put the pen on to the paper and to lift it, but for the present purposes we can ignore this.

In analysing the writing movement we have mainly studied the displacements in the y-direction, because they are larger than those in the x-direction and far less complicated. Preliminary studies had also shown

J. Vredenbregt and Ir. W. G. Koster are with the Institute for Perception Research, Eindhoven.

[1] J. Vredenbregt and W. G. Koster, Some aspects of muscle mechanics *in vivo*, IPO Annual Progress Report **1**, 94-100, 1966. J. Vredenbregt and W. G. Koster, Measurements on electrical and mechanical activity of the elbow flexors, Biomechanics I, 1st Int. Seminar, Zurich 1967, pp. 102-105 (Karger, Basle/New York 1968).

[2] J. J. Denier van der Gon and J. Ph. Thuring, Kybernetik **2**, 145, 1965.

[3] J. S. MacDonald, Quart. Progress Rep. M.I.T.-R.L.E. **76**, 210, 1965.

[4] See also: J. Vredenbregt and W. G. Koster, Analysis and synthesis of handwriting, IPO Annual Progress Report **2**, 157-161, 1967; J. Vredenbregt, W. G. Koster and J. W. Kirchhof, On the tolerances in the timing programme of synthesised letters, *ibid.* **3**, 95-97, 1968; W. G. Koster and J. Vredenbregt, Analysis and synthesis of handwriting, Biomechanics II, 2nd Int. Seminar, Eindhoven 1969, pp. 77-82 (Karger, Basle/New York 1971).

[5] See for example H. J. van Leeuwen and J. Vredenbregt, A muscle stimulator for hemiplegic patients, Philips tech. Rev. **30**, 23-24, 1969.

that further investigation of movements in the x-direction would not yield more information than the investigation of movements in the y-direction.

Fig. 1 gives a schematic view of the instrument used for mechanically recording the y-coordinate of the movement of the pen at point O as a function of time. The electrical signal that is a measure of the displacement in the y-direction is the voltage across the strain gauges Q, attached to one of the leaf springs S. Depending on the displacement of the pen the leaf springs are deflected to a greater or lesser extent by the disc R, thus varying the resistance of the strain gauges.

The construction of the instrument is kept as light as possible to ensure that the mass of the moving parts does not unduly affect the recording. Displacement of the pen in the x-direction is possible because the detection system for the y-coordinate just described can be moved as a whole along a guide shaft A. The movements in the instrument take place practically without friction because a cushion of air is maintained between the block B and the shaft A, and also where the shaft C passes through a hole in the block B. To ensure that block B does not stick because of non-coaxial movement when it is moved along the shaft, the displacement from O in the x-direction is transmitted to B by means of a thin nylon cord which passes over four rollers, i.e. twice around rollers P_1 and P_2.

The frequency characteristic of the analysis system is such that signals of the amplitudes normally encountered in handwriting and up to 25 Hz in frequency are reproduced without distortion. In the frequency range below 25 Hz there is no phase shift between input and output signals. The frequencies encountered in handwriting movements all lie below 15 Hz. An example of the displacement-time diagram for the y-direction of the letter a is shown in *fig. 2*.

A fundamental question to which our analysis of cursive handwriting could supply the answer reads: when different persons write the same character, is it possible to recognize general laws, unconnected with the person writing the character, from the recorded displacement-time diagrams? The results obtained so far indicate that this might indeed be the case. *Fig. 3* shows five characters a written by five subjects with the pen of the analysing instrument described above. Individual differences in the shape of the characters can clearly be seen. Nevertheless, there is a certain similarity of pattern in the displacement-time diagrams, in the times taken to write the successive parts of a character. On dividing the total writing time for the character into intervals of similar direction, by marking the time axis at the points where the y-movement reverses, it is found that the ratios between the successive intervals are about the same for all subjects. This is also

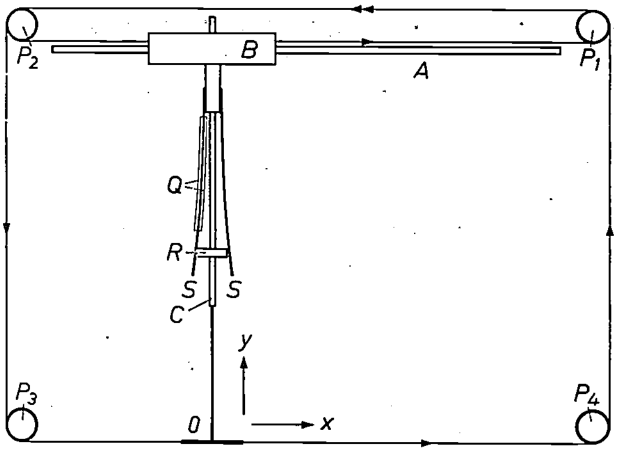


Fig. 1. Instrument for recording the writing movement. O position of the pen. Movements in the direction of writing (the x-direction) are transmitted by a nylon cord to the block B, which can move along the guide shaft A. Movements perpendicular to the x-direction (the y-direction) are transmitted via the shaft C, to which a disc R is fixed, to deflect two leaf springs S to a greater or lesser extent. Q strain gauges, which provide a voltage that is nearly proportional to the deflection of the pen.

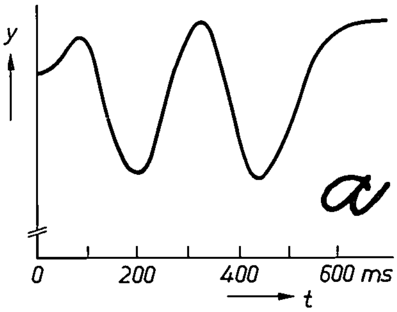


Fig. 2. Displacement-time diagram showing the movements of the pen perpendicular to the x-direction during the writing of the character a.

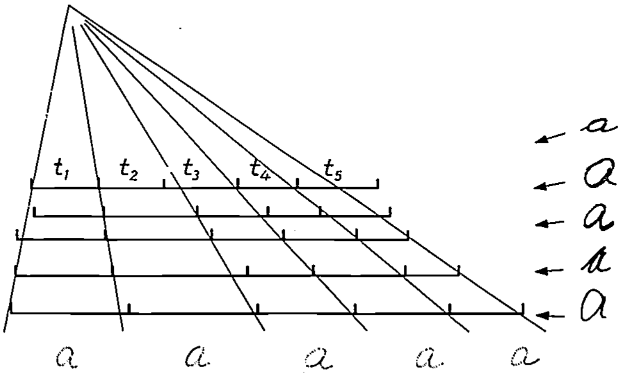


Fig. 3. The character a as written by five subjects, the writing time being divided into intervals t_1-t_5. The intervals are marked by the instants at which the y-component of the movement reverses direction, as seen from the displacement-time diagrams. The ratio between these intervals is virtually constant for all five subjects.

found to be the case for other characters. Thus, in spite of individual differences in the shape of the character, this ratio is more characteristic of the character than of the writer. The same constant ratio is found whether the subject writes a particular character quickly or slowly or large or small.

We shall now take a closer look at the way in which skeletal muscles bring about such movements. A muscle causing a movement in one direction is called the *agonist*, and that causing movement in the opposite direction is called the *antagonist*. When an agonist contracts, the part of the body involved is set in motion and accelerated for as long as this muscle exerts a force as a result of its activity. During this movement the antagonist is passively stretched, which takes up energy and thus offers resistance to the movement. In fast movements, moreover, the activities of the agonist and the antagonist are to some extent interdependent. When the muscular activity ceases, the movement continues for a time because of the characteristics of the muscle and the inertia of the mass of the part of the body. When the antagonist contracts, the movement slows down and changes direction.

Whether and to what extent a muscle is active is easily ascertained because muscular activity is associated with electricity. The electrical effect is due to ion exchange in the membranes surrounding the muscle cells, and with surface electrodes the electrical signals can be detected at the skin as a varying potential difference (electromyography).

Fig. 4 shows the displacement-time diagrams recorded during the writing of the character *a*, together with the associated electromyograms. The arrow on the right of the electromyograms indicates the direction of movement to which the various muscle activities relate. A comparison between the displacement-time diagram and the associated muscle activities shows that the muscular activity indicated by the electromyogram is usually of shorter duration than the movement it causes, and starts at the instants at which the movement must be started, slowed down or changed in direction. In "acquired" movements of this kind, agonist and antagonist are seldom active simultaneously.

Results such as those described above suggest the following hypothetical model for the writing process. Writing can be compared with the operation of a mechanical system that possesses a certain inertia and is controlled by an excitation pattern of short periods of unequal duration. The magnitude of the force, and hence the amplitude of the movement, depends more on the duration than on the strength of the excitation.

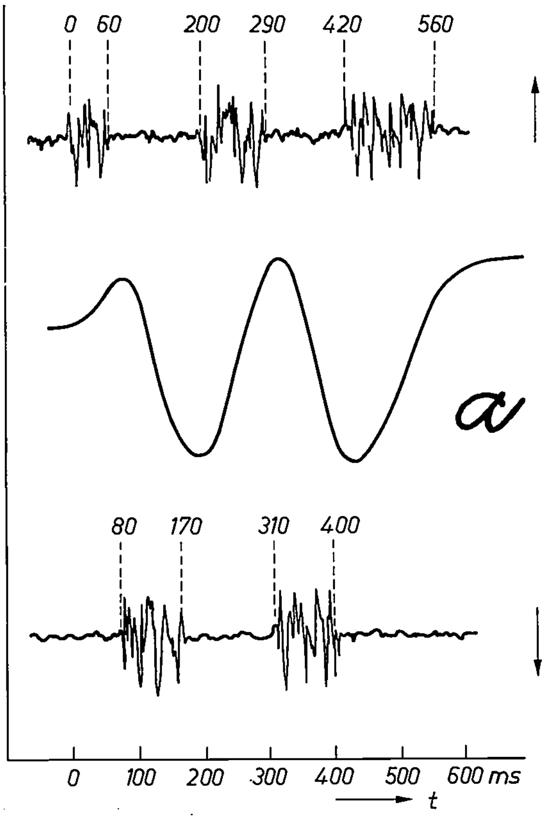


Fig. 4. Displacement-time diagrams for the movement perpendicular to the direction of writing, together with the associated electromyograms recorded during the writing of the character *a*.

Synthesis of written characters

A handwriting simulator

To test this model of the writing process we have built an electromechanical instrument that has the characteristics just described and also has the mechanical characteristics and limitations of the human hand (*fig. 5*). This instrument can indeed produce writing.

The energy is supplied by four d.c. electric motors with starting and stopping characteristics closely resembling those of skeletal muscles. The motors are mechanically coupled in pairs, one pair for moving the stylus in the two x-directions, the other pair for the movement in the two y-directions. The effect of this coupling is that one motor in a pair acts as the "*agonist*" while the other acts as the "*antagonist*". When a particular motor is actuated, its partner operates as a generator, and this energy is dissipated in a resistor, which provides damping for the system (*fig. 6*). The non-actuated motor thus fulfils the role of the passively stretched antagonist. The inertia required for the simulation is provided by the mass of the armature of the motor, the stylus and the other moving parts. The overall translation of hand and forearm along the line is simulated by means of a lead-screw, and the stylus is raised and lowered by means of an electromagnet.

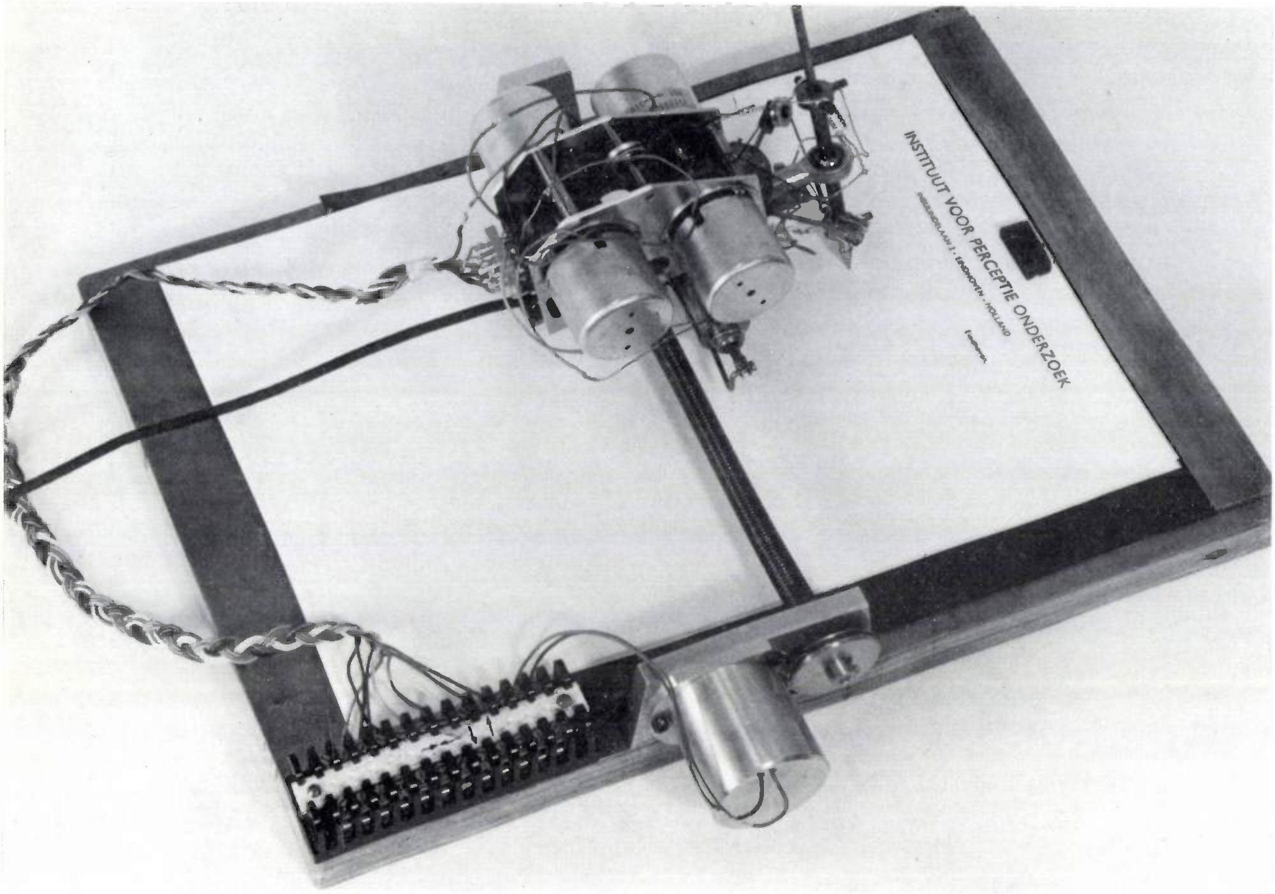


Fig. 5. The handwriting simulator. Two mechanically coupled d.c. motors cause the pen to move in the *x*-direction; the other coupled pair give the displacement in the *y*-direction. The whole system can be moved in the *x*-direction by means of a lead-screw. The pen is raised and lowered by means of an electromagnet.

All four motors are supplied separately with voltage pulses of constant amplitude and variable duration in accordance with a programme which is characteristic of the letter to be written. This is illustrated schematically in *fig. 7* for the character *a*. The exitation — corresponding to the amplitude of the supply voltage — is set such that the maximum velocity at which the stylus moves, both vertically and horizontally, is equal to the actual speed at which a hand-held pen would move.

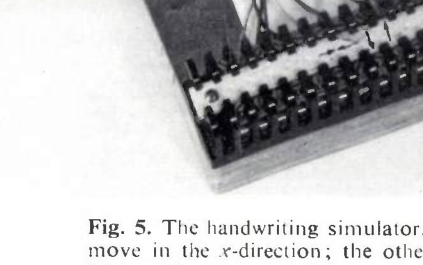


Fig. 6. Waveform of the terminal voltages of a pair of mechanically coupled motors. The actuation pulses from both motors are shown with no shading (e.g. pulse *1*). The shaded area *2* shows the waveform of the voltage generated when the motor continues to run on under the influence of its inertial mass. If the other motor is now actuated (pulse *3*), the first motor starts to turn in the opposite direction and a negative terminal voltage is generated (area *4*). The energy generated in the shaded periods is dissipated in a resistor, which provides damping for the system.

The programme in which the motors are actuated is preset by means of an electronic device [6] that delivers the commands for the actuation pulses at the appropriate instants. The device has various output channels, and the programmed selection of a given output channel determines which motor is actuated. The instant of actuation can be varied in steps of 1 millisecond. It is possible to make the periods of actuation of the various motors overlap.

Timing programme

From the shape of a character and the way in which it is produced when written by hand, the sequence in which the simulator motors have to be actuated may be derived. Determining the timing and duration of the actuation is a process of trial and error, with the "naturalness" of the character produced as the yard-

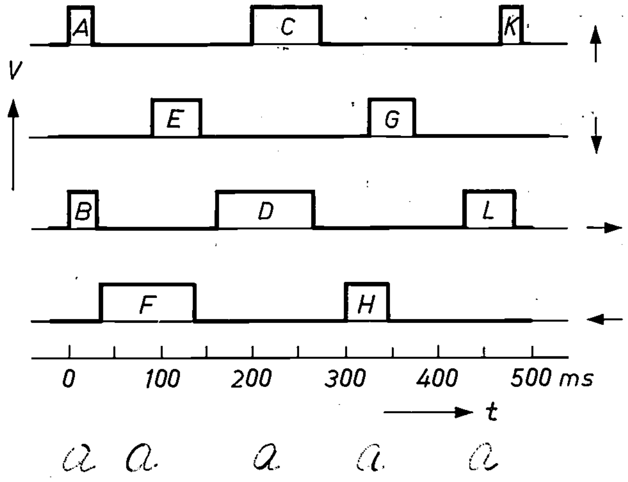


Fig. 7. A timing programme for synthesizing the character *a*. The supply voltage for the four motors is plotted schematically in the vertical direction. The periods of excitation are indicated by the letters *A* to *K*, and an arrow indicates the direction in which the particular motor causes the pen to move. The parts of the character corresponding to the periods are shown below the timing programme.

stick for judging the correctness of the programme. Using the knowledge gained from the analysis the programmer can soon acquire a certain skill.

Fig. 8 shows some characters written by the simulator, giving for each character the variation of the motor terminal voltages and the displacement-time diagrams of the pen. It can be seen that like muscular activity during writing the direction of movement is reversed just after the excitation is switched on, and that when the terminal voltage is switched off the inertia of the system causes the pen to continue moving for a short time. *Fig. 9* shows various simulations of the character *a* from fig. 3. It can be seen that the simulated characters are barely distinguishable from the ones shown in fig. 3. The differences that give handwriting its individuality can be simulated by making relatively small changes in the timing programme, i.e. by varying the starting time and duration of the pulses actuating the motors.

An alteration in the unit of time for the electronic programming switch makes all components of the programme proportionally longer or shorter, resulting in a larger or smaller character with otherwise the same shape. However, since the inertia of the system remains unaltered, larger variation in timing will affect the shape of the character.

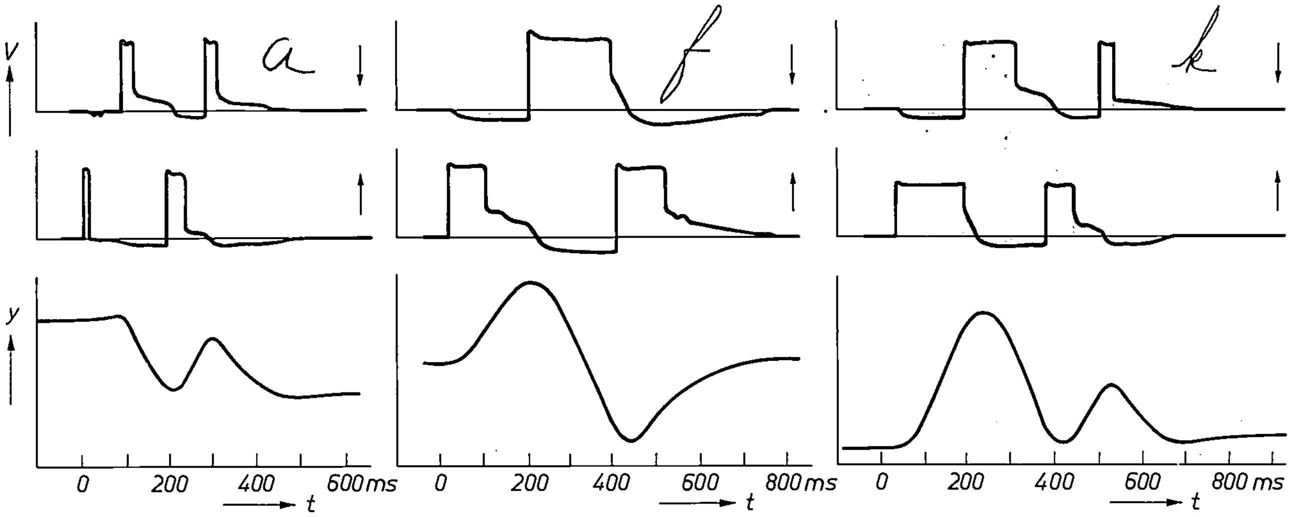


Fig. 8. Some characters produced by the simulator, with the associated displacement-time diagrams and the terminal voltages of the motors. In the shaded periods the motors act as generators.

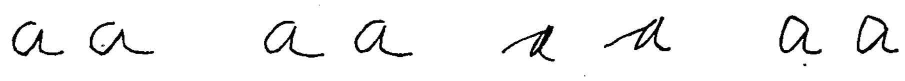


Fig. 9. Simulations of four of the five characters *a* from fig. 3. The differences in shape are produced by making slight changes in the timing programme of the simulator.

An interesting effect has been obtained by giving the motors a bias voltage. The conditions are then comparable with increased muscular tension, resulting in a kind of cramped writing. The characters written by the machine are no longer flowing but look somewhat

[6] G. J. J. Moonen and C. A. Lammers, A preset cascade counter, IPO Annual Progress Report 1, 104-106, 1966.

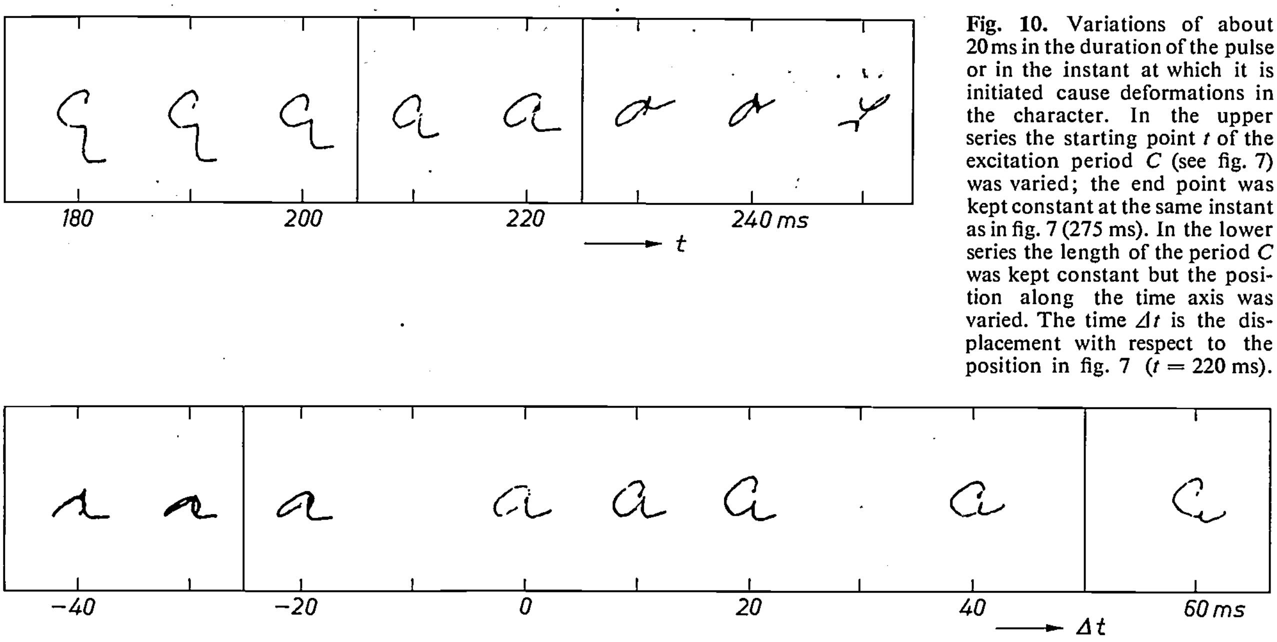


Fig. 10. Variations of about 20 ms in the duration of the pulse or in the instant at which it is initiated cause deformations in the character. In the upper series the starting point t of the excitation period C (see fig. 7) was varied; the end point was kept constant at the same instant as in fig. 7 (275 ms). In the lower series the length of the period C was kept constant but the position along the time axis was varied. The time Δt is the displacement with respect to the position in fig. 7 (t = 220 ms).

uncoordinated, as if they had been written by a person with the faulty coordination of movement found in some slightly spastic subjects.

A study of the effects of changes in the programme on the legibility of the characters showed that the tolerance for a *shift* of a period of excitation along the time axis is at least twice the tolerance for an alteration in the *length* of that period. The most critical periods in the programme for the character *a* appear to be *C*, *E* and *G*. The periods *C* and *E* directly affect the loop of the character *a*, particularly at the point where it should close and the downward stroke begins. The period *G* mainly controls the length of this downward stroke. As a rough guide, changes of 20 milliseconds in the duration of the excitation will often be enough to deform a letter beyond recognition (*fig. 10*).

From the investigations described we conclude that a natural-looking simulation of handwriting can be produced by a system in which the mechanical parameters are constant and only the time parameters of the actuating signal are varied. Some verification that the characters are produced in much the same way as in writing appears from the resemblance between the displacement-time diagram of the simulation and that of the movement of the hand, and also from the agreement between the pattern of the pulses actuating the simulator and the pattern of the electromyographic signals recorded during handwriting.

Since deviations of between 5 and 10 milliseconds in the simulation are enough to cause perceptible differences in the character, it is reasonable to assume that muscles are controlled with the same degree of accuracy.

Summary. The handwriting process was chosen as a topic of study in the context of research on muscular activity in the execution of composite movements. An instrument was built which records displacement-time diagrams during the writing of characters by hand. These diagrams correlate well with simultaneously recorded electromyograms, which give a picture of the muscle acitivity. The writing process can be compared with the action of a mechanical system that possesses a certain inertial mass and is controlled by constant excitation during a number of periods of unequal duration. To test this model a simulator has been built which can produce written characters. It consists of two pairs of d.c. motors which move a pen to and fro in two orthogonal directions. The motors are actuated with constant voltage pulses in accordance with a timing programme. Changes of 5 to 10 milliseconds in the programme characteristic of a particular character cause changes in the character that correspond to individual differences in handwriting. This makes it reasonable to assume that muscles are controlled with comparable accuracy.

Controlling the properties of electroceramic materials through their microstructure

G. H. Jonker and A. L. Stuijts

To some, it might appear that ceramic fabrication is simply an economical method of producing materials which, though possessing acceptable properties, would really have been ideal in the less practicable single-crystal form. It is indeed costly and difficult, if not impossible, to make single crystals of the materials discussed in this article. There are also cases where a single crystal would in fact be ideal, and where the ceramic product is at best an acceptable substitute. In many cases, however, a ceramic process is essential, offering additional degrees of freedom that will make new combinations of properties possible. It is this aspect that is the really fascinating part of modern ceramic technology.

Introduction

Throughout the ages the ceramist has worked to produce ceramic objects that are pleasing in form and of adequate mechanical strength. The two greatest threats to the strength of his final product — a sintered aggregate of powder particles — were porosity and the internal stresses that arise because more than one phase is present. These phases can be crystalline phases that differ chemically or crystallographically, or an amorphous, vitreous phase (see *fig. 1*). The ceramist's ultimate aim could perhaps be expressed as the production of a *single-phase* product sintered *completely solid.*

Since the forties ceramic technology has been developing from an empirical method of fabricating materials to one that is based on scientific knowledge. During this "renaissance" in ceramic technology new fields of application have been opened up, and these developments have been matched by an increasing refinement of the ceramist's objectives [2]. It steadily became clearer that not only the mechanical strength but other properties as well (mechanical, electrical, magnetic) depend to a great extent on the ceramic "microstructure", and that this microstructure has other important aspects apart from porosity and the number of phases. These aspects include the size of the constituent crystallites (or grains), their size distribution, the boundaries between them, and whether or not the crystallites are oriented ("texture"). The modern ceramist therefore tries to give his product an *optimum microstructure* for the properties he has in mind.

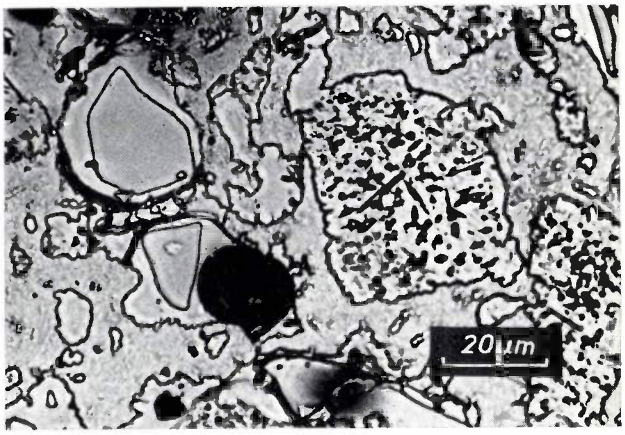


Fig. 1. The microstructure of porcelain, a typical conventional ceramic product. (After S. T. Lundin [1].) A matrix of vitreous clay in which many clay particles are still undissolved contains two quartz crystallites, which can be seen on the left, surrounded by a dissolved edge. A large pore (black) appears to the right of the lower crystallite, and further to the right there is a large residue of feldspar showing the incipient separation of mullite crystals.

This applies particularly to ceramic materials made for electrical applications. These are dielectric, magnetic and semiconducting oxides which often possess a com-

Prof. Dr. G. H. Jonker, formerly with Philips Research Laboratories, Eindhoven, is now Professor in the Technology of Electronic Materials at Twente University of Technology; Prof. Ir. A. L. Stuijts is with Philips Research Laboratories, Eindhoven, and is also Professor Extraordinary in the Technology of Inorganic Materials at Eindhoven University of Technology.

[1] S. T. Lundin, Electron microscopy of whiteware bodies, Trans. IVth Int. Ceramic Congress, Florence, Italy, 1954.

[2] A. L. Stuijts, Renaissance in ceramic technology, Philips tech. Rev. 31, 44-53, 1970. This article is referred to in the following as I.

plex chemical composition. Unlike traditional ceramics, they do not usually contain more than traces of material in a vitreous phase.

To be in a position to optimize the microstructure it is necessary both to understand the relations between microstructure and material properties and to be able to produce the microstructure required. The growing success in the production of the desired microstructures is due to new understanding of the sintering phenomena and to the development of new ceramic processes. In what follows we shall deal with these first [3]. We shall then describe some specific applications of modern electroceramic materials, each of which illustrates the exploitation of a different aspect of the microstructure.

It will become clear that the development of modern electroceramic materials of high quality is possible only through close cooperation between ceramist, physicist, chemist and electrical engineer.

For certain applications the single-crystal form would be ideal. However, making single crystals of the materials discussed here, with their complex chemical composition, is extremely costly and difficult, if not impossible. It is therefore of great practical importance that *ceramic* products that possess very good, though not ideal, properties for these applications can now be made from these materials. The modern ceramist, however, finds a greater fascination in those cases in which it is essential to have a polycrystalline structure.

Modern ceramic methods

The requirements nowadays imposed on both the starting powder and the sintering procedure are still primarily those relating to a densely sintered product. To achieve this, the densification process during sintering must take place as rapidly as possible. This not only simplifies the sintering procedure itself, it also prevents unwanted side-effects, in particular grain growth, from gaining the upper hand and ruining the result.

During the densification (the filling of the pores) in the sintering process material is transported. As discussed at length in I, material transport during sintering is largely due to *bulk diffusion.* Closer investigation shows that the *curvature* of the pore-crystallite interface, especially at the point where two grains meet, is essential to the driving force behind this process. In general terms the process can be said to be due to the decrease in the surface energy that occurs when the total free crystallite surface area decreases. This means that the densification rate usually increases with the crystallite surface area, i.e. decreases with grain size. *The starting powder must therefore be very fine.*

For a given size of crystallite it is also possible, especially in ionic compounds, to increase the rate of sintering by increasing the speed of diffusion, a possibility which depends on the application of *defect chemistry* [4]. As the diffusion mainly takes place via vacancies in the crystal lattice, in compounds it is necessary to create the right concentration of vacancies for each type of ion. In particular, the diffusion can be speeded up by increasing the number of vacancies for the slowest ion. This is done by introducing an exactly controlled deviation from the stoichiometric composition or by adding a controlled amount of a doping agent. On the other hand, unwanted impurities can easily spoil the result of the sintering. To apply this knowledge we must therefore have a starting powder which is not only fine and homogeneous but is also of *extremely high chemical purity.*

Doping can also be used to combat the complication of grain growth which we noted earlier. This growth starts as soon as the densification is so far advanced that large contact areas have formed between the crystallites. Just how drastic this process can be was discussed in I (see I, fig. 7). Grain growth is undesirable because it hinders further densification; moreover, as we shall see later, a small grain size is often one of the particular attributes required of the microstructure. The process can be opposed by adding a "grain-growth inhibiter", an additive which settles at the grain boundaries and checks their mobility, e.g. a vitreous component that envelops all the crystallites with a thin film.

The traditional method of powder preparation, in practice still used almost exclusively today, is not very suitable for obtaining a fine powder which at the same time is chemically extremely pure (*fig. 2a*). In this method a mixture of powders of the constituent materials is first prefired and a solid-state reaction gives the actual starting material to be formed. This prefiring inevitably involves coarsening of the powder grains. After having been prefired the powder therefore has to be reground, and in practice it is equally inevitable that impurities enter the powder because of wear in the grinding equipment. New methods of powder preparation are therefore required.

We shall briefly discuss three new methods that are now being intensively studied: co-precipitation, freeze-drying and spray-drying. These are all wet-chemical methods, in which a solid mixture of salts is prepared from a liquid solution, by either precipitation or evaporation of the solvent, and is then decomposed by heating.

The *co-precipitation* method is already frequently used on a laboratory scale for preparing barium titanate and associated compounds as well as ferrites. One example is the preparation of nickel ferrite. A solution is made of salts containing the Ni and Fe ions in the appropriate proportions, and this is mixed with a solution of sodium carbonate. Nickel-iron carbonate precipitates from the mixture, and the resultant powder is filtered and then washed to remove Na ions. Heating of the powder results in a mixed Ni-Fe oxide, from which nickel ferrite is obtained at relatively low temperature (e.g. 300 °C). In this way a powder can be obtained whose particles are smaller than 10 nm (fig. 2*b*).

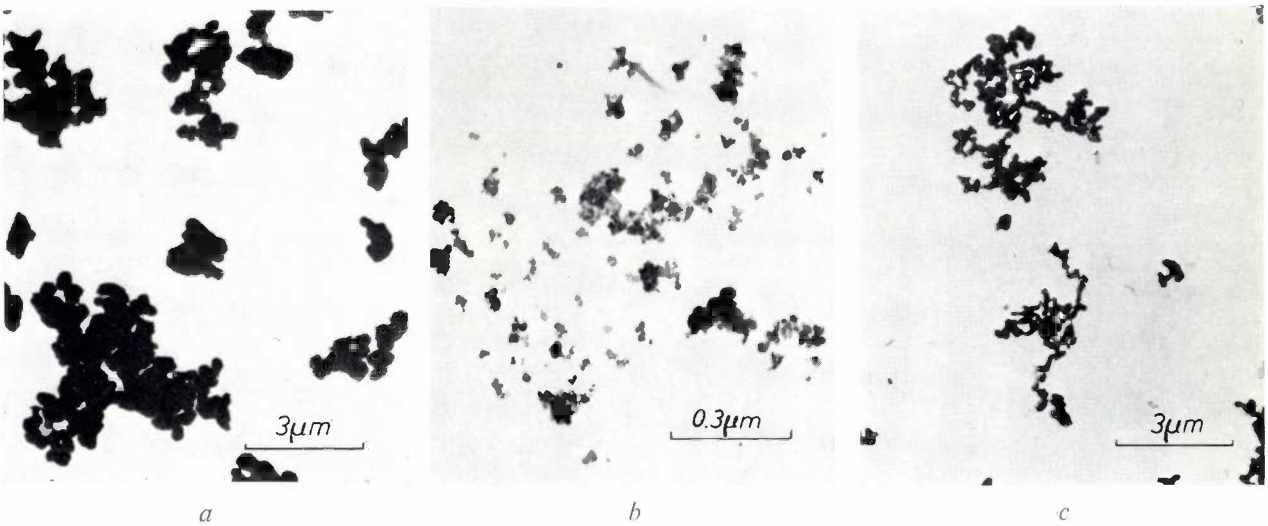


Fig. 2. Electron-photomicrographs of ferrite powders, from left to right: prepared by the conventional method, obtained by co-precipitation, obtained by spray-drying. When spray-drying is used, aggregates of particles are formed. In spite of these aggregates the powder possesses a very high sinter reactivity and can be made completely solid. (There is a difference in magnification between the centre picture and the others.)

In the *freeze-drying* method [5] a finely divided spray of an aqueous solution of salts is directed into a low-temperature bath of a fluid such as hexane. Immediate freezing of the droplets prevents segregation of the ingredients. The water is then removed from the frozen droplets by sublimation in a freeze-drier. The resultant mixture of salts is then heated to obtain the required compound, as in all three methods discussed here.

Finally, in *spray-drying* [6] a finely divided spray of the solution is brought into direct contact with a hot stream of gas (fig. 2*c*). As the droplets are very small (10 to 20 μm) the water is very quickly evaporated so that segregation is limited to a minimum.

The new developments we have so far described all relate to the preparation of the starting powder. A development relating to the sintering process itself is the method known as *hot pressing* [7]. The application of pressure assists the densification process, so that the required densification is obtained *in a shorter time* or *at a lower temperature*. Among the very important advantages of this process there is first of all the possibility of suppressing grain growth, which takes place more readily at a higher temperature. For example, only with hot pressing is it possible to keep the grain size below 1 μm and at the same time to reduce the porosity to lower than 1 % (*fig. 3*), a requirement which, as we shall see later, is nowadays imposed on microwave ferrites. In the second place this is the only way in which certain important materials can be sintered effectively, since at the higher sintering temperature normally required the substance would decompose or components would volatilize.

Once we are able to fabricate a "difficult" product (e.g. one with negligible porosity and small grains), it is obvious that we can also make products that are "less difficult" (e.g. products that are slightly porous and have medium-sized grains). This brings us back to our starting point, which was the achievement of a prescribed microstructure. In this respect the hot-pressing technique looks particularly promising for electroceramics: the microstructure can be controlled very accurately by varying the pressure, the temperature and the length of time in the die.

Defect chemistry of the materials, mentioned earlier, can also be applied for obtaining microstructures of

[3] A more extensive review of new ceramic processes is given in: A. L. Stuijts, New fabrication methods for advanced electronic materials, Science of Ceramics 5, 335-362, 1970.

[4] P. J. L. Reijnen, Nonstoichiometry and sintering in ionic solids, in: Problems of nonstoichiometry (ed. A. Rabenau), North-Holland Publ. Co., Amsterdam 1970, pp. 219-238.

[5] See F. J. Schnettler, F. R. Monforte and W. W. Rhodes, Science of Ceramics 4, 79, 1968.

[6] See J. G. M. de Lau, Amer. Ceramic Soc. Bull. 48, 509, 1969.

[7] See G. J. Oudemans, Continuous hot pressing, Philips tech. Rev. 29, 45-53, 1968. Further literature on hot pressing is listed in the article by A. L. Stuijts [3].

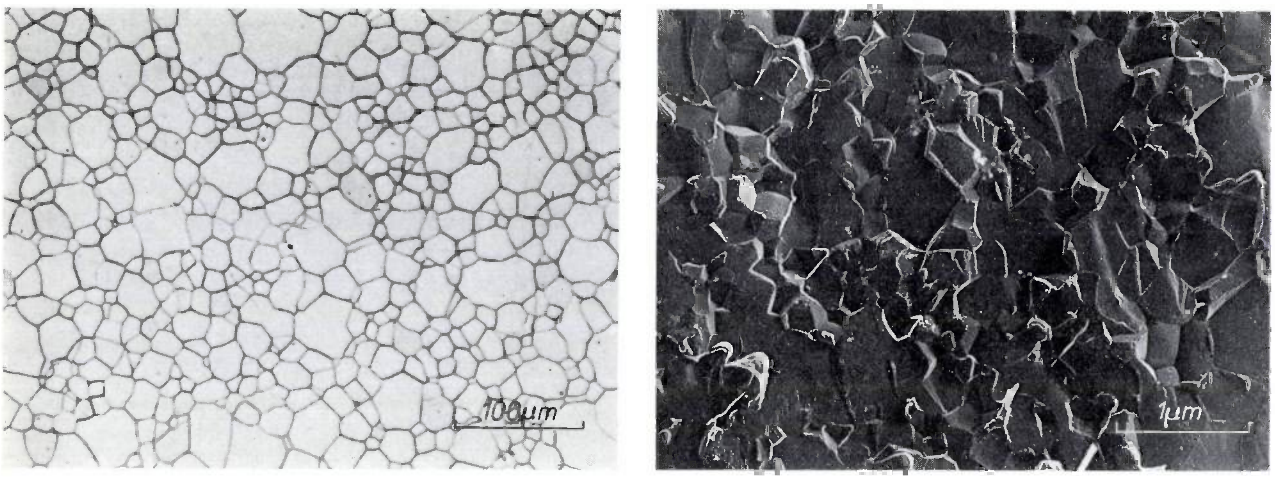


Fig. 3. Photomicrographs of densely sintered ferrites. *a*) Polished and etched surface of a ferrite prepared by spray-drying and conventional sintering. *b*) Electron photomicrograph of the fracture face of a hot-pressed ferrite. In case (*a*) the diameter of the crystallites is about 30 µm, in case (*b*) about 1 µm.

widely different types. Typical examples are shown in *fig. 4*: in one case there are numerous small pores *inside* the grains, in the other there are large pores *between* the grains. In the latter case the pores are carried along with the grain boundaries during the sintering and are swept together. These two structures arise, respectively, in ferrites which are slightly deficient in anions or cations with respect to the stoichiometric composition. We shall return later to the importance of this.

Microstructure and properties

The relation between the microstructure and properties of a ceramic material will be illustrated in the following sections with a number of applications of modern electroceramics. Broadly the picture is as follows.

In *ceramic dielectrics* (e.g. the various titanate oxides) the situation is very straightforward. Grain boundaries are not an unwanted imperfection in this case, and pores, provided they are not freely connected with one another and with the exterior, cause only a dilution effect of minor significance. "Open" porosity, however, is detrimental, since it enables the material to absorb moisture, which greatly increases the dielectric losses. These materials will not be of any further concern to us here.

The materials we are especially interested in are the *ceramic ferromagnetics* (ferrites) and the *ceramic ferroelectrics*, where the relation between microstructure and properties assumes a variety of forms. The crystallites in these materials are split up, below the "Curie

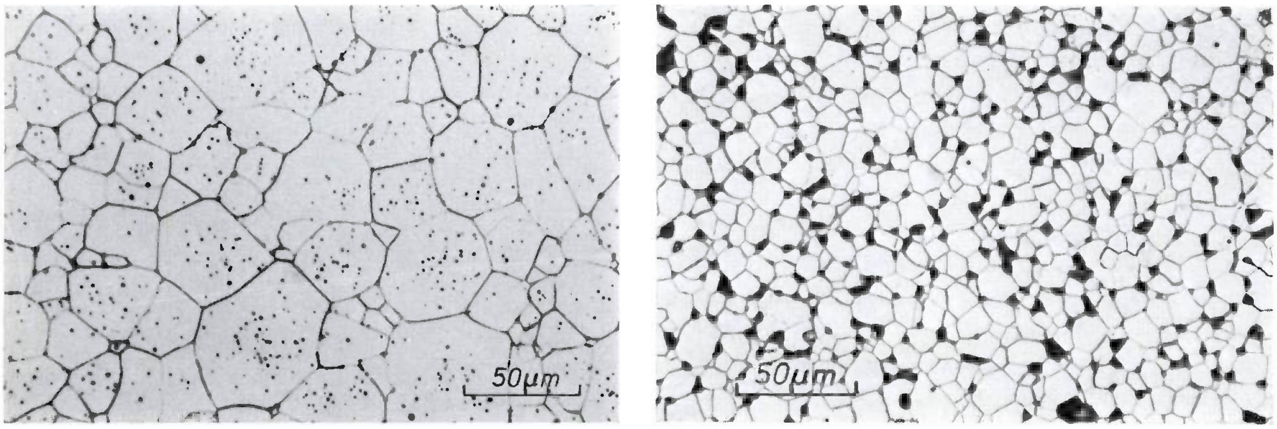


Fig. 4. Two microstructures (photomicrographs of etched surfaces). *a*) Ferrite with an anion deficit. The pores are small, and are located *inside* the crystallites. *b*) Ferrite with a cation deficit. The pores have been "swept" together: they are larger than in (*a*) and are all situated *between* the crystallites.

temperature", into domains, regions of spontaneous electric or magnetic polarization (*fig. 5*). The mobility of the domain (or Bloch) walls is an important factor that can often be influenced through the microstructure. Mobile domain walls favour "soft" properties (e.g. a high permeability or a low coercive field strength, a property that in general corresponds to easy change in the magnetic or electric state), and the material can be given this property by making the crystallites large and fault-free. Small crystallites, on the other hand, favour "hard" properties, such as remanent magnetism. The domain walls then appear to be fixed, or may perhaps be completely absent. Pores can also be introduced to "pin" the walls.

The influence of the microstructure is very pronounced in barium titanate [8], for many years the most familiar example of a ferroelectric. In ceramic form it is only a good ferroelectric (suitable for piezoelectric use) when its crystallites are large enough to consist of many domains. If, on the other hand, the crystallites are small, the ferroelectric character is completely suppressed, as also are the hysteresis losses, so that the material is then a very useful dielectric. Finally, *semiconducting* barium titanate can be used to form an excellent PTC (positive temperature coefficient) thermistor by creating a large number of grain boundaries occupied by oxygen. These boundaries form current barriers that ensure the required behaviour. We shall return later to these subjects at greater length.

As opposed to semiconducting barium titanate, other *oxidic semiconductors* are widely used as NTC thermistors. This application is based on the negative temperature coefficient of the resistance of the material itself. Any perceptible effect which the grain boundaries have on the resistance is now an unwanted effect, so that the problem here is to ensure that *no* current barriers form at the grain boundaries.

Of the various aspects of the microstructure, we have said most about the size of the crystallites. We have also mentioned the grain boundaries, and seen that in the one case current barriers are avoided and in the other case utilized. Another example where current barriers are made use of is that of ferrites for the kHz frequency range, where grain-boundary resistances are used as "internal lamellae" to reduce eddy-current losses.

A third aspect, porosity, was referred to as a means of pinning domain walls. As a rule, however, pores are regarded as undesirable because they decrease domain-wall mobility, cause internal depolarization and make the internal field inhomogeneous. In the preparation of oxidic semiconductors *open* porosity assists the adsorption or desorption of oxygen by the grain boundaries, and whether or not this is desired depends on whether or not resistance barriers are required. We have already mentioned that open pores in dielectrics are detrimental.

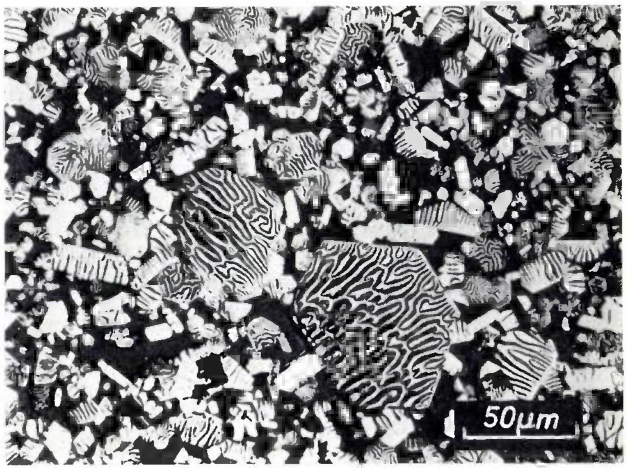


Fig. 5. Domain walls in ferroxdure crystallites. The crystallites are embedded in a vitreous medium (black). The large crystallites required for this photograph were obtained by prolonged firing of a composition of this type. The black lines in the crystallites are the domain walls, made visible by means of a suspension o a finely distributed magnetic powder which concentrates on the walls.

Finally, texture is important in materials where the magnetization has a strong preference for one particular crystallographic axis. By orienting the crystallites along this axis the remanent magnetization can be increased. This treatment is applied in ferroxdure.

The following discussion of various specific properties and applications is arranged in terms of materials and areas of application. This means that the ferromagnetic and ferroelectric materials will be treated separately, even though the effects in these materials are closely analogous.

Ferrites

Before taking a look at a number of ferrite products it will be useful to comment briefly on the starting materials. For, however much we may be concerned with the microstructure, we must not forget that the choice of the material itself — the chemical composition, which implies a certain crystal structure — is obviously the primary consideration when we have a particular application in view [9]. For instance, the Mn-Zn ferrites are appropriate starting materials when soft magnetic properties are required. These materials are "soft by nature", their spins being only weakly tied to a preferred crystallographic direction; in other words, the

[8] See G. H. Jonker, Ber. Dtsch. Keram. Ges. **44**, 265, 1967.

[9] A detailed treatment of the relation between magnetism, microstructure and crystal chemistry is given in: A. Broese van Groenou, P. F. Bongers and A. L. Stuijts, Mat. Sci. Engng. **3**, 317, 1968/69.

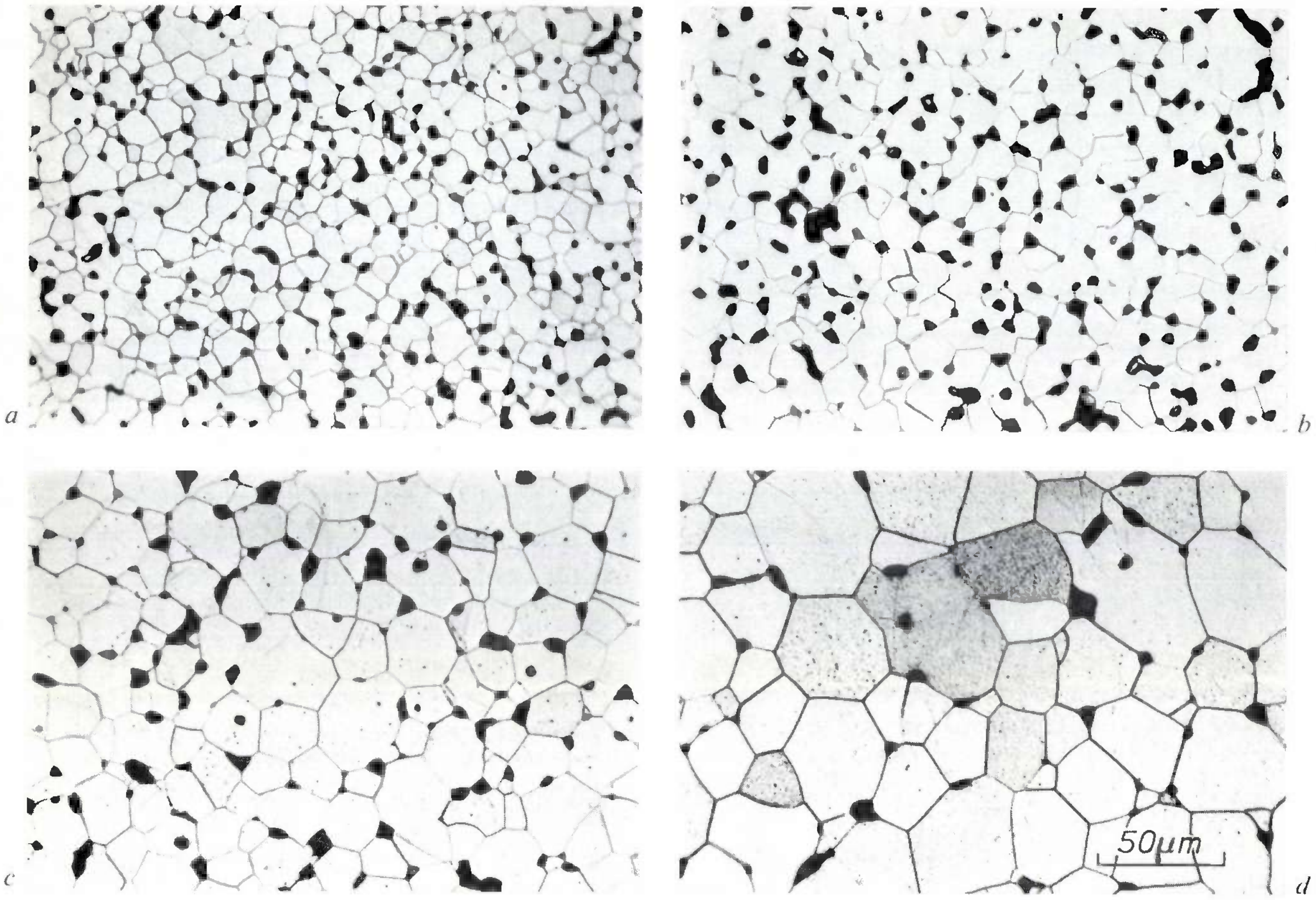


Fig. 6. The microstructure of high-permeability ferrites. (After D. J. Perduijn and H. P. Peloschek [10].) The structure is of the type shown in fig. 4*b*: the pores are situated entirely on the grain boundaries. The relative initial permeability μ_i is higher the larger the average grain diameter: *a*) μ_i = 6500, *b*) 10 000, *c*) 16 000, *d*) 21 500. Larger grains are obtained by more prolonged firing. The magnification is the same in all four cases.

magneto-crystalline anisotropy of the material is weak. The Ni-Zn ferrites are magnetically somewhat harder, and some hexagonal ferrites are extremely hard. The magnitude of the crystalline anistropy is an intrinsic property of the material. In addition the material may show magnetic "stress anisotropy" as a consequence of mechanical stresses if the material is magnetostrictive. A measure of the total magnetic anisotropy is the anisotropy field H_A. A spin is bound to its preferred or "easy" direction as if there were a magnetic field H_A in that direction. The value of H_A ranges from less than one oersted in the softest ferrites to roughly 10000 oersteds in the very hard ferrites. (1 Oe = 79.6 A/m.)

The magnetic anisotropy of the starting material is reflected in a variety of ways in the final product. Looking, for example, at the permeability, a weak anisotropy is manifested in the first place in the fact that the spins are easily deflected from their preferred direction by an external field ("spin rotation"): the material is then said to have a high "rotation permeability". In high-permeability ferrite products, however, the main contribution to the permeability does not come from rotation permeability but from "wall permeability". This is a measure of the ease with which the polarization changes as a result of wall displacements, where one domain grows at the expense of another. Unlike the rotation permeability, the wall permeability is of course very sensitive to the microstructure; it does however also depend on the intrinsic properties of the material. This is so because a domain wall is a transitional layer between two domains, in which the magnetic polarization gradually changes from one preferred direction to another. In the middle of the wall the polarization deviates considerably from the preferred direction, and the wall therefore contains energy. This energy clearly increases with the anisotropy. High-energy walls are less easily created and adhere more strongly to obstacles; they are thus less mobile than walls with lower energy. Consequently a high wall permeability can only be produced in materials possessing low magnetic anisotropy.

High-permeability ferrites

To have a high permeability the material must contain highly mobile domain walls. This implies that the crystallites must be large and perfect, i.e. without pores or other crystal imperfections, and the crystalline anisotropy must be small.

As we have said, the Mn-Zn ferrites are suitable in this case because of their small crystalline anisotropy. In practice the anisotropy can be reduced even further by substituting Fe^{2+} ions for a small fraction of the Mn^{2+} and Zn^{2+} ions, in other words by making mixed crystals of the original ferrite with Fe_3O_4. In the correct mixture the average crystalline anisotropy as a function of temperature may even go through zero at room temperature. Moreover Fe_3O_4 has a high positive magnetostriction and therefore partly compensates the low negative magnetostriction of the unmixed Mn-Zn ferrites.

We now return for a moment to the two porosity structures of fig. 4, which can be obtained by introducing deviations from the stoichiometric composition: a) a large number of small pores *inside* the crystallites, and b) a small number of large pores *between* the crystallites. Although, with the same sintering procedure, the crystallites are smaller and the total volume of pores larger in case (b) than in case (a), the microstructure of (b) nevertheless gives higher permeabilities because the perfect crystallites in (b) are larger than the distances between the pores in (a). If we now carefully choose the sintering procedure so as to obtain large crystallites in case (b) as well, very high permeabilities can be achieved (see *fig. 6*) [10]. It can be seen from *fig. 7* that the permeability rises linearly with crystallite diameter. Ferrites with a relative permeability of more than 10 000 are now commercially available.

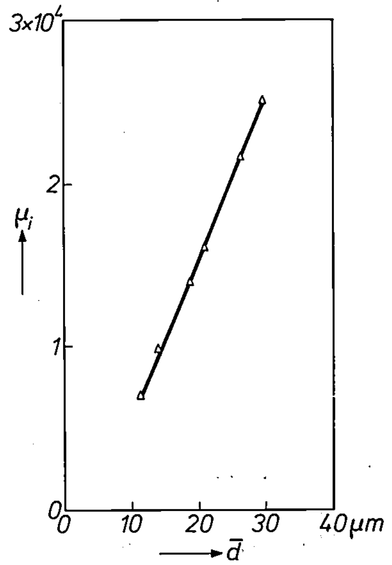


Fig. 7. Relative permeability μ_i as a function of average grain diameter $\bar{d}$. (After D. J. Perduijn and H. P. Peloschek [10].)

These considerations might lead to the conclusion that single crystals would really be ideal. In a constant or slowly varying field this is true. It is no longer true, however, at higher frequencies of the field, because losses then occur [11]. The polycrystalline structure is now essential to prevent these losses. The problem and the answer differ in the various frequency ranges. Some familiar solutions of practical importance are mentioned below.

The kHz range

In the frequency range up to about 0.1 MHz the Mn-Zn ferrites form a good starting point. Substituting Fe^{2+} ions in the lattice as mentioned above gives the ferrite a relatively low resistivity, because the extra electron of the ferrous ion can easily jump from one Fe site to another. Because of this, eddy-current losses soon occur as the frequency rises. These losses can be suppressed by means of the "internal lamellae" referred to earlier. These are introduced by adding to the starting material a small quantity of a silicate mixture, which settles on the grain boundaries during sintering, forming an insulating layer. Although this grain-boundary layer causes a slight decrease in the permeability μ, the loss factor tan δ decreases so much more that the net result is a better (i.e. lower) value of the technically significant figure of merit $(\tan \delta)/\mu$ [12].

The MHz range

In the MHz range, two types of loss occur, one of which is essentially due to "ferromagnetic resonance". The atomic magnetic moments responsible for the rotation permeability precess around the local internal field (the Larmor precession); in many cases this field consists mainly of the anisotropy field H_A. The closer the frequency of the r.f. field approaches the natural frequency of the precession (the resonant frequency), the more strongly the precession is excited and the greater is the associated dissipation of energy. The resonant frequency is proportional to the local field strength, i.e. roughly proportional to H_A. The Mn-Zn-Fe^{2+} ferrites can no longer be used, because the anisotropy and hence H_A is made small, so that the resonant frequency is low (about 1 to 10 MHz, depending on the Mn-Zn ratio) and the resonance losses already become perceptible at relatively low frequencies. The Ni-Zn ferrites have stronger internal fields, and the ferromagnetic resonance there-

[10] D. J. Perduijn and H. P. Peloschek, Proc. Brit. Ceramic Soc. **10**, 263, 1968; see also, by the same authors, IEEE Trans. MAG-4, 453, 1968.

[11] See section 7, "Damping" (p. 364) of the article by A. Broese van Groenou *et al.* [9].

[12] T. G. W. Stijntjes, A. Broese van Groenou, R. F. Pearson, J. E. Knowles and P. Rankin, Effects of various substitutions in MnZnFe ferrites, Proc. Int. Conf. on Ferrites, Kyoto, Japan, 1970, pp. 194-198.

fore lies at much higher frequencies (10 to 1000 MHz). This means that these ferrites are a better starting material for megahertz applications, although their permeability is smaller. In addition, the resistivity is much higher, so that eddy currents present no problem.

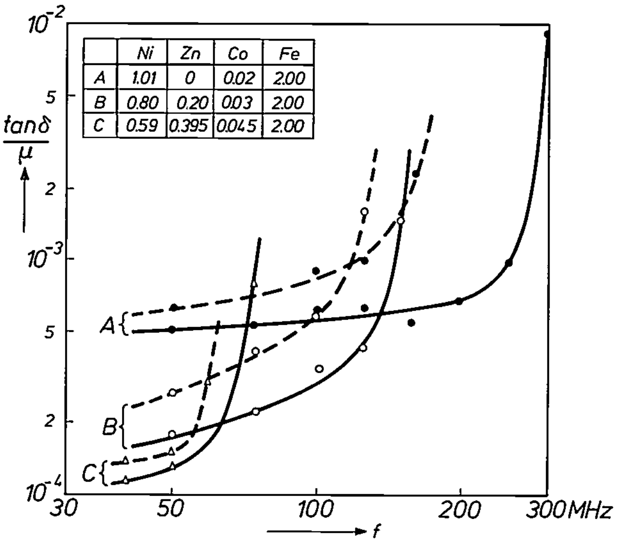
At frequencies that are about a tenth of the resonant frequency — i.e. in the MHz range for a Ni-Zn ferrite — high losses may occur in the form of damping of domain-wall movements. To avoid this contribution to the losses the contribution of wall displacements to the permeability should be suppressed and the magnetization process should be restricted to spin rotations. This can be achieved through the microstructure, by making the crystallites small enough to minimize the mobility of any domain walls that may be present [13]. The classical means of doing this is the addition of a "grain-growth inhibiter", but considerably better results have been obtained by hot pressing (*fig. 8*).

The GHz range

Ferrites for microwaves (the GHz range), which are mainly used in microwave circulators and phase shifters, are an entirely different case. The ferromagnetic resonance here is turned to practical application, but now in material that is saturated by an external field; the spins are thus oriented, and at resonance they precess together around this field. In applications of these ferrites it is important that a microwave magnetic field in the x-direction, with a frequency near resonance, excites polarization not only in the x-direction but also in the y-direction, provided that the total local field (the sum of the external and the internal field) is in the z-direction. This is a direct consequence of the precessional nature of the spin motion. Since the absorption is too high at the top of the resonance curve, it is necessary to use this effect on the slope of the resonance curve. This method yields useful results only if the line is narrow, i.e. if the total field, which locally determines the resonant frequency, varies only slightly with position.

Pores are therefore to be seen as undesirable, because they make the internal field inhomogeneous and hence broaden the line. Another irregularity is formed by the randomly oriented anisotropy fields. Attempts have been made to improve this by orienting the crystallites. In the ferrites used the ⟨111⟩ axis is a preferred direction, and by producing a ⟨111⟩ texture it has in fact been found possible to obtain a smaller line-width [14]. At the present time, however, the removal of the crystalline anisotropy and magnetostriction by means of suitable substitutions in the starting material is proving more successful, and then the only resonance-determining factor is the applied field.

With the microwave ferrites the situation so far discussed is one in which a single-crystal structure would seem to be ideal: it has no pores and the axes are completely oriented. Here again, however, there is one effect that would make a single crystal completely useless. This is the parametric excitation of spin waves at higher microwave powers. We shall not go into this effect here, but just note that it is a non-linear effect, so that the dissipation associated with it shows a disproportionate increase with increasing power. These spin waves are strongly damped if the material is finely distributed [15]. If the grain size is reduced from about 15 μm to about 2 μm the microwave power which the material can handle without additional high losses is increased by a few orders of magnitude (*fig. 9*). This is another application in which hot pressing has proved to be a highly effective method of obtaining the optimum microstructure, i.e. small crystallites and low porosity.



	Ni	Zn	Co	Fe
A	1.01	0	0.02	2.00
B	0.80	0.20	0.03	2.00
C	0.59	0.395	0.045	2.00

Fig. 8. Comparison of the high-frequency losses in a normally sintered and in a hot-pressed material, indicated by broken and solid lines respectively, for three Ni-Zn ferrites with Co doping. (After J. G. M. de Lau [13].) δ loss angle, μ relative permeability f frequency. The beneficial effect of hot pressing is greatest near the ferromagnetic resonant frequency (where the curves rise steeply).

Ferrites for permanent magnets

A good permanent magnet is characterized by a high coercive field strength and a high remanent polarization (*fig. 10*).

A high coercive field strength is obtained, as will be clear from what we have said above, by starting from a highly anisotropic material and making from it a product that has small crystallites. The high anisotropy will give high-energy domain walls and in small crystallites there will either be no domain walls or they will be firmly pinned to barriers. In the hard ferrites the grains can be kept small enough for this with the conventional sintering process, i.e. without exerting pressure on the material. Nevertheless, in all practical materials, it is domain-wall displacements that finally cause a reverse in polarization when the field strength is increased. This is evident from the fact that the coercive field is always considerably weaker than the anisotropy field H_A that would be needed to change the polarization of the spins through "uniform spin rotation" (about 10 000 oersteds in the hard ferrites). A coercive field strength closely approaching the anisotropy field has been found only in very exceptional cases, for example in iron whiskers [16], assumed to consist of virtually ideal single crystals.

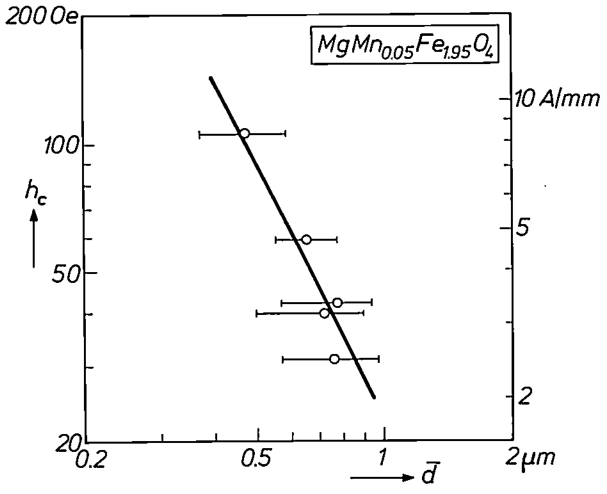


Fig. 9. Effect of the average grain diameter $\bar{d}$ in a microwave ferrite on microwave absorption at higher powers. (After Q. H. F. Vrehen, H. G. Beljers and J. G. M. de Lau [15].) The critical amplitude h_c of the microwave field above which the microwave absorption shows a disproportionate increase is plotted vertically. This increase is attributed to the excitation of spin waves. When the grains are smaller the spin waves are more strongly damped, giving the material a higher power-handling capacity.

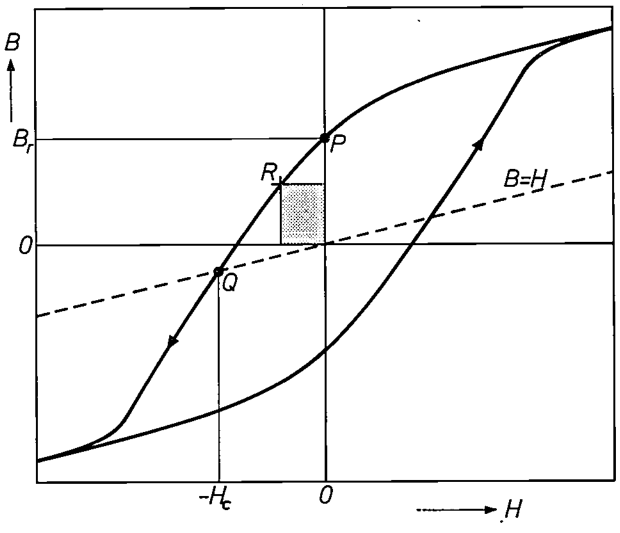


Fig. 10. Hysteresis loop in the B-H plane. A good permanent magnet is characterized by a high remanence B_r and a high coercive field strength H_c. In the Gaussian system, $B = H + 4\pi I$, where I is the polarization. For $H = 0$ (at point P) the flux density and the polarization have the remanent values B_r and I_r ($B_r = 4\pi I_r$). The absolute value of the field strength H reaches the coercive field strength H_c for $I = 0$, and hence for $B = H$ (point Q). A frequently used measure of the quality of a permanent magnet is $(BH)_{max}$, which is the maximum of $|BH|$ on demagnetization, i.e. when the loop goes through the section PQ. This maximum is reached at R, and is equal to the area of the grey rectangle. In SI units, $B = \mu_0 H + J$. The dashed line then corresponds to the equation $B = \mu_0 H$, and there is no factor 4π between induction and polarization at remanence ($B_r = J_r$).

If the material has a high uniaxial anisotropy the remanent polarization can be improved by means of a crystallographic texture in which the crystallites are oriented along the preferred axis of polarization. After saturation, on reducing the field the local polarizations generally return to the closest preferred axis, and if this axis is oriented the polarizations remain oriented as well. A single crystal, which might be thought to be better in this respect, cannot of course be used since its coercive field is weak because it is split up into domains.

Strong, uniaxial magneto-crystalline anisotropy is found in some hexagonal ferrites, in which the hexagonal axis is the preferred direction of magnetization. The classic example is $BaFe_{12}O_{19}$, known in its ceramic form as "ferroxdure" [17]. The orientation has been brought about in this case (see I, fig. 4) by introducing a suspension of finely ground powder particles into a magnetic field, filtering them and then compacting them [18]. A fortunate circumstance here is that the subsequent sintering process actually enhances the degree of orientation, since the few crystallites that are out of alignment are incorporated in their neighbours during the crystallization process. This process is controlled with the aid of an appropriate grain-growth inhibiter (which may again be a silicate mixture). The improvement that can be obtained in this way is illustrated in *fig. 11*, which presents a comparison between isotropic and crystal-oriented ferroxdure.

Ferrites for memory systems

The ferrites used for core stores in computers resemble permanent magnets in that they have to possess well defined remanent polarization that will withstand field fluctuations (in this case small), while at the

[13] J. G. M. de Lau, Proc. Brit. Ceramic Soc. **10**, 275, 1968 and IEEE Trans. MAG-5, 291, 1969.

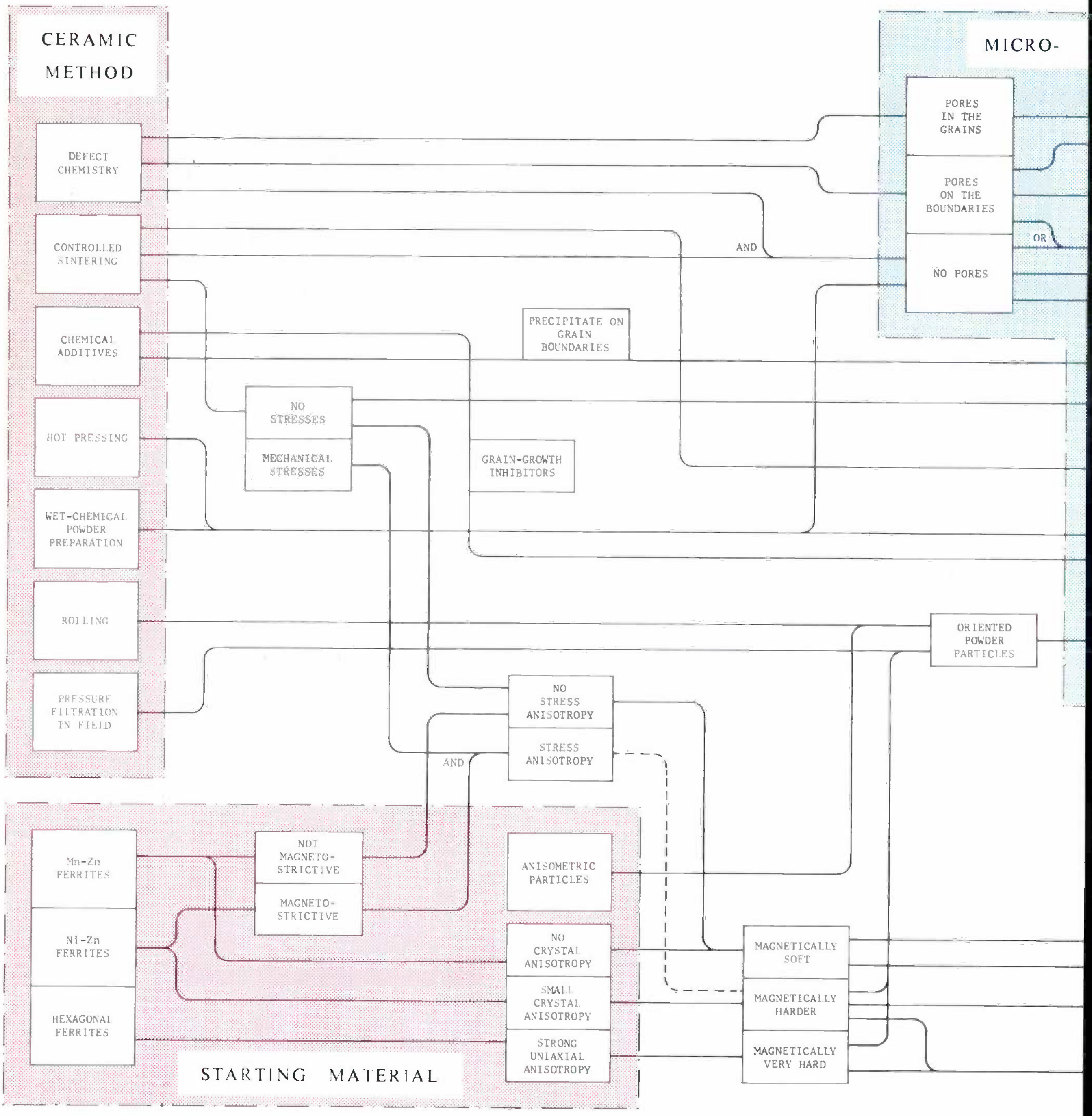
[14] A. Broese van Groenou, H. G. Beljers and F. C. M. Driessens, J. appl. Phys. **40**, 1424, 1969.

[15] Q. H. F. Vrehen, H. G. Beljers and J. G. M. de Lau, IEEE Trans. MAG-5, 617, 1969.

[16] R. W. DeBlois and C. P. Bean, J. appl. Phys. **30**, 225S, 1959.

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

[18] A. L. Stuijts, G. W. Rathenau and G. H. Weber, Philips tech. Rev. **16**, 141, 1954/55.


CERAMIC METHOD
DEFECT CHEMISTRY
CONTROLLED SINTERING
CHEMICAL ADDITIVES
HOT PRESSING
WET-CHEMICAL POWDER PREPARATION
ROLLING
PRESSURE FILTRATION IN FIELD
MICRO-
PORES IN THE GRAINS
PORES ON THE BOUNDARIES
NO PORES
AND
OR
PRECIPITATE ON GRAIN BOUNDARIES
NO STRESSES
MECHANICAL STRESSES
GRAIN-GROWTH INHIBITORS
ORIENTED POWDER PARTICLES
NO STRESS ANISOTROPY
STRESS ANISOTROPY
AND
Mn-Zn FERRITES
Ni-Zn FERRITES
HEXAGONAL FERRITES
NOT MAGNETO-STRICTIVE
MAGNETO-STRICTIVE
ANISOMETRIC PARTICLES
NO CRYSTAL ANISOTROPY
SMALL CRYSTAL ANISOTROPY
STRONG UNIAXIAL ANISOTROPY
STARTING MATERIAL
MAGNETICALLY SOFT
MAGNETICALLY HARDER
MAGNETICALLY VERY HARD

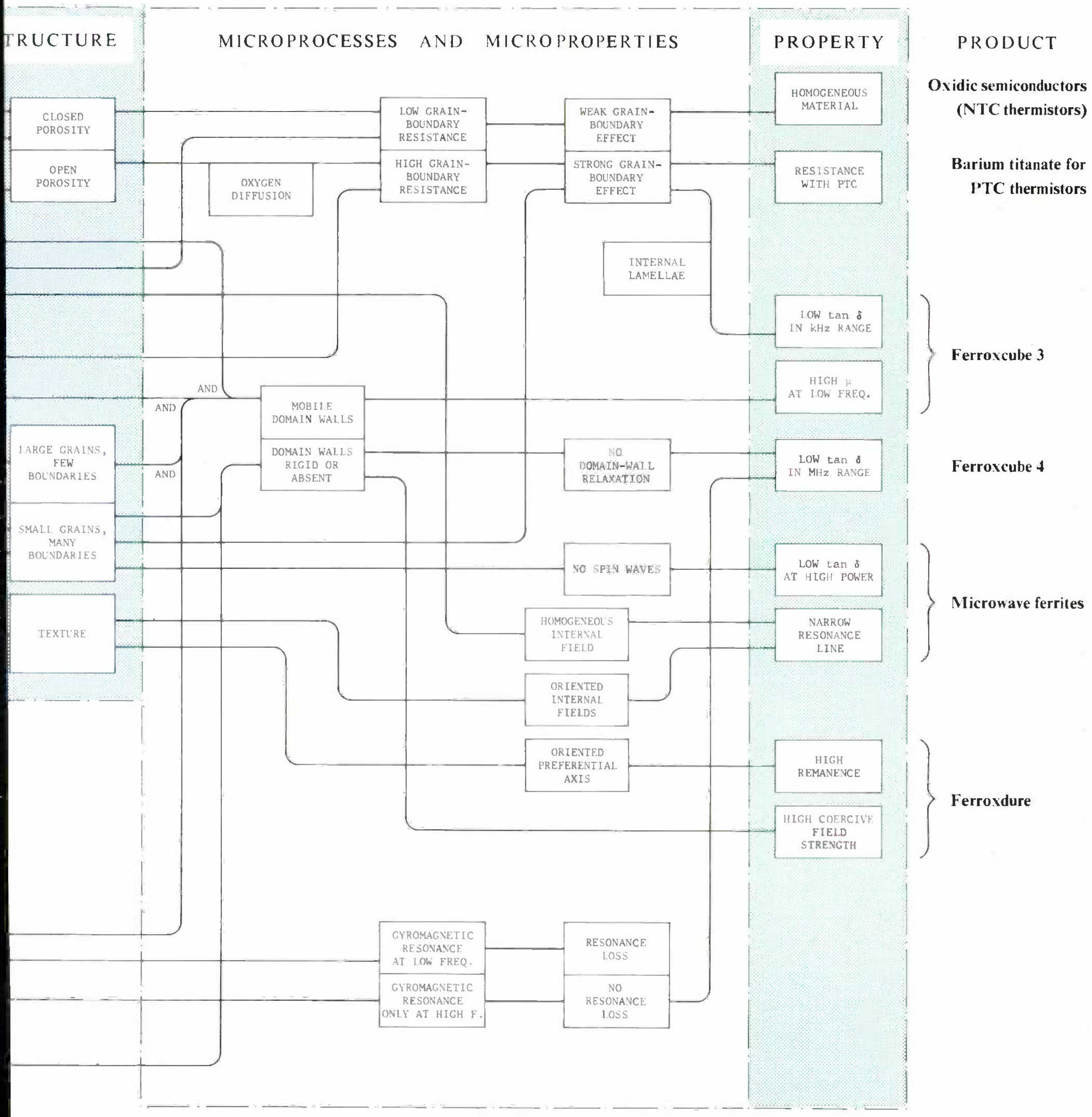

TRUCTURE
MICROPROCESSES AND MICROPROPERTIES
PROPERTY
PRODUCT
CLOSED POROSITY
OPEN POROSITY
OXYGEN DIFFUSION
LOW GRAIN-BOUNDARY RESISTANCE
HIGH GRAIN-BOUNDARY RESISTANCE
WEAK GRAIN-BOUNDARY EFFECT
STRONG GRAIN-BOUNDARY EFFECT
HOMOGENEOUS MATERIAL
RESISTANCE WITH PTC
Oxidic semiconductors (NTC thermistors)
Barium titanate for PTC thermistors
INTERNAL LAMELLAE
LOW tan δ IN kHz RANGE
HIGH μ AT LOW FREQ.
Ferroxcube 3
AND
AND
AND
MOBILE DOMAIN WALLS
DOMAIN WALLS RIGID OR ABSENT
LARGE GRAINS, FEW BOUNDARIES
SMALL GRAINS, MANY BOUNDARIES
TEXTURE
NO DOMAIN-WALL RELAXATION
LOW tan δ IN MHz RANGE
Ferroxcube 4
NO SPIN WAVES
LOW tan δ AT HIGH POWER
HOMOGENEOUS INTERNAL FIELD
NARROW RESONANCE LINE
Microwave ferrites
ORIENTED INTERNAL FIELDS
ORIENTED PREFERENTIAL AXIS
HIGH REMANENCE
HIGH COERCIVE FIELD STRENGTH
Ferroxdure
GYROMAGNETIC RESONANCE AT LOW FREQ.
GYROMAGNETIC RESONANCE ONLY AT HIGH F.
RESONANCE LOSS
NO RESONANCE LOSS

Diagram of relationships, see p. 94

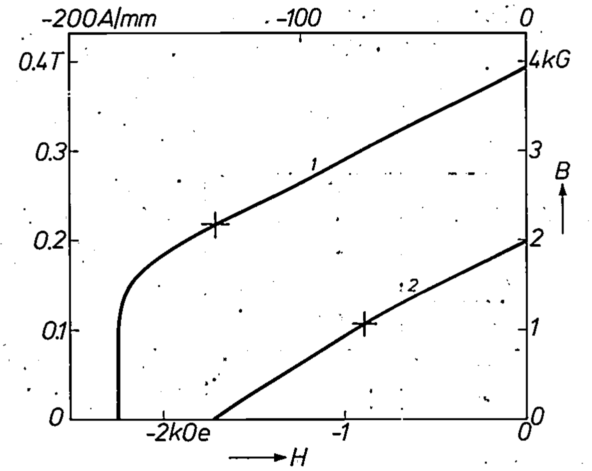


Fig. 11. Demagnetization curves of grain-oriented ferroxdure (*1*, after F. G. Brockman [19]) and of isotropic ferroxdure (*2*). The cross indicates where $|BH|$ reaches the value $(BH)_{\max}$. In the grain-oriented material $(BH)_{\max}$ is about 4 times as high as in the isotropic material.

same time, as in high-permeability ferrites, it must be possible to reverse the spins by a weak field [20].

Two wires run through a ferrite ring. The magnetization of the ring must reverse when both wires are energized, but not when only one of the wires is energized; in other words, the reversal must take place at a field $H = H_{\rm m}$, but not at $H = \frac{1}{2}H_{\rm m}$. Broadly speaking, this means that the hysteresis loop must be rectangular and that the coercive field strength $H_{\rm c}$ must lie between $\frac{1}{2}H_{\rm m}$ and $H_{\rm m}$; therefore $H_{\rm c} \approx \frac{3}{4}H_{\rm m}$. The switching field $H_{\rm m}$ is thus determined by the coercive field strength: $H_{\rm m} \approx \frac{4}{3}H_{\rm c}$.

The switching speed is of particular importance. The switching or reversal process is a kind of frictional process whose speed increases with the difference between the "stimulus" $H_{\rm m}$ and the "threshold" $H_{\rm c}$. For the switching time τ — which is inversely proportional to the switching speed — we may therefore write:

$$1/\tau = (H_{\rm m} - H_{\rm c})/s. \qquad (1)$$

We shall return presently to the proportionality constant s. Combining this expression with the value of $H_{\rm m}$ found above, we have:

$$1/\tau \approx H_{\rm c}/3s. \qquad (2)$$

The coercive field strength thus determines not only the switching field but also the switching speed. The optimum value of the coercive field therefore depends on the design in which the cores are to be used.

For the ferrite technologist this means that it must be possible to select the coercive field strength, and this can be done — as follows from the above — by means of the grain size: the smaller the crystallites, the greater the coercive field strength. Domain walls are apparently more strongly pinned when there are more grain boundaries.

To give a clear distinction between "switching" and "no switching" of the magnetization the sides of the hysteresis loop must be steep. The threshold fields for the various obstacles to domain-wall movement must therefore have about the same value; this means that *homogeneity* is of great importance here, and more especially a *uniform grain size.*

Finally there is the question as to whether, for a given coercive field strength, the switching speed can be further increased, in other words whether the constant s in equations (1) and (2) can be reduced. In the ferrites at present available the smallest values found for s are in the region of 0.3 μsOe. It is evident that the switching speed is mainly determined by two factors: the domain-wall velocity v and the distance the walls travel. The velocity v is found by experiment to be given approximately by the relation $v = a(H - H_{\rm c})$, and is thus proportional to the "driving force". This indicates that the wall movement is inhibited by a kind of frictional process. Although it is not yet clear what factors determine the value of a — which is of the order of 10 m/sOe — it looks as if there is not much that can be done to change it. At the few oersteds applied here the wall velocity thus remains limited to a few metres per second. What might perhaps be done is to reduce the distances which the walls have to travel by creating a large number of nucleation centres for domain walls. A small grain size, possibly combined with a low and finely distributed porosity, would favour this.

Ceramic ferroelectrics with perovskite structure

In dealing with the ferroelectrics we shall confine ourselves to the perovskites, compounds whose structure corresponds to the mineral perovskite ($CaTiO_3$). These compounds have a very high dielectric constant. They are also piezoelectric in a particular temperature range, a property directly related to the ferroelectricity. For years barium titanate was the best known ferroelectric material. It is still very important for dielectric applications, but for piezoelectric applications mixed crystals of lead titanate and lead zirconate are nowadays preferred.

The ferroelectrics are remarkably like ferromagnetics in many ways: they also give spontaneous polarization, hysteresis, a domain structure, wall displacements, etc., below a Curie temperature $T_{\rm C}$ (in the case of barium titanate this is between 120 and 130 °C). In other ways they are of course different, and we shall now briefly deal with some of these differences.

In the first place there is a difference in the nature of the primary polarization process. In the ferromagnetics the process is one of "spin rotation", i.e. orientation of permanent dipole moments, the "spins". In the ferroelectrics it is a displacement of the mean centre of the positive ions in the crystal lattice with respect to the mean centre of the negative ions. In the (field-free) dielectric state above T_C these centres coincide, but in the ferroelectric state below T_C they spontaneously assume a certain separation. In the spontaneously polarized ferroelectric state, the crystal lattice maintains a high sensitivity to further polarization by an external applied field, so that this state still has a high dielectric constant, of the order of several hundred. In a single-domain ferromagnetic state, on the other hand, all the spins are oriented, and there can be no further increase in polarization.

Another difference, but only a quantitative one, has a considerable effect on ceramic practice; this is the extent to which a crystal changes shape when going from the unpolarized state into a single spontaneously polarized domain. This change of shape is very slight in most ferrites, but in the perovskites it is considerable. In the dielectric state the perovskites are all cubic, a ferroelectric domain is tetragonal or rhombohedral. For example, a single-domain crystal of barium titanate at room temperature is tetragonal with a c/a axial ratio of 1.01. This difference of 1 % from the cubic structure implies a change of shape on polarization which is a few orders of magnitude greater than in the ferromagnetics. In lead titanate the axial ratio on polarization changes by as much as 6 %.

This effect is of paramount significance for the properties of the ceramic material. In a piece of sintered barium titanate that has been cooled to below the Curie temperature the crystallites are wedged tightly between their neighbours; spontaneous polarization can only take place if the external form of the crystallites is almost exactly preserved. Under favourable conditions the material spontaneously satisfies this condition by forming a fine pattern of domains where the longitudinal axis of each domain differs by 90° from that of its neighbour, i.e. where there are 90° walls between the domains (*fig. 12*). Crystallites are then formed with ridges on the surface. In the most favourable case the ridges of two neighbouring crystallites fit together. In this kind of polarization, however, stresses are bound to occur, as appears from the line broadening invariably observed in X-ray diffraction. Crystallites may be too small to split into domains and may then often be incapable of being polarized, since in their constrained position they cannot achieve the required change of shape.

The last difference between ferroelectrics and ferromagnetics that we shall mention is the behaviour of the domain walls. Although the walls in the perovskites can also move at a considerable velocity under favourable conditions, most of these materials exhibit an ageing effect in which the walls become immobile and the domain pattern rigidifies. As we shall see, in certain applications it is considered satisfactory if the domain pattern can be changed just once. This rigidification has hitherto limited the application of ferroelectrics as switchable storage elements. Recently compositions have been produced in which it is possible to avoid this effect [22].

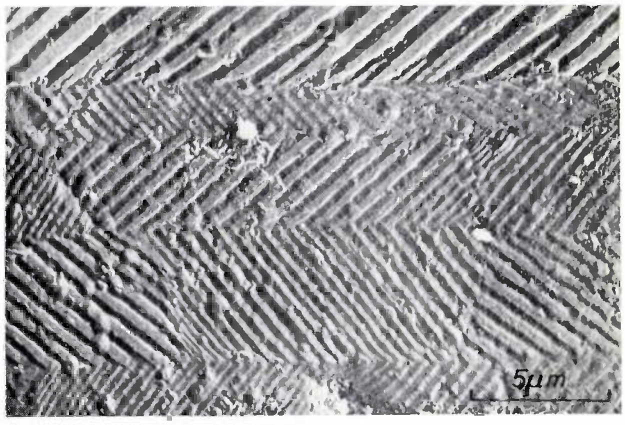


Fig. 12. Electron-photomicrograph of an etched surface of coarse-grained ceramic barium titanate. (After G. H. Jonker and W. Noorlander [21].) The domain pattern, in which neighbouring domains differ from one another in their direction of polarization by 90°, shows up because there is a difference in etching speed between the positive and negative sides of a domain.

Ceramic piezoelectrics

A perovskite crystal consisting of a single spontaneously polarized domain is piezoelectric: the electric polarization changes when the crystal is subject to mechanical pressure. Virgin ceramic material does not possess this property. Since the domain polarizations are randomly oriented, the changes in polarization when a uniform pressure is applied are also random, and the net polarization remains zero. In a polycrystalline material, therefore, the domain polarizations have to be oriented if a piezoelectric effect is required; in other words, the material must be given a remanent polarization. This is done by subjecting it once to a strong electric field.

This "poling" process is only effective, however, if the crystallites are large enough for them to consist of a large number of domains, for only then is it possible to change the polarization while preserving the shape fairly accurately. Moreover, small crystallites often

[19] F. G. Brockman, Amer. Ceramic Soc. Bull. **47**, 186, 1968.

[20] J. E. Knowles, Philips tech. Rev. **24**, 242, 1962/63. See also A. Broese van Groenou *et al.* [9].

[21] G. H. Jonker and W. Noorlander, Science of Ceramics **1**, 255, 1962.

[22] G. H. Haertling and C. E. Land, J. Amer. Ceramic Soc. **54**, 1, 1971.

show no polarization at all. Since the width of the domains (the distance between the walls) is about 1 μm, the crystallites should be many microns in diameter. In this case a single crystal would be ideal, but ceramic material with a grain size of about 10 μm is found very satisfactory in practice.

If the environment of the crystallites allows no change in shape at all, then 90° rotations of polarization can no longer occur and only 180° rotations are possible. The polarizing of virgin ceramic material would then produce a remanent polarization equal to at most one-third of that of a single crystal. In barium titanate this is in fact the case. In the mixed crystals of lead titanate and lead zirconate, which are of much greater practical importance, this fraction may be much larger. The situation with these materials is more complicated. In some cases a rhombohedral deformation occurs, in which the crystal structure has eight easy directions of polarization. A range of composition has also been found in which tetragonal and rhombohedral forms exist side by side, and where one may change into the other [23].

A complication encountered in the preparation of ceramic lead titanate-zirconate is that lead oxide is highly volatile and easily escapes at the usual sintering temperature (about 1200 °C). One of the methods of minimizing this is hot pressing; with this technique a material is quickly formed in which any pores that may still be present are closed, preventing further escape of the lead oxide. However, at the temperature of about 1000 °C currently used for hot pressing — a temperature at which lead oxide is less volatile, which would be an added advantage here — the crystallites are too small (1-3 μm) for piezoelectric use (*fig. 13*). To obtain the desired grain size (> 5 μm) the hot pressing should be carried out at a higher temperature (1100 to 1200 °C) [24].

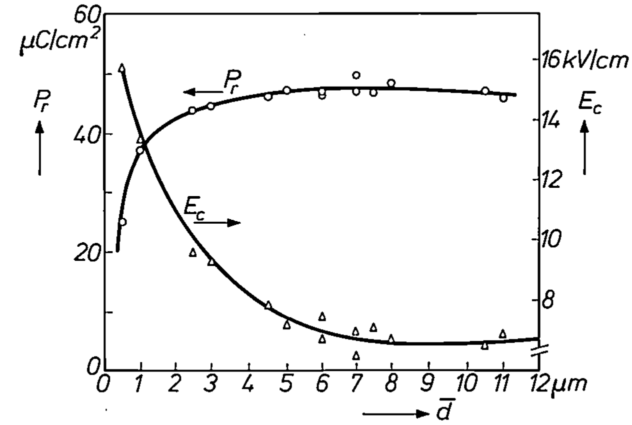


Fig. 13. Remanent electric polarization P_r and coercive electric field E_c as a function of average grain diameter $\bar{d}$ in hot-pressed lead zirconate-titanate. (After G. H. Haertling [24].) A large P_r and a small E_c are desirable for piezoelectric applications. P_r reaches its maximum at about 7 μm, near where E_c is at its minimum. At larger $\bar{d}$ values there is little further change in P_r and E_c. The grain diameter in this material depends mainly on the temperature during hot pressing; it should be 1100 to 1200 °C to obtain $\bar{d}$ = 7 μm.

Ceramic barium titanate as a dielectric

When a ferroelectric material is used as a dielectric in a capacitor the main role is played by the "normal" polarization process, in which the positive ions are collectively displaced slightly with respect to the negative ions. The contribution from wall displacements is slight. At small field amplitudes wall movements can still take place without losses (i.e. reversibly), but at higher amplitudes wall movements give rise to hysteresis losses, which increase steeply with the amplitude of the field. The rigidification mentioned earlier of the domain pattern as the material ages appears as a decrease in the dielectric constant and in the dielectric losses.

With barium titanate the two requirements which a dielectric should meet — high dielectric constant and low losses — are both satisfied by producing the material as a ceramic with small grains [21] [25]. This in fact suppresses the ferroelectric behaviour: as we have seen, there is no spontaneous polarization at all in small crystallites that are clamped tightly between their neighbours and too small to split up into domains.

In the first place this completely eliminates the principal losses found in coarse-grained material, the ferroelectric hysteresis losses. But small grains also favour a high dielectric constant. When a crystal changes from the dielectric to the ferroelectric state, there is a stiffening of the lattice in the direction of what becomes the polar axis: in other words the positive and negative ions become less easy to displace with respect to one another. Consequently the dielectric constant in coarse-grained material (with grains of 50 to 100 μm) is relatively low: when the material is cooled, the value decreases from about 10000 around the Curie temperature to 1400 at room temperature. In fine-grained material (with grains of about 1 μm) the decrease is much smaller and the final (room temperature) value is 3000 to 4000 (*fig. 14*).

Ceramic semiconductors

Some of the oldest semiconductors are various metal oxides, sulphides and iodides, which were formerly mainly used in rectifiers. In the transistor era of today they have been largely superseded as semiconductors by germanium and silicon; nevertheless, the oxides in particular have kept a place in their own right. In practical respects they differ mainly from germanium and silicon in that they function well in ceramic form, are relatively inexpensive to make and can be used at high temperatures. A familiar application is the NTC ther-

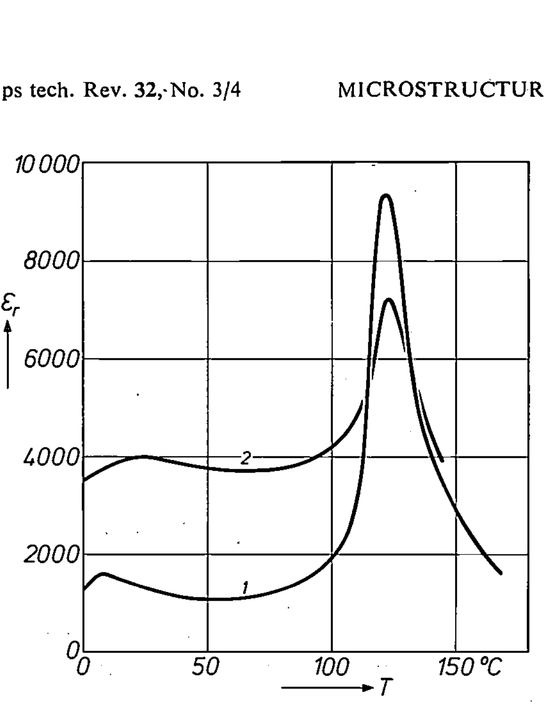


Fig. 14. The relative dielectric constant ε_r of ceramic barium titanate as a function of temperature. (After Jonker and Noorlander [21].) Curve *1* for grains larger than 50 μm, curve *2* for grains of about 1 μm. The fine-grained material has a considerably higher ε_r in the temperature range below T_C — owing to the suppression of the ferroelectric behaviour — and is therefore more suitable as a dielectric.

mistor, based on the negative temperature coefficient of the resistance, which is an inherent property of semiconductors.

The conduction in these oxides arises through a change of the valence of some of the metal ions. For instance, NiO becomes a *P*-type conductor when Li^{1+} ions are substituted for a number of Ni^{2+} ions. Electroneutrality is preserved because an equal number of Ni^{2+} ions becomes trivalent. Hole conduction arises because of the occurrence of Li^{1+}-Ni^{3+} pairs as acceptor centres, capable of receiving an electron from the valence band. Another *P*-type conductor is Li-substituted CoO. Familiar *N*-type conductors are Fe_2O_3 with Ti substitution and TiO_2 with Nb substitution, in which Ti^{4+}-Fe^{2+} and Nb^{5+}-Ti^{3+} are the respective donor centres.

It is interesting to pause here for a moment to consider why it is that these materials can be used as semiconductors in the ceramic, i.e. polycrystalline form, whereas germanium and silicon semiconductors always have to be single crystals. In germanium and silicon the charge carriers in the unperturbed crystal have a very high mobility (> 1000 cm^2/Vs), which corresponds to a mean free path of many thousands of interatomic distances. In many applications (e.g. the transistor) a mean free path of this magnitude is essential. Since charge carriers are scattered at grain boundaries, the mean free path in polycrystalline material is smaller, and therefore the polycrystalline form cannot be used. In the oxidic semiconductors, on the other hand, the mobility of charge carriers is very low even in the unperturbed crystal, corresponding to a mean free path of the order of one interatomic distance, and grain boundaries cannot reduce it very much further. The important thing here, therefore, is not a large mean free path but a particular conductivity (temperature-dependent in the NTC thermistor). In spite of the mobility being so very much smaller, this conductivity has a value that can compare with that of germanium and silicon, because of a much greater concentration of charge carriers.

Thus, grain boundaries are in themselves permissible. However, accurate control of their state is essential. In *N*-type semiconductors, for example, adsorption of oxygen at the grain boundaries can be harmful, since the adsorbed oxygen atoms form O^- or O^{2-} ions by taking up the required electrons form the material on both sides of the grain boundaries. This produces layers of high resistance between the grains, which can greatly reduce the total conductivity. This also makes the conductivity dependent on frequency and voltage. Conversely, in *P*-type conductors the removal of oxygen from the grain boundaries has a detrimental effect (*fig. 15*).

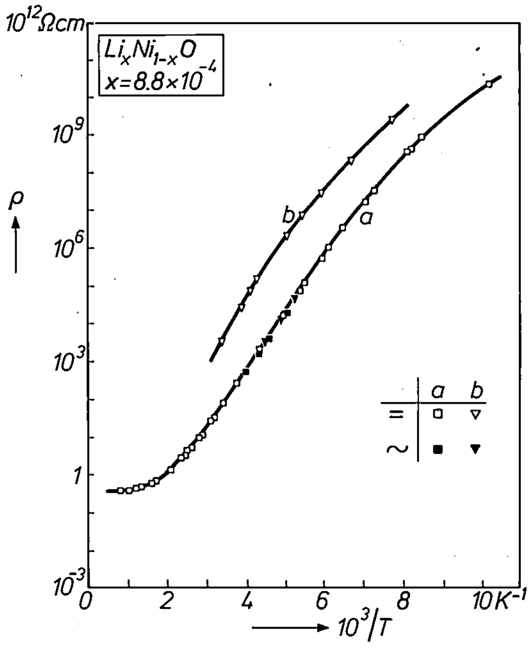


Fig. 15. Resistivity ρ of Li-doped NiO, a *P*-type semiconductor, as a function of the reciprocal temperature, for both d.c. (=) and a.c. (~). (After A. J. Bosman and C. Crevecoeur [26].) Sample *a* was cooled in air after sintering, *b* was cooled in nitrogen. In case *b* oxygen has been extracted from the grain boundaries, causing these to become current barriers. At higher frequencies the resistance of *b* coincides with that of *a*, which proves that the increased resistance is in fact due to the grain boundaries.

[23] K. Carl and K. H. Härdtl, Strukturelle und elektromechanische Eigenschaften La-dotierter $Pb(Ti_{1-x}Zr_x)O_3$-Keramiken, Ber. Dtsch. Keram. Ges. **47**, 687-691, 1970.

[24] G. H. Haertling, Amer. Ceramic Soc. Bull. **43**, 875, 1964.

[25] H. Kniepkamp and W. Heywang, Z. angew. Physik **6**, 385, 1954.
W. R. Buessem, L. E. Cross and A. K. Goswami, J. Amer. Ceramic Soc. **49**, 36, 1966.

[26] A. J. Bosman and C. Crevecoeur, Phys. Rev. **144**, 763, 1966.

It is therefore necessary to prevent oxygen from being taken up or released at the grain boundaries. This means that the microstructure should not have open pores, and also that the gas atmosphere should be accurately controlled both during sintering and cooling.

This reasoning would seem to imply that the crystallites should be large and that a single crystal would really be ideal. In practice, however, ceramic material has been found to be more reliable — even for research purposes — than the single crystals at present available, since these single crystals are often not as homogeneous chemically as has sometimes been assumed.

Ceramic barium titanate for PTC thermistors

In semiconducting barium titanate for PTC thermistors [27] the situation is the converse: here it is the grain-boundary effect that is used to good advantage.

The semiconductivity is obtained here by doping with an ion of higher valency: La^{3+} for Ba^{2+}; Sb^{5+} or Nb^{5+} for Ti^{4+}. In combination with a Ti^{3+} ion an ion of this type acts as a donor centre. On increasing the temperature, the resistance of a ceramic N-type semiconductor obtained in this way may increase by a factor of 10^3 to 10^5 in a range of some tens of degrees around the Curie point (*fig. 16*). The positive temperature coefficient may reach a value of 60% per degree.

This remarkable property may be explained from the fact that the dielectric constant ε varies considerably in the region of the Curie point, in combination with the assumption that oxygen is adsorbed at the grain boundaries [28]. As in the other oxidic N-type semiconductors, this oxygen takes up conduction electrons from the environment. A grain boundary thus becomes a negatively charged barrier layer with a positive space-charge layer on both sides, and hence a symmetrical potential barrier for the remaining conduction electrons. The electrons can only cross this barrier if they have sufficient thermal energy. The resistance is therefore proportional to exp $(e\phi/kT)$, where ϕ is the height of the barrier. Now ϕ is theoretically proportional to $1/\varepsilon$, and above the Curie point ε decreases sharply with increasing temperature, since according to the Curie-Weiss law ε is inversely proportional to $T - T_C$. This explains the observed increase of resistance.

For $T < T_C$ the value of ε is also much smaller than just above the Curie point, and it might therefore again be supposed that the grain boundaries would form high barriers, so that the resistance would be high. This is not the case; the barriers here are probably destroyed by the spontaneous polarization.

In a $BaTiO_3$ crystal doped for example with La the extra oxygen adsorbed at the end of the sintering process would indeed be expected to remain at the grain boundaries and not diffuse into the crystallites. This is because the substitution of the higher-valency La^{3+} ions for Ba^{2+} ions is compensated not only by conduction electrons but also by metal-ion vacancies. As a result there are no oxygen vacancies, and the diffusion of oxygen is not possible.

The optimum microstructure in this case will now be clear: there should be a large number of grain boundaries, i.e. a fine-grained material is required, and there should be effective distribution of oxygen over the grain boundaries, which is best achieved with a slightly porous material. Since there are no oxygen vacancies the sintering process in this material takes place relatively slowly. Because of this, both conditions are to some extent fulfilled automatically.

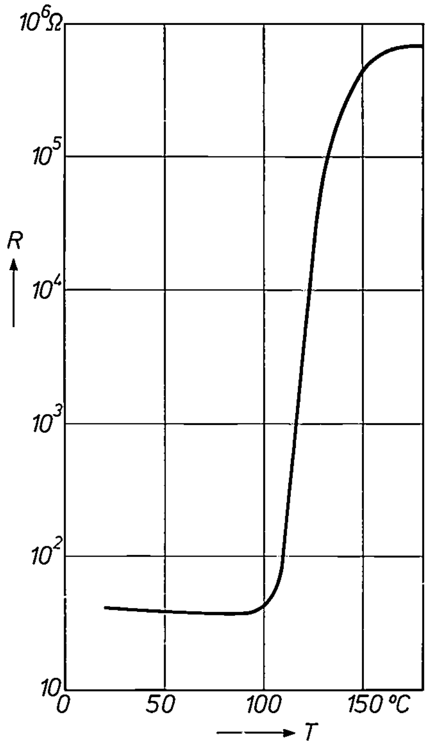


Fig. 16. The resistance of semiconducting barium-titanate as a function of temperature. (After E. Andrich [27].)

Diagram of relationships

That ends our account of the properties and applications that illustrate the importance of the microstructure. On pages 88 and 89 a diagram is presented which depicts the multiple relationships existing between method of preparation, starting material, microstructure and product. The diagram, which lays no claim to rigour or completeness, is best read from right to left. Most of the connecting lines can then be taken to mean "because". To take an example we consider the Ni-Zn ferrite "ferroxcube 4" in the "product" column on the right. This is characterized by low losses in the MHz range. These losses are low *because* there is no domain-wall relaxation and no resonance loss. There are no resonance losses *because* gyromagnetic resonance only occurs at higher frequencies *because* the material possesses magnetic anisotropy, in short *because* it is an Ni-Zn ferrite. Domain-wall relaxation is absent *because* there are no mobile domain walls *because* the crystallites are small *because* the material has been hot-pressed.

Other examples discussed in this article can also be identified in the diagram.

Summary. The scope for optimizing the microstructure of a ceramic material to give the product certain desired properties has greatly increased in recent years. This is because of advances in the understanding of sintering processes, such as the effect of certain slight deviations from the stoichiometric composition (defect chemistry), and in new methods of sintering (hot pressing) and of preparing the starting powder (co-precipitation, freeze-drying, spray-drying).

In ceramic ferromagnetics a structure with large, fault-free crystallites favours domain-wall mobility, which is desirable for high-permeability ferrites; domain-wall movements are not possible in a fine-grained structure, which is desirable for permanent magnets and high-frequency ferrites; a structure with small grains and low porosity is also required for microwave ferrites, as it suppresses the unwanted excitation of spin waves.

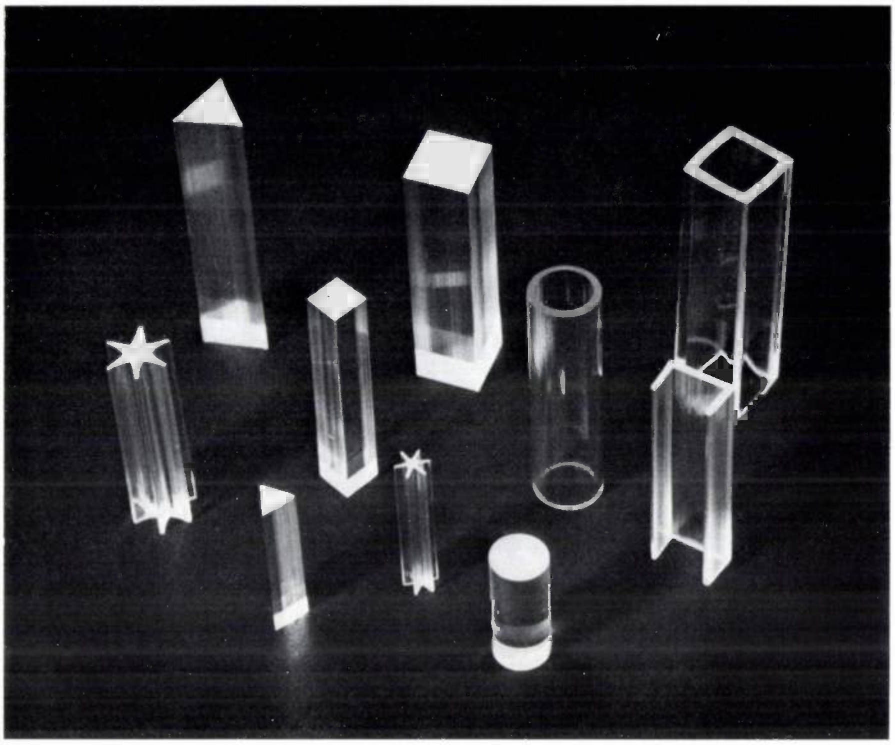
In ceramic ferroelectrics, at any rate in perovskites, the change in shape of the crystals on polarization is of overriding significance. Small crystallites wedged between their neighbours can neither split up into domains nor undergo the requisite change in shape, so that polarization is only possible in coarse-grained material. Large crystallites are therefore required if the ferroelectric effect is to be utilized (piezoelectric applications), but the grains should be small if this effect is to be suppressed (e.g. in barium titanate as a dielectric).

Grain boundaries can act as current barriers, which may be undesirable for some applications (e.g. in NTC thermistors of semiconducting oxides) but can also be desirable (in barium titanate used for PTC thermistors or in Mn-Zn ferrites for the kHz range).

Grain orientation (or "texture") is used in ferroxdure to obtain a higher $(BH)_{\max}$.

[27] For applications see, for example, E. Andrich, Philips tech. Rev. 30, 170, 1969.

[28] O. Saburi, J. Phys. Soc. Japan **14**, 1159, 1959.
W. Heywang, Solid-State Electronics **3**, 51, 1961 and Z. angew. Physik **16**, 1, 1963/64.
J. B. MacChesney and J. F. Potter, J. Amer. Ceramic Soc. **48**, 81, 1965.
G. H. Jonker, Solid-State Electronics **7**, 895, 1964 and Mat. Res. Bull. **2**, 401, 1967.
E. Andrich and K. H. Härdtl, Philips tech. Rev. **26**, 119, 1965.



Extrusion of glass

E. Roeder

In recent years many new types of glass have become available which do not always lend themselves readily to the conventional methods of shaping. Increasing interest has therefore been shown in new methods. The article below discusses the application of extrusion techniques to glass. Extrusion is shown to be especially suitable for glasses that have a strong tendency to crystallize.

In the metal and plastics industry extrusion is nowadays a widely used method of manufacturing rods and tubes of various profiles [1]. There are two main forms of this essentially simple method: direct and indirect (or inverse) extrusion. In *direct* extrusion a quantity of material contained in a cylinder is subjected to pressure by a plunger or punch, which forces it through the relatively small aperture of a die; see *fig. 1a*. The material is heated before extrusion to increase its plasticity. The profile of the bars extruded in this way is determined by the shape of the die aperture. Tubes instead of bars can be obtained by fitting a mandrel in the die aperture (fig. 1*b*). During this process the punch and the extruded rod or tube move in the same direction.

In the *indirect* extrusion process a *hollow* punch is used, to which the die is fixed. Pressure on the die causes it to move towards the material, so that the material is forced through the hollow punch (fig. 1*c*). The compressed material thus flows in the direction opposite to that of the punch. There are no frictional forces at the cylinder wall in this case, since there is no relative motion between the cylinder and the material. Although this method therefore requires less compressive force, the use of a hollow punch involves so many difficulties that indirect extrusion today is employed only in special cases.

Dr. Ing. E. Roeder is with Philips Forschungslaboratorium Aachen GmbH, Aachen, Germany.

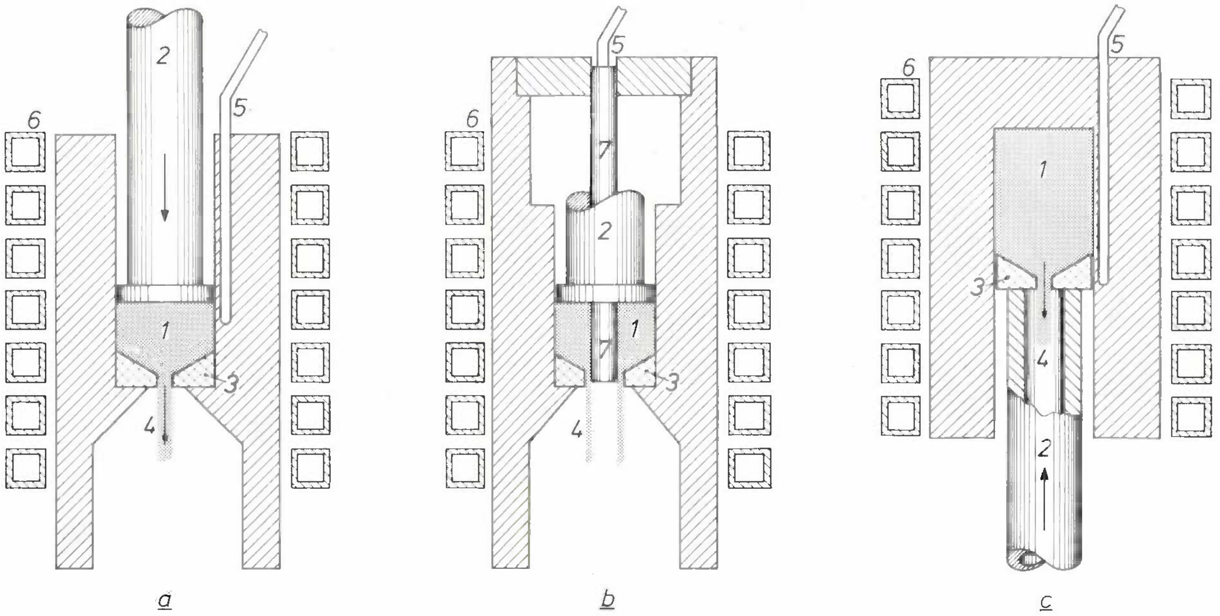


Fig. 1. The three types of extrusion equipment (schematic): *a*) for direct extrusion of rods (punch and glass rod move in the same direction), *b*) for direct extrusion of tubes and *c*) for indirect extrusion of rods (punch and rod move in opposite directions). *1* glass billet. *2* punch. *3* die. *4* extruded product. *5* thermocouple (only partly visible in *b*; the weld here is at the end of the hollow mandrel *7*). *6* high-frequency coil for induction heating.

In the glass industry extrusion is not yet a widely used method. This is presumably because the glasses used until the present in industry have mainly been types that can be shaped in an economic way by the conventional methods, such as moulding, blowing, drawing, pressing or rolling. Recently, however, there has been an increasing demand for new types of glass with special chemical and physical properties, and often of unusual compositions. Extrusion lends itself better to the shaping of such glasses than the conventional methods [2], and in some cases it is in fact the only possible method.

Just as with metals and plastics, extrusion of glass also provides a simple means of producing tubes and rods of widely different shapes and cross-sections. Some typical extruded glass products are shown in the title photograph.

Which types of glass are most suitable for extrusion?

Short glasses

In the working of glass, one of the most important properties is its viscosity. The viscosity range that can be covered is no less than 17 powers of ten. As an example *fig. 2* shows the viscosity-temperature curve of a common soda-lime-silica glass. The various shaping processes have to take place within specific viscosity limits, that determine the "working range". The figure shows the working ranges for mechanical blowing (*b*), pressing (*p*) and drawing (*d*).

If the viscosity varies strongly with temperature the temperature range within which the material can be worked is very narrow. Glasses of this type are called short glasses. The fact that large temperature changes are not permissible in the working of such glasses forms a problem in hand-drawing, for example, owing to the

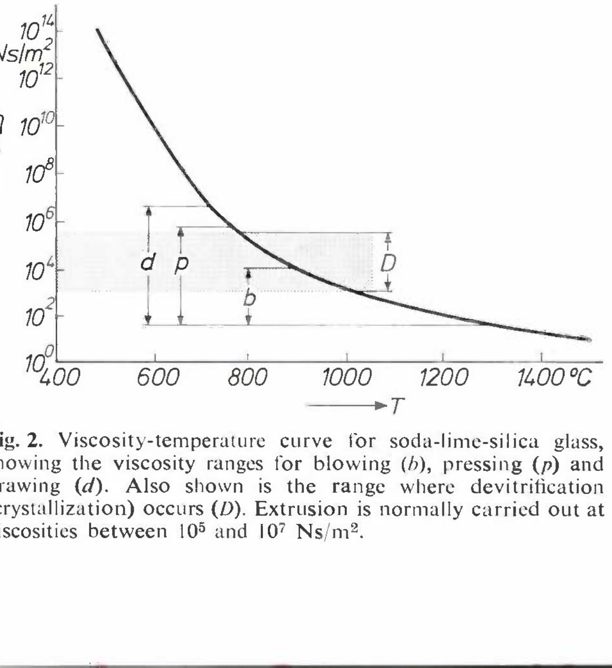


Fig. 2. Viscosity-temperature curve for soda-lime-silica glass, showing the viscosity ranges for blowing (*b*), pressing (*p*) and drawing (*d*). Also shown is the range where devitrification (crystallization) occurs (*D*). Extrusion is normally carried out at viscosities between 10^5 and 10^7 Ns/m^2.

[1] See for example C. E. Pearson and R. N. Parkins, The extrusion of metals, Chapman and Hall, London 1960; R. Chadwick, Metallurg. Rev. 4, 189, 1959; E. C. Bernhardt, Processing of thermoplastic materials, Reinhold, New York 1959.

[2] B. Frank, E. Roeder and S. Scholz, Ber. Dtsch. Keram. Ges. 45, 231, 1968; E. Roeder, J. non-cryst. Solids 5, 377, 1971.

considerable cooling that occurs during this process. In the extrusion process, on the other hand, the temperature of the zone in which the actual shaping takes place remains fairly constant. This makes extrusion highly suitable for working short glasses.

Glasses that crystallize easily

In the extrusion process the glass is enclosed on practically all sides by the solid walls of the cylinder, the punch and the die. Because of this, greater deformation forces can be exerted on the glass than would be possible in the case of unidirectional loading, as for example in the drawing process. This means that in the extrusion method the glass can be shaped at a higher viscosity, and hence at a lower temperature, than in a conventional method. Applying pressures of the order of 10^8 N/m^2 (a few thousand atmospheres) it is possible to go up to a viscosity of 10^7 Ns/m^2 (10^8 poise); see fig. 2. This is a great advantage with glasses that readily devitrify, i.e. crystallize. As can be seen in fig. 2, the viscosity range D in which the crystallization tendency is very marked lies in the middle of the working range of the conventional methods. The application of such methods to glasses of this type is therefore limited. Extrusion makes it possible to stay "above" the dangerous region.

The viscosity range is somewhat narrower than for the conventional glasses (about 10^6 to 10^7 Ns/m^2). The upper limit of the working range is, of course, the same, but the lower limit — i.e. the upper limit of the crystallization range, which is determined by the ion mobility — is higher in the case of unstable glasses. (In the case of conventional glasses this limit lies at only about 10^5 Ns/m^2.)

The extrusion of unstable glasses requires special measures for filling the cylinder; the cylinder cannot be filled, in the usual way, with a billet pre-cast in a mould, since the glass will already have crystallized to a considerable extent. It is therefore necessary to make "frit", which is done by quenching small quantities of glass rapidly to room temperature. The rapid cooling suppresses crystallization. This can also be done by heating the starting material in a high-temperature source (e.g. a plasma torch), resulting in molten droplets which are quenched as they fall. In some cases the glass area of a system can even be enlarged in this manner [3].

Fig. 3 illustrates the microstructure of a glass after extrusion starting from frit. This structure is highly temperature-dependent. In general, the higher the extrusion temperature, the greater will be the fusion of the original glass grains. In fig. 3*b* the "grain boundaries" can still clearly be distinguished.

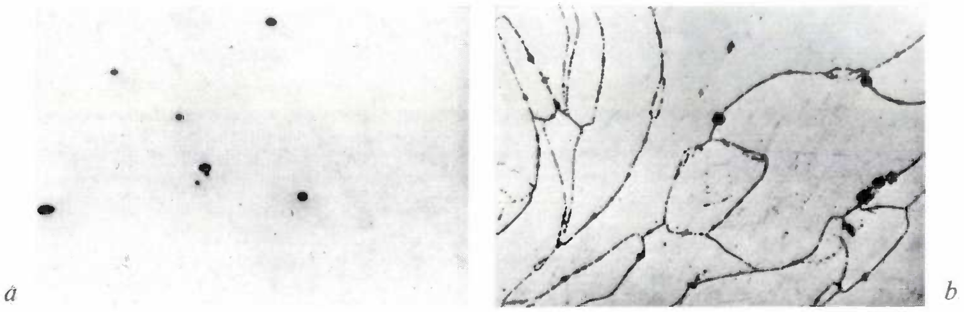


Fig. 3. Polished sections of products obtained by extruding "frit". *a*) Boron-silicate glass, extruded at a temperature of 860 °C, a pressure of 10^7 N/m^2 and a viscosity of $10^{4.7}$ Ns/m^2. *b*) Calcium-aluminate glass, extruded at a temperature of 880 °C, a pressure of 3×10^7 N/m^2 and a viscosity of $10^{6.7}$ Ns/m^2. In the boron-silicate glass there is almost complete fusion of the original grains; in the calcium-aluminate glass the grain boundaries are still clearly visible in the product. A distance of 1 cm in the photograph is approximately 100 microns. The etchant used was concentrated HNO_3.

High-melting glasses

We have already indicated that working at a higher viscosity than in the conventional methods is equivalent to working at a lower temperature. Depending on the steepness of the viscosity-temperature curve a working temperature 70 to 250 °C lower may be chosen (*fig. 4*). In terms of energy saving, this is an advantage for all types of glass, but it is particularly advantageous for glasses with a high softening point. The energy saving is then considerable, and moreover it is not so difficult to control the temperature. For example, quartz glass can be extruded at a temperature of only 1750 °C, whereas a temperature of about 2000 °C is required for melting and drawing quartz glass.

Working at a lower temperature, and hence at a higher viscosity, reduces the effect of surface tension, enabling products to be obtained that have sharp edges; see the title photograph.

[3] B. Frank and J. Liebertz, Glastechn. Ber. 41, 253, 1968.

The equipment

The equipment we have developed for glass extrusion does not differ in essentials from that used for extruding metals or plastics. The equipment is made of a heat-resistant Cr-Ni steel; for higher temperatures and pressures a nickel-base alloy is used, which is capable of withstanding a pressure of 10^8 N/m^2 at temperatures up to about 950 °C. For extrusion at a temperature higher than this the metal tools must be replaced by graphite, and the pressure must not exceed 3×10^7 N/m^2.

The pressure is applied to the punch hydraulically and is kept constant during extrusion by means of a regulating valve. The cylinder may be heated either by high-frequency induction, with a heating element or by means of a gas burner. The extrusion temperature is monitored and regulated by a thermocouple near the die aperture. An important design consideration is that the temperature distribution inside cylinder and die must possess rotational symmetry, otherwise the extruded products will leave the die aperture on the skew.

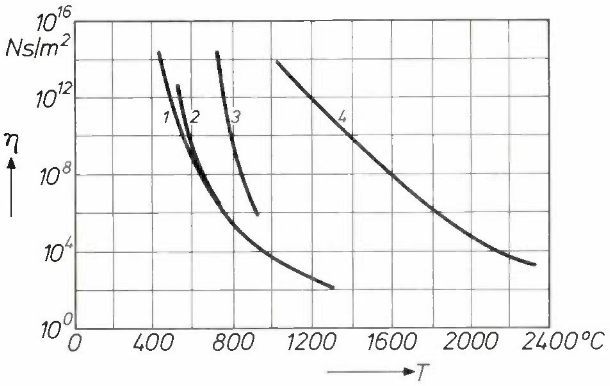


Fig. 4. Viscosity-temperature curves for various extruded glasses. Curve *1* a boron-silicate glass (Philips No. 28). Curve *2* an alkali lime-silicate glass. Curve *3* a calcium-aluminate glass. Curve *4* quartz glass. The maximum viscosity at which the glass can be shaped is higher in the extrusion technique than in the conventional methods, which means that a lower working temperature can be used.

The flow behaviour of the glass

An understanding of the flow behaviour of glass at the extrusion temperature can be obtained by studying the flow pattern of the glass. *Fig. 5* illustrates the simple manner in which such flow patterns can be obtained. The cylinder is filled with discs of glass between which a pigment, e.g. green Cr_2O_3 powder, is sprinkled. After some extrusion a rod is obtained as illustrated in the figure. It can clearly be seen that the material flows much faster in the inner zone than at the edge. The flow pattern is approximately parabolic in form. The question therefore arises whether the velocity distribution is in accordance with the Poiseuille-Hagen theory of the laminar flow of viscous liquids through long tubes:

$$v(x) = \frac{\Delta p}{4\eta l}\,(r^2 - x^2). \tag{1}$$

Here $v(x)$ is the velocity at a distance x from the axis

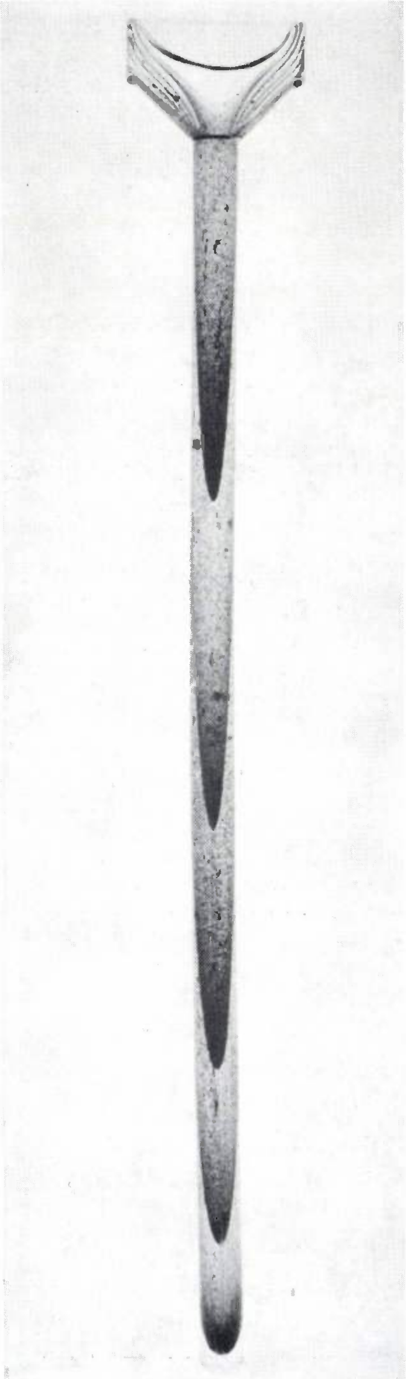


Fig. 5. Flow pattern during the extrusion of an alkali-lime-silicate glass through a graphite die with an aperture angle of 90°. Packed in the cylinder are seven transparent glass discs with a pigment, Cr_2O_3 powder, sprinkled between them. The displacement of the interfaces during the extrusion can clearly be followed in this way. Extrusion temperature 750 °C; viscosity $10^{6.2}$ Ns/m^2; pressure 10^7 N/m^2; diameter of die aperture 5 mm.

of the tube (the die channel), Δp is the pressure difference between the two ends of the tube, l is the length and r the radius of the tube and η the dynamic viscosity. At the wall ($x = r$) this law states that $v(x)$ is zero.

Whether the flow due to the extrusion of glass does in fact obey this theory can be ascertained by plotting measured velocity values as a function of the pressure difference Δp (*fig. 6*). A straight line is then obtained in accordance with the well-known Poiseuille-Hagen equation for the volume flow V, given after integration of equation (1) by:

$$V = \frac{\pi \Delta p r^4}{8 \eta l}. \tag{2}$$

The theoretical curve (chain-dotted line) agrees well with the experimental curve *1*, obtained using a die made of a nickel-base alloy.

If however a die of graphite or boron nitride is used (curves *2* and *3*) the curves, though still straight, are very much steeper. This may be explained by assuming "slip" of the glass along the wall of these self-lubricating materials, resulting in a higher velocity.

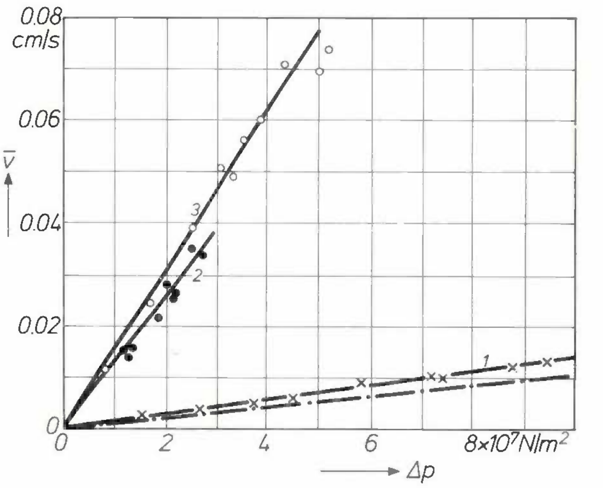


Fig. 6. The extrusion rate $\bar{v}$ (the volume flow V divided by πr^2) of an alkali lime-silicate glass plotted as a function of the pressure difference Δp between the ends of the die channel. Low extrusion rates were used to keep the experimental conditions constant. Extrusion temperature 680 °C; viscosity $10^{7.67}$ Ns/m^2; diameter and length of the die channel 4 and 10 mm. The chain-dotted line is the theoretical curve corresponding to the Poiseuille-Hagen theory. *1*, *2* and *3* are experimental curves obtained using dies made of three different materials: a nickel-base alloy, boron nitride and graphite. The steeper slope of curves *2* and *3* is connected with the occurrence of slip along the wall.

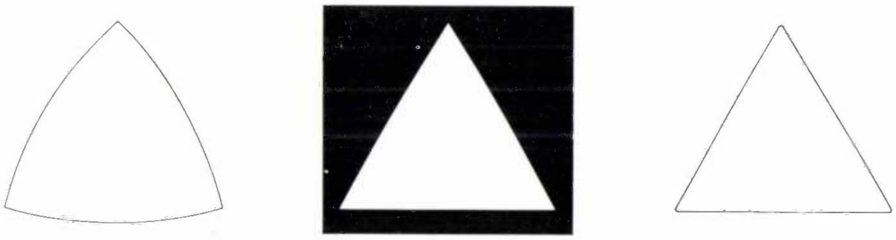


Fig. 7. Cross-section of extruded alkali lime-silicate glass, using a Cr-Ni steel die (on the left) and a graphite die (on the right), compared with the cross-section of the die aperture. It can be seen that the extruded glass "swells" in the first case but not in the second case. Extrusion temperature 750 °C; viscosity $10^{7.4}$ Ns/m^2; pressure 10^7 N/m^2; side of triangle 10 mm.

The speed of the extrusion process may be as high as a few tens of centimetres per minute for glasses that do not crystallize easily (in which case a high temperature can be chosen). For glasses that do crystallize easily, however, lower temperatures and therefore lower speeds have to be accepted.

Accuracy of shape

In the extrusion of glass through metal dies of circular cross-section it is almost always found that the diameter of the product is about 10% greater than that of the die aperture. This is not the case, however, in extrusion through a graphite die.

With round profiles these changes of diameter are easy to correct by changing the die aperture. Non-circular cross-sections, however, give rise to more complicated deviations, which are not so easy to correct. For example, a triangular aperture in a die of Cr-Ni steel results in a rod with a cross-section as illustrated in *fig. 7* (on the left), showing a "swelling" of the side walls. This swelling effect is not found with the same die made of graphite (fig. 7, on the right).

This effect may possibly be attributable to the cooling of the emerging mass of glass, giving rise to a colder, more viscous outside layer which is consequently "blown up" [4] as a result of the central part of the mass flowing at a higher speed. When a graphite die is used, the difference in speed between the outside layer and the bulk of the material is so much less, owing to slip at the walls, that the effect is negligible.

[4] Visco-elastic effects may also be involved here; see for example A. C. Merrington, Nature **152**, 663, 1943, and J. Serrin in Encyclopedia of Physics VIII/1, Springer, Berlin 1959, in particular p. 242.

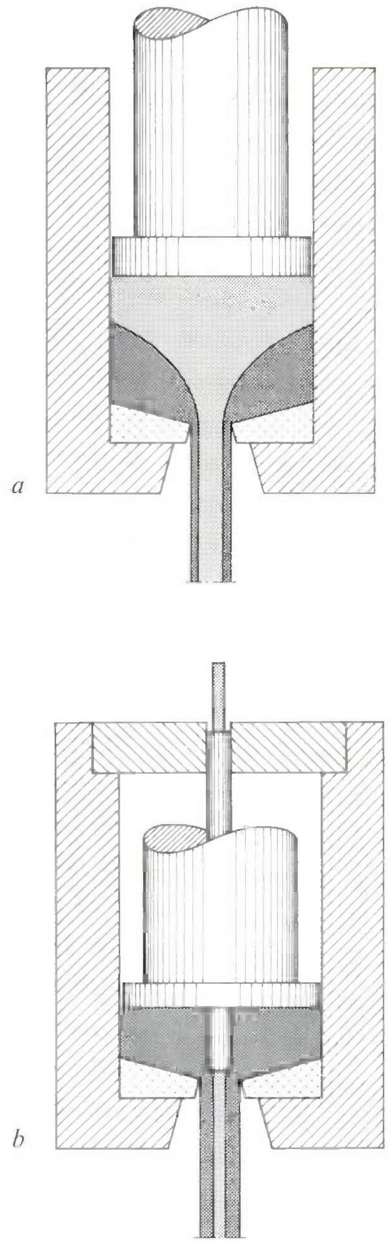


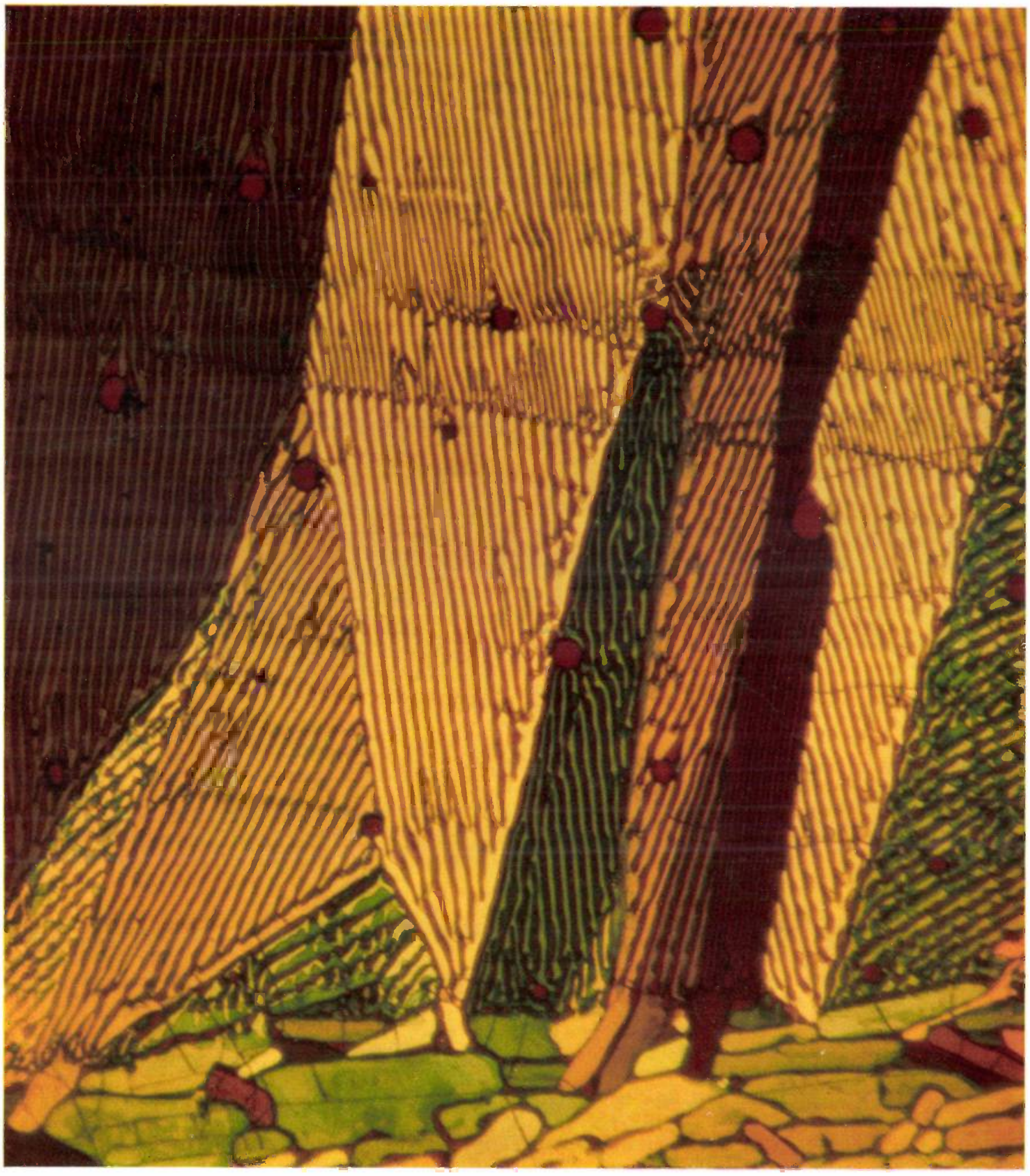
Fig. 8. Principle of two methods of extrusion cladding. In (*a*) two discs of different glass are used (light and dark grey). The lower disc provides the cladding, the upper disc the core. In (*b*) the core material is introduced beforehand into a hollow mandrel.

Extrusion cladding

To conclude this article we shall briefly discuss an interesting variant of the glass extrusion process, made possible by the characteristic flow pattern in extrusion. This variant, which we have called extrusion cladding, is a method of extruding bars or tubes with a sleeve of a different type of glass. In its simplest form the method consists in extruding two discs of dissimilar glass placed one on top of the other (*fig. 8a*). There are two conditions to be met by this method of extrusion cladding: the viscosities of the two types of glass must not differ too widely, and the coefficient of expansion of the cladding glass must not be greater than that of the core glass. If the cladding glass shrinks more than the core glass during the cooling, then tensile stresses are created and cracks appear in the surface.

It is easy to see that the thickness of the cladding is not constant over the whole length of the product extruded in this way. A product with a cladding of constant thickness can be obtained using the method illustrated in fig. 8*b*. A rod of core glass is introduced beforehand into a hollow mandrel. The cladding glass is then applied around it and heated to its appropriate extrusion temperature. In this method the viscosity of the core glass must be much higher than that of the cladding glass. In a similar way a metal wire can be coated with glass. The relative coefficients of expansion must then meet the same requirements as for glass on glass.

Summary. The extrusion technique has much to offer for the shaping of "short" glasses, glasses that easily crystallize (devitrify) and high-melting glasses. Rods and tubes of widely different profile can be produced with this technique. Extrusion rates can be of the order of a few tens of cm/min. A good picture of the flow pattern during extrusion can be obtained by filling the extrusion cylinder with glass discs and applying a pigment between them. If the extrusion die is made of Cr-Ni steel the flow behaviour follows the Poiseuille-Hagen theory. If it is made of graphite, the glass exhibits "slip" along the walls. In the first case the extruded product "swells" upon emerging from the die aperture, in the second case it does not. The article concludes with a description of a variant of the technique, called extrusion cladding, which produces a glass rod or tube coated with glass of a different type. The method can also be used for coating a metal wire with glass.



Growth of "composites" by unidirectional solidification

Composites are anisotropic materials in which two (or possibly more) different phases are arranged in a periodic pattern, for example in the form of lamellae or needles. They can be obtained by unidirectional solidification of a melt of eutectic composition. This requires a flat solid/liquid interface moving at a constant rate (e.g. 10 cm/h). The segregating phases then arrange themselves into a periodic structure (order of magnitude of the period 1 micron). Certain relationships characteristic of the eutectic are found between the growth direction, the crystallographic directions of the phases and the geometry of the crystals. There is usually a small angle between the lamellae or needles and the growth direction.

In investigations of these solidification processes at Philips Research Laboratories the solid/liquid interface where the crystal-growth process takes place is directly observed with a microscope. A thin layer (about 10 μm) of a molten eutectic mixture between fused-silica plates is moved at a suitable rate under the objective in a stationary temperature field; the solid/liquid interface is then also stationary with respect to the objective. By using eutectics with at least one birefringent phase, differences in crystal orientation can be made visible as differences in colour by using polarizers.

The photograph (magnification 1200×) shows a thin layer of this type after solidification. It consists of a eutectic in which one phase is birefringent and the other (the phase that is mainly brown) is not. The unidirectional solidification started at the lower polycrystalline layer. It can be seen that with the upward movement of the solid/liquid interface from seed crystals in the original structure various regions have grown, each of which has its own typical lamellar orientation. During the solidification process some regions are gradually ousted by others (competitive growth), so that finally only a few remain. Traces of growth rate fluctuations, splitting up the lamellae, can be seen just above the centre of the photograph.

“Controlled cascading”, a new open-loop control principle for adjustable frequency dividers

G. Kaps

A technique widely favoured by control engineers is that of frequency dividing. Now that modern digital-circuit blocks are available, frequency dividing can be more extensively and more simply applied, in applications ranging from speed control of simple electric motors to flow measurements in complex chemical processes.

A detailed analysis of conventional multistage divider circuits has led the author to a new kind of internal control principle. This “controlled cascading” suppresses most of the pulse bunching that often degrades measurement data processed by conventional circuits.

Frequency dividing: why and how?

In many measuring processes and control-technique problems quantities are represented by periodic pulse signals. Since the frequency of such a signal is a measure of the original quantity, any interference has less effect than it would with analog signals. Moreover, modern digital equipment can measure frequencies very accurately and will also permit simple frequency-data processing. In this kind of work frequency-dividing circuits are often used. An adjustable frequency divider can for example convert a signal at a fixed reference frequency into a number of signals at lower frequencies. This will give a variable frequency standard, which is often necessary in frequency control. Frequency dividers can also be used to change frequencies that represent measured quantities in such a way that the new numerical values are equal to the quantities expressed in the correct units. This kind of frequency conversion is referred to as “scaling”. Scaling is frequently used in chemical-process instrumentation; for example, when measuring a volume flow rate, the measuring circuit can be used to divide the frequency of the pulse signal by the calibration factor, i.e. the integer (> 1) that corresponds to 1 litre per minute. The output frequency is then equal to the number of litres per minute.

There are two kinds of dividing circuit. A “true” dividing circuit counts the pulses of the input signal up to a predetermined number, then gives a single output pulse, resets itself to zero and starts counting again. The output frequency is then equal to the input frequency divided by the predetermined number.

The second kind, the “digital rate multiplier”, produces an output signal by transmitting only M pulses from each train of say 10^3 input pulses, all the others being suppressed. The frequency of the output signal is then equal to that of the input signal multiplied by $M \times 10^{-3}$, where M is any integer from 1 to 999. These circuits multiply frequencies by numbers smaller than 1; hence the name. In practice digital rate multipliers meet the needs of most users, largely because of their wide range: for example it is easy to produce any decimal fraction to three places of decimals by cascading three adjustable frequency dividers.

A difficulty with the “true” dividing circuits is that they will only give integral calibration factors. Moreover, large calibration factors seriously degrade the discrimination of the measurements. Thus, if a detector measuring a volume flow gives 2364 pulses for each 1000 litres and the calibration factor 2364 is used, then the minimum detectable variation of volume flow is no less than 1000 litres.

On the other hand, if a digital rate multiplier were used in the example just mentioned, it would improve the discrimination by a factor of 10^3. With this kind of scaler arranged to give 423 output pulses for every 1000 at the input (the fractions 1000/2364 and 423/1000 are almost exactly equal) there would be 1000 output pulses per 1000 litres. Changes of one litre become discernible, provided that the output pulses appear at equal intervals, i.e. without bunching. Obtaining output signals without this loss of periodicity is the par-

G. Kaps, Dipl. Ing., now with the Philips Medical Systems Division, Eindhoven, was formerly with Philips Forschungslaboratorium Hamburg GmbH, Hamburg, Germany.

ticular problem that is alleviated by "controlled cascading".

In existing circuits, measuring accuracy is affected by the output-pulse bunching — present even when the input signal is purely periodic; Table I shows a striking example of this. In frequency-control devices transient response and even the stability can be seriously affected by bunching.

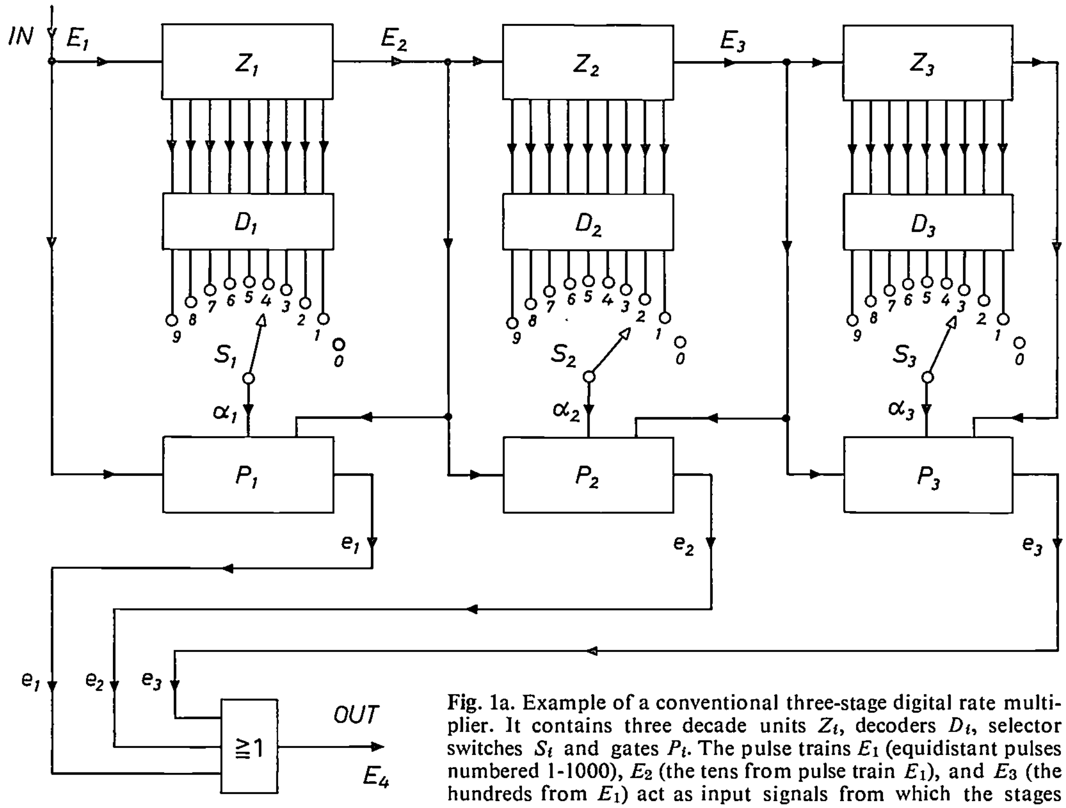


Fig. 1a. Example of a conventional three-stage digital rate multiplier. It contains three decade units Z_i, decoders D_i, selector switches S_i and gates P_i. The pulse trains E_1 (equidistant pulses numbered 1-1000), E_2 (the tens from pulse train E_1), and E_3 (the hundreds from E_1) act as input signals from which the stages generate three partial pulse trains e_1, e_2, e_3. The OR circuit combines the partial pulse trains to produce the output signal E_4, which consists of 423 pulses from each 1000 input pulses E_1. $\alpha_{1,2,3}$ control signals for the gates P_i.

To begin with we shall carefully analyse the operation of a conventional adjustable three-stage decimal divider circuit. From this we shall derive two clearly distinguishable kinds of bunching. The first, and rather trivial, kind is counteracted in existing circuits by using pulse-selection gates. We shall then show how the new control principle eliminates the second kind of bunching [1].

To simplify multistage arrangements we have developed a special frequency-dividing decade unit [2]. This circuit, which is only briefly discussed here, has been used for a four-stage digital rate multiplier we have built, incorporating controlled cascading. Frequencies up to 20 kHz can be accurately multiplied by the factor $M \times 10^{-4}$, where M is any integer from 1 to 9999.

The new control principle is a kind of "hierarchy": the first divider stage produces the complete output signal, controlled by the succeeding stages. In conventional circuits, on the other hand, the different stages in principle each perform the same function.

Conventional digital rate multipliers

The following example is intended to clarify the principles of existing digital rate multipliers.

The operation of a three-stage multiplier can be seen from *figs. 1a* and *b*, the block diagram and the corresponding pulse diagram. Each stage is made up from a decade unit Z_i, a switch S_i, a decoder D_i, and a gate P_i. We assume that the input signal E_1 applied to the first stage is purely periodic. The switches determine the

[1] For a more detailed treatment, see G. Kaps, Scaling of frequency analogous measured values, Proc. IFAC Symp. on Pulse-Rate and Pulse-Number Signals, Budapest, April 1968, p. 148.

[2] G. Kaps and H. P. Kunert, Dekadisch aufgebauter Frequenzteiler, German patent application 1285538 (23 May 1967).

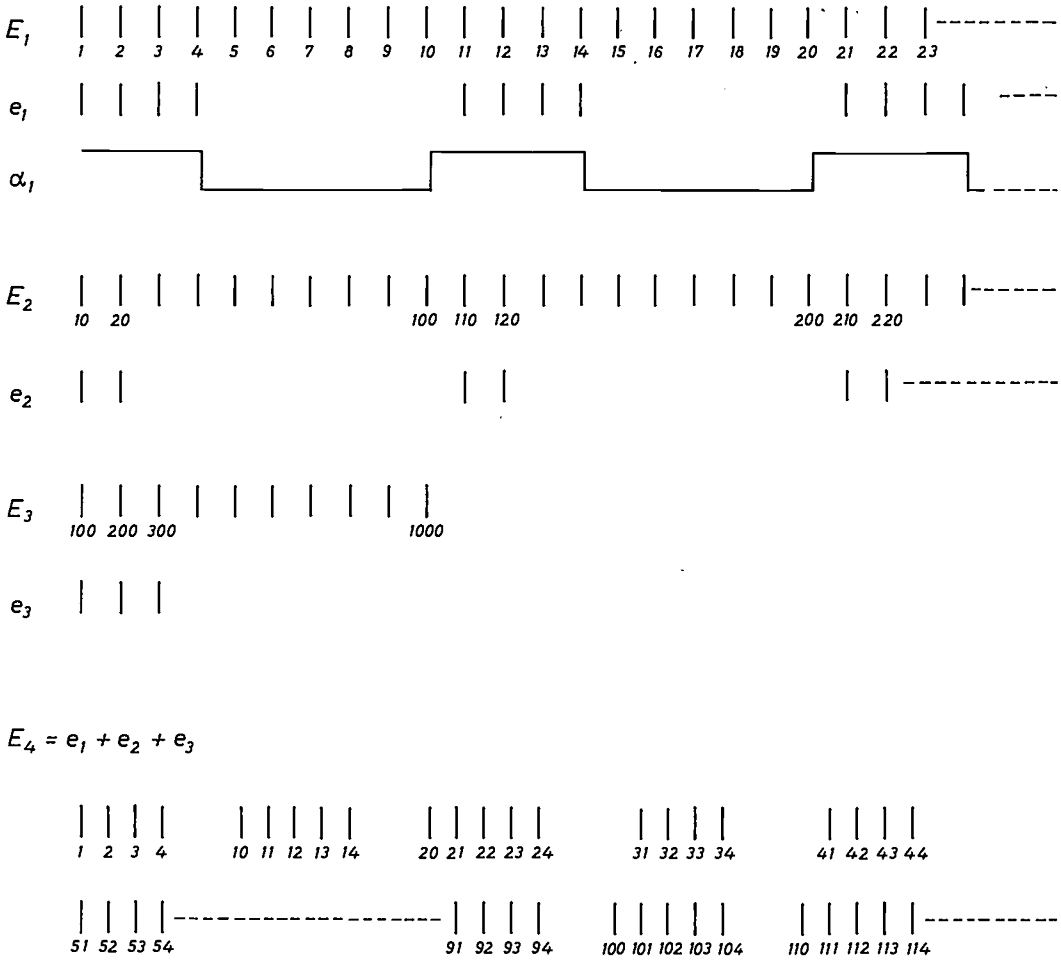


Fig. 1b. Pulse diagram for fig. 1*a* for the multiplication of 1000 Hz by the factor 0.423. Pulses, shown as vertical dashes, are labelled by the number indicating their position in the input pulse train E_1. The numbers of the pulses in the output signal E_4 illustrate pulse bunching. The significance of the other symbols is the same as in fig. 1*a*.

factor by which the frequency of E_1 is multiplied; in the case illustrated the factor is equal to 0.423. By means of the partial pulse trains e_i, 423 pulses out of each group of 1000 input pulses reach the output terminal, via the OR circuit and without noticeable delay. These pulses form the output signal E_4 with the decreased frequency required. In fig. 1*b* pulses are indicated by numbered vertical dashes. The pulse numbered 1 arrives at the input terminal first. The numbers of the pulses present in the output signal clearly demonstrate the occurrence of bunching.

Table I contains data from another case, in which a volume flow had to be measured in litres. Suppose that the detector gives 91 pulses per litre. The rate multiplier then has to multiply the frequency by 0.011 (switches S_1, S_2, S_3 in fig. 1*a* set at 0, 1, 1), since the fraction 11/1000 is for practical purposes equal to 1/91. The input E_1 (see fig. 1*b*) acts as a timing signal; instants

Table I. Example of a flow measurement in which pulse bunching causes a serious discrepancy between the actual instants in time when listed volumes pass and the recorded times (see the fourth column). The detector used here gives 91 pulses per litre. A scale transformation with a calibration factor of 91 is obtained by means of a digital rate multiplier of the type shown in fig. 1*a* and set to 0.011. The times are labelled by the numbers of the corresponding pulses. The output pulse numbered 100 originates in the third stage, all the other pulses in the second column come from the second stage (the pulse train e_2 in fig. 1*a*).

Flow volume in litres	Recorded elapsed time	True elapsed time	Ratio
1	10	91	9.10
2	100	182	1.82
3	110	273	2.58
4	210	364	1.73
5	310	455	1.47
6	410	546	1.33
7	510	637	1.25
8	610	728	1.19
9	710	819	1.15

in time are therefore labelled by the number of the corresponding pulse. The effect of the large measuring error caused by bunching appears in the discrepancy between the recorded elapsed times in the second column and the actual elapsed times in the third column.

In this circuit, bunching arises both when a partial pulse train is being formed, and during the superposition of these pulse trains to obtain the output signal.

Fig. 1*b* shows that from each group of ten pulses admitted to the first stage, only the first four pass the gate, forming the corresponding partial pulse train. The signal α_1 controls the gate. The input pulse numbered 10 reopens the closed gate and functions as the carry pulse for the following stage. The partial pulse train produced by the second stage comprises the first two pulses from each group of ten input pulses, similarly in the last stage the pulse train consists of the first three pulses. Clearly, such poor distributions imply bunching in the output signal. There is one exception: if only one pulse in ten passes a gate, then the train is of course periodic.

Here, perhaps, we should explain how the most uniform distribution is found for k pulses over a period of ten pulses, where k is the integer 1, 2, . . . , 9. For this purpose the relationship $m \geqslant 10\,p/k$ is used; m represents the number of the pulse to be selected and p is equal to 1, 2, . . . , k.

Table II lists the numbers m found from this relation for all values of k; purely periodic distributions arise only for k = 1, 2, 5.

Table II. The crosses denote the most uniform distribution of $k(= 1, 2, \ldots,$ or 9) pulses over an interval of ten periods; m is the number of the pulse to be selected.

m k	1	2	3	4	5	6	7	8	9	10
1										×
2					×					×
3				×			×			×
4			×		×			×		×
5		×		×		×		×		×
6		×		×	×		×		×	×
7		×	×		×	×		×	×	×
8		×	×	×	×		×	×	×	×
9		×	×	×	×	×	×	×	×	×

The pulse distribution within the partial pulse trains can be improved by placing pulse-selection gates between the decoders and the switches (fig. 1*a*). The sequences of voltage levels occurring at the terminals of a decade unit are simultaneously converted into nine pulse trains improved in the way shown in Table II. The conversion is effected by selection gate and decoder together. Each selection gate has nine output terminals carrying the improved pulse trains of 1, 2, 3, . . . , or 9 pulses. The correspondingly numbered contacts of the switch are connected to these output terminals.

The second kind of bunching is due to the superposition of the partial pulse trains (e_i in fig. 1). Bunching is usually made worse whenever the OR circuit adds a pulse from the second or third stage to the train formed in the first stage. In fig. 1*b* this happens when the pulses numbered 10, 20, 110, 120, 210, 220, etc. are added. Again, in Table I the pulse numbered 100 forming the partial pulse train of the third stage degrades the distribution in the output signal.

In the next section we shall discuss the elimination of this second kind of bunching by controlled cascading.

Controlled cascading

The previous section showed that superposing partial pulse trains leads to bunching. This is because the pulses from the second- and third-stage partial pulse trains are added to the first-stage pulse train at times that are fixed and not uniformly distributed. Adding such a pulse, e.g. from the second stage, changes the output signal of the circuit in a "discontinuous" way: the output suddenly contains one pulse more than the number selected by the first-stage switch. The distribution of these $s_1 + 1$ pulses is poor, even if the distribution of the s_1 pulses over a period of ten first-stage input pulses is correct. In short, the solution offered by controlled cascading is to raise the number selected by the first-stage switch temporarily by one. Controlled cascading restricts the production of the *complete* output signal to the first stage of the circuit. The addition of partial pulse trains and the associated bunching are eliminated. The pulses in the partial pulse trains from the following stages function as control signals, modifying the operation of the first stage. To put it more exactly: the appearance of a pulse in the second-stage partial pulse train starts a period of ten first-stage input pulses during which the first stage produces a modified partial pulse train. The modified pulse train contains one pulse more than the number selected by means of the first-stage switch. As soon as the control signal from the second stage returns to zero the first stage resumes normal operation. The first stage of the circuit thus transmits from each group of ten input pulses a partial pulse train consisting of either the number of pulses selected by the switch or of that number plus one. In similar fashion the second stage is controlled by the third.

For further clarification let us apply the new principle to the example of fig. 1*b*. The switches S_1, S_2 and

S_3 in fig. 1*a* are set at 4, 2 and 3 respectively. The input signal E_1 consists of 1000 pulses, ten of them also functioning as input to the third stage. The third-stage partial pulse train consists of three pulses, which control the second stage in such a way that the second-stage switch is effectively moved three times from the original setting 2 to 2 + 1 and remains 10 − 3 times at the setting 2. The input signal to the second stage consists of 100 pulses; control from the third stage results in a second-stage partial pulse train of 23 pulses (i.e. 3×(2 + 1) + (10 − 3)×2, which is of course equal to 10×2 + 3). Without controlled cascading, however, the second-stage partial pulse train would consist of 20 (i.e. 10×2) pulses. In turn the 23 pulses from the second stage control the first, making it transmit five instead of four pulses from each group of ten. From the remaining 77 groups of 10 — the total input E_1 consists of 100 groups of 10 — the number selected (four) is transmitted to the partial pulse train. The latter therefore consists of 423 pulses per 1000 input pulses; thus the first-stage partial pulse train can act as output from the complete circuit.

The arithmetic making controlled cascading feasible in this example includes the equalities

$$423 = 23\times(4 + 1) + (100 - 23)\times 4$$

and

$$23 = 3\times(2 + 1) + (10 - 3)\times 2.$$

The pulse-selection gate in the first stage ensures the optimum distribution of the five or four transmitted pulses over the corresponding period of ten input pulses. The second-stage pulse-selection gate determines the distribution in time of the four- and five-pulse groups forming the first-stage partial pulse train. The distribution selected should give the minimum loss of periodicity. Again it can be shown that the distribution has to be taken from Table II. The third-stage pulse-selection gate operates in a completely analogous way: it optimizes the distribution of the two- and three-pulse groups that form the second-stage partial pulse train.

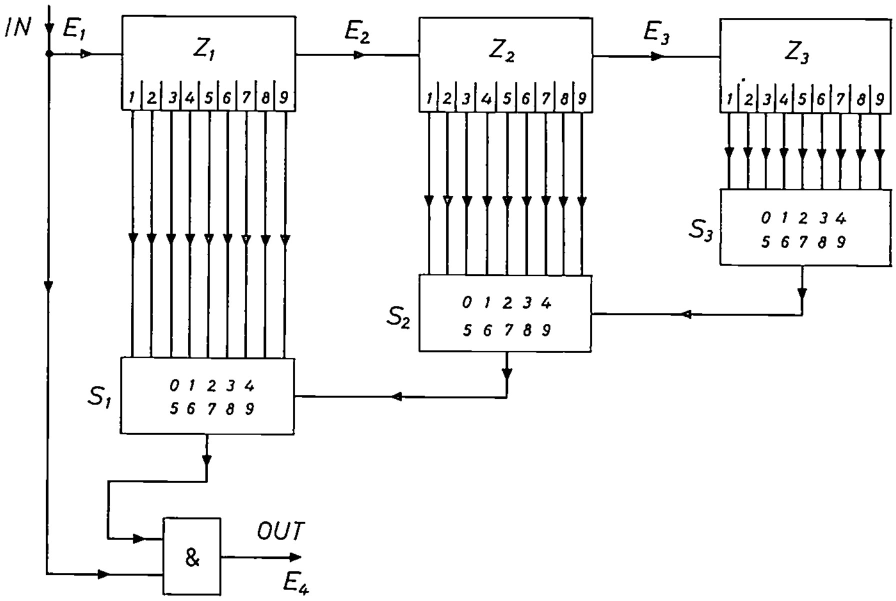


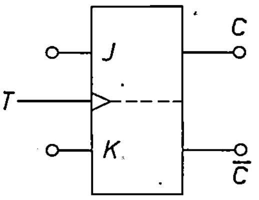
Fig. 2. Block diagram of a three-stage scaler with "controlled cascading". The stages consist of a decade unit Z_i and a control circuit S_i. Each control circuit receives simultaneously nine partial pulse trains of 1, 2, . . . , 9 pulses for each 10 input pulses. E_1 input signal to the circuit. The first figure of the multiplication factor is selected by means of S_1 and the second and third figures are selected by means of S_2 and S_3. The output E_4 appearing at *OUT* is produced in an AND circuit from the signal E_1 and the first-stage partial pulse train. The partial pulse trains produced in the other stages control S_1. The inputs E_2 and E_3 are the tens and hundreds respectively from E_1.

The next section discusses how a multistage digital rate multiplier can be designed with the aid of the new control principle, and includes a brief description of the novel decade stage.

Digital rate multipliers with controlled cascading

Fig. 2 shows a block diagram of a three-stage circuit based on the new control principle. The separate stages contain a decade unit Z_i and a control circuit S_i. The decade units receive the input signals E_i. The output E_4 is produced in the AND circuit from the input signal and the partial pulse train from the first stage. Apart from the circuits for controlled cascading, the control circuits include the switches that select the multiplication factor. Assuming again that the multiplication is by 0.423, S_1 is set at 4, S_2 at 2, and S_3 at 3. The second and the third decade units are fed by signals whose frequencies are a tenth and a hundredth of the input frequency respectively. Each decade unit simultaneously produces nine partial pulse trains consisting of 1, 2, . . . , or 9 pulses for each 10 input pulses. These pulse trains, which appear at the correspondingly numbered output terminals, feed the control circuits. Controlled cascading is obtained by connecting the three control circuits as a cascade, with the third circuit controlling the second, which in turn controls the first. A control pulse lasts as long as a pulse train of ten input pulses reaching the stage to be controlled. The reason is that during the complete ten-pulse period the switch of the controlled stage should be set at either the required number plus one, or the required number itself. These settings are effected by control-amplitude levels "1" and "0" respectively. More details will be given later.

Adding the input signals of the second and third stages to their own outputs, although necessary in fig. 1*a*, can be omitted in the circuit of fig. 2. In this respect the new circuit is simpler than the conventional circuits. The absence of separate decoders and pulse-selection gates in fig. 2 represents an even more striking simplification. This is due to the specially developed frequency-divider decade unit; these units themselves select the optimum distribution of the pulses for the partial pulse trains. The proper sequences of voltage levels appear simultaneously at the nine output terminals of a decade unit, making separate decoders and pulse-selection gates unnecessary.



J	K	C(n+1)
0	0	C(n)
0	1	0
1	0	1
1	1	C̄(n)

Fig. 3. Bistable circuit (flip-flop) with condition inputs J and K. The output signal appears at C, its complement at $\bar{C}$. T input for clock pulse. The truth table shows the effect of J and K upon C. The expression $C(n+1)$ represents the voltage level at C when the clock pulse $n+1$ has arrived and $C(n)$ represents the level at C when the clock pulse n has arrived. For example, the level is maintained, i.e. $C(n+1) = C(n)$, when $J = K = 0$.

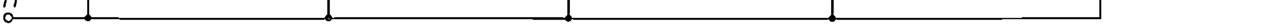


Fig. 4. Diagram of the new decade unit; separate decoders and pulse-selection gates are not required. The circuit consists of five bistable circuits with condition inputs J and K (Philips FCJ 101, see fig. 3); the input signal arrives at the five clock-pulse inputs T_i simultaneously. Output partial pulse trains consisting of 1, 9, 2, 8, etc. pulses per 10 input pulses appear at C_1, $\bar{C}_1$, C_2, $\bar{C}_2$, etc. respectively. All the K inputs are set permanently at "1" via H. The AND and NOT circuits that provide the J signals are not shown.

The decade unit consists of five bistable circuits (flip-flops) of the JK type. Besides a clock-pulse input, these bistable circuits possess two condition inputs, labelled J and K (see *fig. 3*). There are two output terminals C and $\bar{C}$, which carry complementary signals. The effect of the condition inputs upon the behaviour of a bistable circuit follows from the truth table. It can be seen that J sets the element to the "1" state and K resets it to the "0" state; in addition the combination $J = K = 1$ is allowed.

A circuit diagram of the new decade unit is shown in *fig. 4*. This diagram also shows bistable circuits with two J inputs. The voltage level to be used for J in the truth table is then equal to the output level of an OR circuit whose inputs are the signals at the double J input. All the bistable circuits function synchronously, i.e. each input pulse arrives at the five clock-pulse inputs simultaneously. The terminals C_1, C_2, . . . , C_5 carry the partial pulse trains consisting of 1, 2, . . . , 5 pulses respectively for each 10 clock pulses. All the K inputs are kept at voltage level "1" continuously to ensure that each bistable circuit cannot stay in state "1" any longer than the time elapsing between two consecutive clock pulses, corresponding to the pulse trains required at the terminals C_1, C_2, . . . , C_5. The terminals $\bar{C}_4$, $\bar{C}_3$, $\bar{C}_2$ and $\bar{C}_1$ carry the other pulse trains consisting of 6, 7, 8 and 9 pulses respectively. *Table III* gives the nine partial pulse trains.

The distribution in time of the pulses deviates slightly from the distribution shown in Table II, with the important advantage that the decade-unit circuit is far simpler. The remaining degree of bunching is not noticeably worse, and a good compromise is obtained between suppressing pulse bunching and simplifying the circuits. Apart from the five bistable circuits, all that is required is a few AND and NOT circuits to produce the J signals. These signals are indicated in fig. 4; to make the presentation clearer, the circuits are not shown. The signal at C_1, with only one pulse for each ten input pulses, functions as clock signal for the next decade unit.

Fig. 5 is a detailed diagram for a digital rate multiplier with three decade units of the type described above, with controlled cascading. With the aid of this diagram we shall now discuss in more detail how controlled cascading operates. Switch S_3 has one array of ten contacts, the other switches have two arrays of ten contacts. In each group of ten, one contact always carries the "0" level, enabling zero to be included as a figure of the desired decimal fraction. The other contacts of the switches are connected to the terminals of the decade units, which carry the output pulse trains.

Table III. The partial pulse trains of voltage levels "1" and "0" as they appear at the output terminals C_1, C_2, . . . , $\bar{C}_1$ of the new decade unit. The first column gives the numbers of the clock pulses.

Terminal Number of clock pulse	C_1	C_2	C_3	C_4	C_5	$\bar{C}_4$	$\bar{C}_3$	$\bar{C}_2$	$\bar{C}_1$
1	0	0	0	1	1	0	1	1	1
2	0	0	1	0	0	1	0	1	1
3	0	1	0	0	1	1	1	0	1
4	0	0	0	1	0	0	1	1	1
5	0	0	0	0	1	1	1	1	1
6	0	0	1	1	0	0	0	1	1
7	1	0	0	0	1	1	1	1	0
8	0	1	0	0	0	1	1	0	1
9	0	0	1	1	1	0	0	1	1
10	0	0	0	0	0	1	1	1	1

For the "lower" arrays of S_1 and S_2 the number of pulses in the partial pulse trains is equal to the switch contact number; for the "higher" arrays, however, the number of pulses in a pulse train exceeds the contact number by one.

The input signal is applied together with the control signal s_1 to an AND circuit, whose output signal represents the output pulse train at *OUT*. The control signals in the first and second stages (s_1 and s_2) are produced by an OR circuit fed by two AND circuits. The input signals to these two AND circuits are the

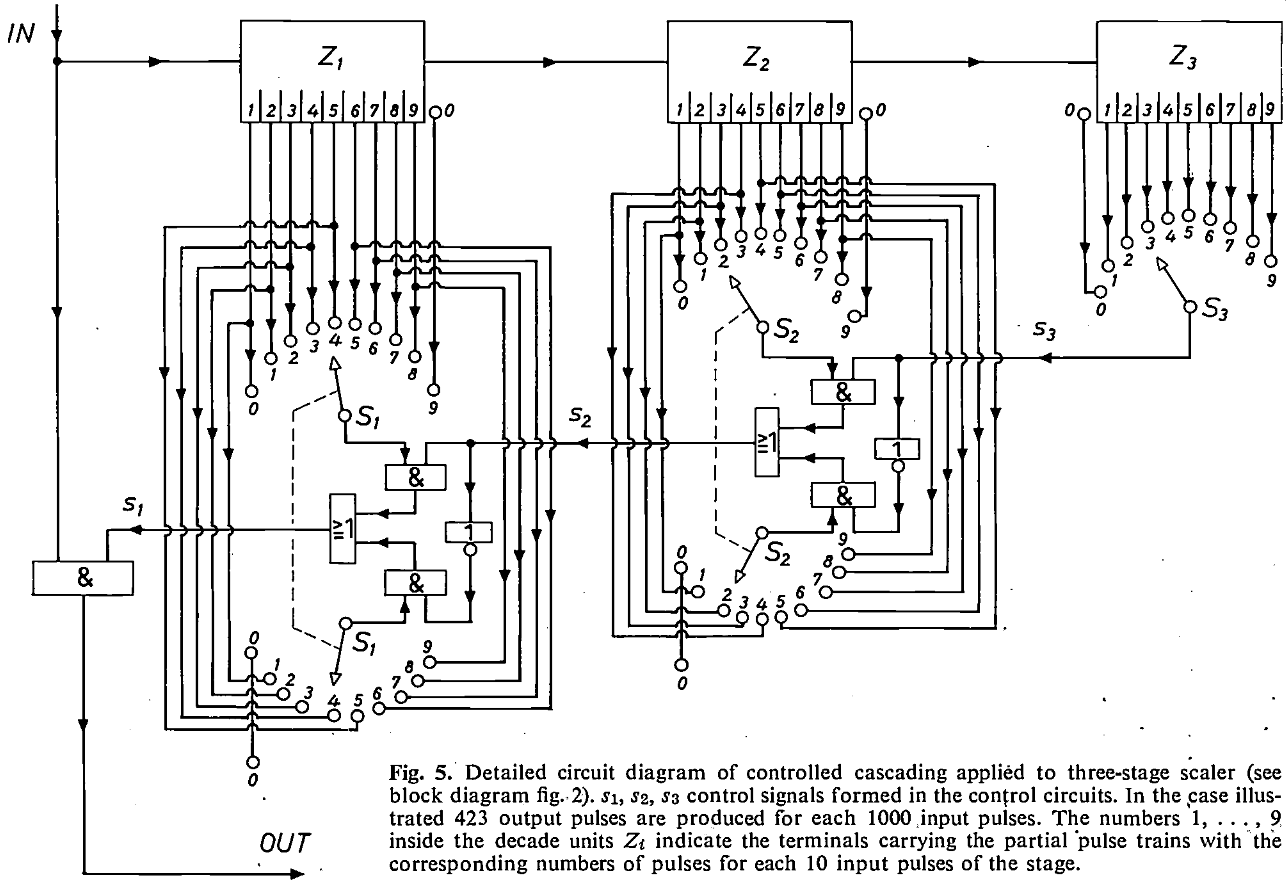


Fig. 5. Detailed circuit diagram of controlled cascading applied to three-stage scaler (see block diagram fig. 2). s_1, s_2, s_3 control signals formed in the control circuits. In the case illustrated 423 output pulses are produced for each 1000 input pulses. The numbers 1, . . . , 9 inside the decade units Z_i indicate the terminals carrying the partial pulse trains with the corresponding numbers of pulses for each 10 input pulses of the stage.

two partial pulse trains selected by the corresponding switch S_1 or S_2, and the control signal produced in the next stage, both inverted and uninverted.

As we saw earlier the controlled cascading in fig. 5 starts from the third stage; there are two initial conditions: control signal s_3 is equal to either "1" or "0". Assume that the switches S_1, S_2 and S_3 are set at j, k and l respectively, where all integers 1, 2, . . . , 9 and 0 are possible choices. The initial conditions occur l and $(10 - l)$ times for each 1000 first-stage input pulses. During the time that the control signal s_3 assumes the value "1" or — of equal duration — the value "0", the partial pulse trains produced in the second decade contain respectively $k + 1$ or k times the value "1" (i.e. a pulse), and correspondingly $(9 - k)$ or $(10 - k)$ times the value "0" (no pulse). This last statement also holds for the first decade unit, provided that s_3 is replaced by s_2 and k by j. In *fig. 6* the number of times the output of the circuit assumes the value "1" or "0" is given as a function of the control signal s_2. In an analogous way fig. 6 also shows the number of times that "1" and "0" occur in the signal s_2, related to the two initial conditions s_3 equal to either "1" or "0". These distributions of "1" and "0" are easily found by an inspection, using AND and OR truth tables, of what happens to the signals "1" and "0" originating from the third switch. The controlled cascading can then clearly be seen: when the control signal is "1" there are always $j + 1$ and not j pulses in the output signal during any period of ten input pulses.

To illustrate the improvement that can be achieved with the new three-stage rate multiplier, let us look again at the flow measurement of Table I. Applying controlled cascading in this case gives the results of *Table IV*. The first and third columns, the volume passed and the true times are the same as those in Table I. The second and fourth columns clearly show the much better agreement between the times indicated by the output pulses and the true times. For every 1000 first-stage input pulses the control signal from the third stage consists of the pulse numbered 700 alone. The numbers of the control pulses from the second stage are 70, 170, 270, 370, 470, 570, 670, 730, 780, 870, 970. The occurrence of 730 and 780 instead of 770 is due to the third-stage control.

Table IV. Results of the flow measurement of Table I, but now using the improved three-stage digital rate multiplier of fig. 5. The agreement between true times and the recorded times is much better (see also the fourth column). The output pulses numbered 737 and 787 are formed from the third-stage control signal via the second-stage signal ($s_3 = 1$, $s_2 = 1$, see fig. 6), the other output pulses are due to second-stage control signals only ($s_3 = 0$, $s_2 = 1$).

Flow volume in litres	Recorded time	True time	Ratio
1	77	91	1.18
2	177	182	1.03
3	277	273	0.98
4	377	364	0.97
5	477	455	0.95
6	577	546	0.95
7	677	637	0.94
8	737	728	0.99
9	787	819	1.04

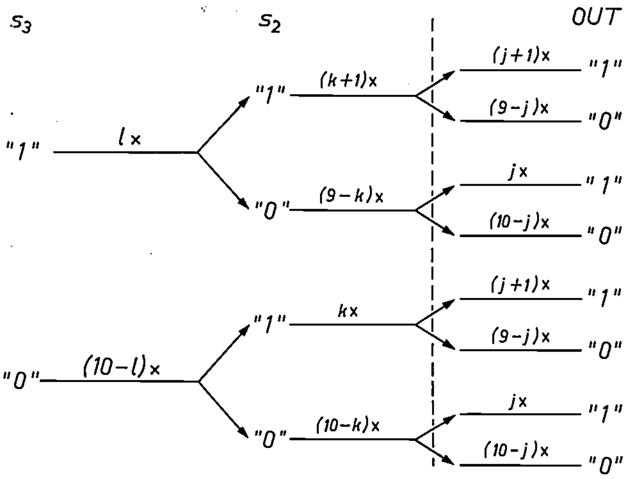


Fig. 6. Distribution of levels "1" (pulse) and "0" (no pulse) in the output signal OUT and in the control signals s_3 and s_2 for the scaler of fig. 5. The total number of output pulses per 1000 input pulses is: $l\{(k + 1)(j + 1) + (9 - k)j\} + (10 - l)\{k(j + 1) + + (10 - k)j\} = 100j + 10\,k + l$, where j, k, l are the settings of the selector switches S_1, S_2, S_3 (see fig. 5). × number of times that the levels occur.

The restriction to the three stages shown in fig. 5 is by no means an essential feature. We have more recently constructed a four-stage digital rate multiplier of the new type. The predicted performance has been achieved, and the circuit will multiply frequencies up to about 20 kHz by any factor with four places of decimals from 0.0001 to 0.9999.

Summary. Multistage scalers based upon digital multiplication, and used in frequency measurement and control, generally give some pulse bunching at the output. Bunching occurs both when partial pulse trains are formed in the separate stages, and when pulse trains are combined to form the output signal.

"Controlled cascading", a new control principle, eliminates pulse bunching when the pulse trains are combined. The principle leads to a first-stage partial pulse train consisting of groups of either s or $s + 1$ pulses for each ten input pulses. The number s is the first figure of the desired decimal multiplication fraction. In controlled cascading the second-stage partial pulse train (influenced by the third, etc.) regulates the order of s and $s + 1$ selections in the first stage in such a way that the partial pulse train from the first stage represents the complete output signal with the optimum periodicity.

An improved decade unit consisting of five synchronously operating JK bistable circuits is also described. The ten output terminals carry simultaneously partial pulse trains of 0, 1, 2, . . ., 9 pulses (for each 10 input pulses), each with optimized periodicity. Pulse-order selection is built in by means of a small number of logic elements; separate decoders are not necessary.

Three- and four-stage digital rate multipliers have been built on the new principle and using the improved decade units. The four-stage version multiplies frequencies up to 20 kHz by any factor with four decimal places from 0.0001 to 0.9999, with optimum reduction of pulse bunching.

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 Forschungslaboratorium Aachen GmbH, Weißhausstraße, 51 Aachen, Germany *A*
Philips Forschungslaboratorium Hamburg GmbH, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (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: Amplificateur de tension quasi continue très bas niveau à dérive extrêmement faible.
Bull. TRACE **4**, 382-391, 1970 (No. 12). *L*

J. R. A. Beale: Microelectronics.
Physics Bull. **21**, 60-65, 1970 (Feb.). *M*

K. H. Beckmann & N. J. Harrick (Philips Laboratories Briarcliff Manor, N.Y., U.S.A.): Hydrides and hydroxyls in thin silicon dioxide films.
J. Electrochem. Soc. **118**, 614-619, 1971 (No. 4). *H*

K. Bethe & F. Welz: Preparation and properties of (Ba,Sr)TiO$_3$ single crystals.
Mat. Res. Bull. **6**, 209-217, 1971 (No. 4). *H*

R. N. Bhargava, S. K. Kurtz (both with Philips Laboratories Briarcliff Manor, N.Y., U.S.A.), **A. T. Vink & R. C. Peters:** Spectroscopic observation of a vacancy complex in GaP.
Phys. Rev. Letters **27**, 183-185, 1971 (No. 4). *E*

G. Blasse & A. Bril: Luminescence in some tantalate host lattices.
J. Solid State Chem. **3**, 69-74, 1971 (No. 1). *E*

P. M. Boers, G. A. Acket, D. H. Paxman & R. J. Tree: Observation of high-field domains in *n* type indium phosphide.
Electronics Letters **7**, 1-2, 1971 (No. 1). *E, M*

H. van den Boom: The application of a double modulation technique in spectrometers for electron paramagnetic resonance.
Rev. sci. Instr. **42**, 524-526, 1971 (No. 4). *E*

G. A. Bootsma & H. J. Gassen: A quantitative study on the growth of silicon whiskers from silane and germanium whiskers from germane.
J. Crystal Growth **10**, 223-234, 1971 (No. 3). *E*

A. Boucher, J. P. Chané & E. Fabre: Contrôle du dopage dans la croissance épitaxiale d'arséniure de gallium.
Rev. Physique appl. **6**, 5-10, 1971 (No. 1). *L*

H. Bouma (Institute for Perception Research, Eindhoven): Enige aspecten van normale leesprocessen.
T. Orthopedagogiek **10**, 2-13, 1971 (No. 1).

C. J. Bouwkamp: On some special squared rectangles.
J. combin. Theory B **10**, 206-211, 1971 (No. 3). *E*

C. J. Bouwkamp: Numerical computation of the radiation impedance of a rigid annular ring vibrating in an infinite plane rigid baffle.
J. Sound Vibr. **17**, 499-508, 1971 (No. 4). *E*

M. J. G. Braken: Micro-plasmalasproeven met temperatuurmeting.
Lastechniek **37**, 43-46, 1971 (No. 3). *E*

P. C. Brandon: Inhibition of photophosphorylation by β-bromo-β-nitrostyrene.
FEBS Letters **14**, 153-156, 1971 (No. 3). *E*

P. B. Braun, J. Hornstra & J. I. Leenhouts: The crystal structure of the carotenoidal compound 1,14-bis-(2′,6′,6′-trimethylcyclohex-1′-enyl)-3,12-dimethyltetradeca-1,3,5,7,9,11,13-heptaene-6,9-dinitrile.
Acta cryst. **B 27**, 90-95, 1971 (No. 1). *E*

J. C. Brice: The variation of interface segregation coefficients with growth rate of crystals.
J. Crystal Growth **10**, 205-206, 1971 (No. 2). *M*

J. C. Brice, O. F. Hill, P. A. C. Whiffin & J. A. Wilkinson: The Czochralski growth of barium strontium niobate crystals.
J. Crystal Growth **10**, 133-138, 1971 (No. 2). *M*

J. J. Brissot & C. Belin: Preparation of artificial calcite single crystals by solvent zone melting.
J. Crystal Growth **8**, 213-215, 1971 (No. 2). *L*

J. van den Broek, W. Kwestroo, C. Langereis & A. Netten: Physical properties of lead-tin monoxide, a new photoconductive layer material.
Proc. 3rd Int. Conf. on Photoconductivity, Stanford 1969, pp. 195-198; 1971. *E*

A. Broese van Groenou: De rol van de anisotropie in magnetische filtermaterialen.
Ned. T. Natuurk. **37**, 317-324, 1971 (No. 12). *E*

K. H. J. Buschow, H. J. van Daal, F. E. Maranzana & P. B. van Aken: Kondo sidebands in CeAl$_3$ and related pseudobinary compounds.
Phys. Rev. B **3**, 1662-1670, 1971 (No. 5). *E*

K. H. J. Buschow & A. S. van der Goot: The crystal structure of rare-earth aluminium compounds R$_2$Al.
J. less-common Met. **24**, 117-120, 1971 (No. 1). *E*

K. H. J. Buschow & A. S. van der Goot: Composition and crystal structure of hexagonal Cu-rich rare earth - copper compounds.
Acta cryst. **B 27**, 1085-1088, 1971 (No. 6). *E*

K. H. J. Buschow & R. P. van Stapele: Magnetic properties of the intermetallic compounds RFe$_2$.
J. Physique **32**, C1/672-674, 1971 (Colloque No. 1 Vol. II). *E*

F. J. du Chatenier: Some properties of vapour-deposited layers of red lead monoxide.
Proc. 3rd Int. Conf. on Photoconductivity, Stanford 1969, pp. 199-203; 1971. *E*

M.-C. Chen (National Chiao Tung University, Hsinchu, Taiwan), **J. O. Artman, D. Sengupta** (both with Carnegie-Mellon University, Pittsburgh, Pa.) & **J. C. M. Henning:** Site occupancy of the PbWO$_4$:Co and PbMoO$_4$:Co systems.
Phys. Rev. B **4**, 1387-1389, 1971 (No. 4). *E*

J. B. Coughlin: Clean rooms in a contaminated environment.
Research **4**, 59-63, 1971 (No. 3). *M*

C. Crevecoeur & H. J. de Wit: Dielectric losses in As$_2$Se$_3$ glass.
Solid State Comm. **9**, 445-449, 1971 (No. 8). *E*

H. J. van Daal, F. E. Maranzana & K. H. J. Buschow: Kondo side-bands in cerium intermetallic compounds.
J. Physique **32**, C1/424-431, 1971 (Colloque No. 1, Vol. I). *E*

H. Dammann & K. Görtler: High-efficiency in-line multiple imaging by means of multiple phase holograms.
Optics Comm. **3**, 312-315, 1971 (No. 5). *H*

H. Dammann, G. Groh & M. Kock: Restoration of faulty images of periodic objects by means of self-imaging.
Appl. Optics **10**, 1454-1455, 1971 (No. 6). *H*

H. Dammann & M. Kock: Removal of nonperiodic structures from a periodic image by means of spatial filtering.
Optics Comm. **3**, 251-253, 1971 (No. 4). *H*

P. Delsarte: On cyclic codes that are invariant under the general linear group.
IEEE Trans. **IT-16**, 760-769, 1970 (No. 6). *B*

F. Desvignes: Rayonnement terrestre et senseurs d'horizon.
Acta Electronica **13**, 227-247, 1970 (No. 3). *L*

F. Desvignes, C. Hily & Mme J. Michel: Senseur d'horizon pour un satellite à attitude stabilisée selon trois axes en orbite basse.
Acta Electronica **13**, 249-279, 1970 (No. 3). *L*

R. H. Dijken (Philips Domestic Appliances Division, Drachten, Netherlands): Optimization of small AC series commutator motors.
Thesis, Eindhoven 1971.

C. Z. van Doorn & D. J. Schipper: Luminescence of O$_2^-$, Mn^{2+} and Fe^{3+} in sodalite.
Physics Letters **34A**, 139-140, 1971 (No. 3). *E*

W. F. Druyvesteyn & F. A. de Jonge: A special kind of magnetic domain: hollow bubble with at its centre a bubble.
Physics Letters **36A**, 1-2, 1971 (No. 1). *E*

H. Duifhuis (Institute for Perception Research, Eindhoven): Audibility of high harmonics in a periodic pulse, II. Time effect.
J. Acoust. Soc. Amer. **49**, 1155-1162, 1971 (No. 4, Part 2).

F. L. Engel (Institute for Perception Research, Eindhoven): Visual conspicuity, directed attention and retinal locus.
Vision Res. **11**, 563-575, 1971 (No. 6).

U. Enz, R. Metselaar & P. J. Rijnierse: Photomagnetic effects.
J. Physique **32**, C1/703-709, 1971 (Colloque No. 1, Vol. II). *E*

W. G. Essers, G. Jelmorini & G. W. Tichelaar: Metal transfer from coated electrodes.
Metal Constr. Brit. Welding J. **3**, 151-154, 1971 (No. 4). *E*

W. G. Essers, A. C. H. J. Liefkens & G. W. Tichelaar: Plasma-MIG-welding.
Proc. Conf. on Advances in Welding Processes, 1970, pp. 216-219; publ. The Welding Institute, Cambridge 1971. *E*

K. G. Freeman, R. N. Jackson, P. L. Mothersole (Mullard Central Appl. Lab., Mitcham, England) & **S. J. Robinson:** Some aspects of direct television reception from satellites.
SERT J. **5**, 75-80, 1971 (No. 4). *M*

R. Genève: Aspects physiques de la thermographie médicale et instrumentation.
Rev. gén. Thermique **10**, 239-254, 1971 (No. 111). *L*

C. J. Gerritsma, W. J. A. Goossens & A. K. Niessen: The helical twist in a cholesteric Grandjean-Cano pattern.
Physics Letters **34A**, 354-355, 1971 (No. 7). *E*

J. A. Geurst: Generalised vorticity in the theory of liquid crystals.
Physics Letters **36A**, 63-64, 1971 (No. 1). *E*

J.-M. Goethals: Codes linéaires définis par des polynomes.
Rev. CETHEDEC **7**, No. 22, 1-13, 1970. *B*

A. S. van der Goot & K. H. J. Buschow: Lattice constants and Curie temperatures of $CaCu_5$-type thorium-cobalt compounds.
Phys. Stat. sol. (a) **5**, 665-668, 1971 (No. 3). *E*

H. C. de Graaff: A.c. small-signal model for bipolar transistors in saturation.
Electronics Letters **7**, 73-75, 1971 (No. 3). *E*

C. A. A. J. Greebe, P. A. van Dalen, T. J. B. Swanenburg & J. Wolter: Electric coupling properties of acoustic and electric surface waves.
Physics Repts. **1C**, 235-268, 1971 (No. 5). *E*

G. Groh & M. Kock: Hologramme von Röntgenbildern.
Röntgenblätter **24**, 451-455, 1971 (No. 9). *H*

A. J. Guest & R. Pook: The applications of channel electron multipliers.
Proc. Electro-Optics '71 Int. Conf., Brighton 1971, pp. 277-289. *M*

H. W. Hanneman: The systematic and the random errors due to element tolerances of electrical networks.
Philips Res. Repts. **26**, 414-423, 1971 (No. 5). *E*

H. Haug & K. Weiss: Hydrodynamic equations for the condensate and the depletion of helium II.
Phys. Rev. A **3**, 717-724, 1971 (No. 2). *E*

H. Haug, K. Weiss & M. van Hove: Heat exchange in liquid helium by phonon tunneling through very thin plates.
J. low Temp. Phys. **4**, 263-271, 1971 (No. 3). *E*

E. E. Havinga & M. H. van Maaren: Oscillatory dependence of superconductive critical temperature on number of valency electrons in alloys of mainly non-transition metal compounds.
Proc. 12th Int. Conf. on Low Temperature Physics, Kyoto 1970, pp. 355-356. *E*

N. Hazewindus & J. M. van Nieuwland: Een automatische meetopstelling voor het onderzoek van het centrum van een cyclotron.
Ned. T. Natuurk. **37**, 292-295, 1971 (No. 11). *E*

J. C. M. Henning: Weak exchange interactions in chromium doped $ZnGa_2O_4$.
Physics Letters **34A**, 215-216, 1971 (No. 4). *E*

D. Hennings: The range of existence of perovskite phases in the system PbO-TiO_2-La_2O_3.
Mat. Res. Bull. **6**, 329-339, 1971 (No. 5). *A*

E. P. Honig & P. M. Mul: Tables and equations of the diffuse double layer repulsion at constant potential and at constant charge.
J. Colloid Interface Sci. **36**, 258-272, 1971 (No. 2). *E*

E. P. Honig, G. J. Roebersen (University of Utrecht) **& P. H. Wiersema** (Univ. Utrecht): Effect of hydrodynamic interaction on the coagulation rate of hydrophobic colloids.
J. Colloid Interface Sci. **36**, 97-109, 1971 (No. 1). *E*

F. N. Hooge & J. L. M. Gaal: Experimental study of $1/f$ noise in thermo e.m.f.
Philips Res. Repts. **26**, 345-358, 1971 (No. 5). *E*

S. van Houten: Fysische aspecten van enkele moderne display-technieken.
Ingenieur **83**, ET 40-45, 1971 (No. 11). *E*

D. van Houwelingen (Philips Electronic Components and Materials Division (Elcoma), Eindhoven) **& A. A. Kruithof** (Eindhoven University of Technology): Transition probabilities of some ArI and ArII spectral lines.
J. quant. Spectroscopy rad. Transfer **11**, 1235-1243, 1971 (No. 8).

J. P. Hurault, K. Maki (Faculté des Sciences d'Orsay) **& M. T. Béal-Monod** (Fac. Sci. Orsay): Fluctuations of the order parameter in small superconducting samples.
Phys. Rev. B **3**, 762-768, 1971 (No. 3). *L*

B. B. van Iperen & H. Tjassens: Influence of carrier velocity saturation in the unswept layer on the efficiency of avalanche transit time diodes.
Proc. IEEE **59**, 1032-1033, 1971 (No. 6). *E*

R. N. Jackson: Visual displays: a preview.
Proc. Electro-Optics '71 Int. Conf., Brighton 1971, pp. 521-538. *M*

J. Janse: Numerical treatment of electron lenses with perturbed axial symmetry.
Optik **33**, 270-281, 1971 (No. 3). *E*

C. J. G. F. Janssen, H. Jonker & A. Molenaar: Photofabrication methods in PD processes.
Plating **58**, 42-46, 1971 (No. 1). *E*

W. H. de Jeu, C. J. Gerritsma & A. M. van Boxtel: Electrohydrodynamic instabilities in nematic liquid crystals.
Physics Letters **34A**, 203-204, 1971 (No. 4). *E*

F. A. de Jonge, W. F. Druyvesteyn & A. G. H. Verhulst: Observations and properties of a new domain: hollow bubble.
J. appl. Phys. **42**, 1270-1272, 1971 (No. 4). *E*

J. T. Klomp: Heat-resistant ceramic-to-metal seals.
Welding J. **50**, 88s-90s, 1971 (No. 2). *E*

H. Koelmans: Application of semiconducting thin films.
Thin Solid Films **8**, 19-33, 1971 (No. 1). *E*

E. Krätzig: Ultrasonic investigation of superconducting properties induced by proximity effects.
Solid State Comm. **9**, 1205-1209, 1971 (No. 14). *H*

D. J. Kroon: Analyse van de buitenlucht.
Natuur en Techniek **39**, 245-256, 1971 (No. 6). *E*

H. K. Kuiken: The effect of normal blowing on the flow near a rotating disk of infinite extent.
J. Fluid Mech. **47**, 789-798, 1971 (No. 4). *E*

J. Lebrun: A new CdTe thin film solar cell.
Conf. Rec. 8th IEEE Photovoltaic Specialists Conf., Seattle 1970, pp. 33-39. *L*

J. I. Leenhouts: The crystal and molecular structure of 5,7-dibromo-2-(3,5-dibromo-2-hydroxyphenyl)-2-methoxymethoxy-3(2H)-benzofuranone.
Rec. Trav. chim. Pays-Bas **90**, 385-388, 1971 (No. 5). *E*

M. Lemke, H. J. Schmitt & D. Vollmer: Empfänger für das 12-GHz-Fernsehen.
Rundfunktechn. Mitt. **15**, 163-166, 1971 (No. 4). *H*

K. Löhn: Nichtlinearitäten in Modulations- und Mischsystemen.
Ann. Physik (7) **26**, 345-348, 1971 (No. 4). *A*

F. K. Lotgering: Exchange interactions in semiconducting chalcogenides with normal spinel structure from an experimental point of view.
J. Physique **32**, C1/34-38, 1971 (Colloque No. 1, Vol. I). *E*

F. K. Lotgering: On the magnetic interaction between Fe^{3+} and Ni^{2+} ions occupying octahedral sites in oxyspinels.
Solid State Comm. **9**, 1309-1312, 1971 (No. 15). *E*

F. E. Maranzana & P. Bianchessi: Study of the temperature dependence of the resistivity in the Kondo side-band model.
Phys. Stat. sol. (b) **43**, 601-610, 1971 (No. 2). *E*

D. Mateika: Ein isothermes Schmelztropfen-Verfahren zur Herstellung von großen, sehr reinen und homogenen Blei-Zinn-Tellurid-Einkristallen aus der Dampfphase.
J. Crystal Growth **9**, 249-254, 1971. *H*

R. Memming & F. Möllers: Spectroscopic studies of electrochemical reactions of adsorbed dye layers.
Symp. Faraday Soc. **4**, 145-156, 1970. *H*

R. Metselaar & P. Rem: Radio frequency sputtering of ferrimagnetic thin films.
Czech. J. Phys. **B 21**, 558-562, 1971 (No. 4/5). *E*

F. Meyer & M. J. Sparnaay: Ellipsometric and gas-volumetric investigation of adsorption reactions on clean silicon and germanium.
Symp. Faraday Soc. **4**, 17-26, 1970. *E*

K. Mouthaan & H. P. M. Rijpert: Nonlinearity and noise in the avalanche transit-time oscillator.
Philips Res. Repts. **26**, 391-413, 1971 (No. 5). *E*

J. Neirynck: Conditions on the group delay of transmittances.
Electronics Letters **7**, 7-8, 1971 (No. 1). *B*

J. M. van Nieuwland: Versnellers voor hoogenergetische zware ionen.
Ned. T. Natuurk. **37**, 91-99, 1971 (No. 6). *E*

W. J. Oosterkamp, W. J. L. Scheren & J. J. F. de Wijk: Characteristic properties of scintillation counters for exposure measurements between 8 and 50 keV.
Coll. int. Problèmes de radioprotection liés à l'émission de rayons X parasites par des systèmes électroniques, Toulouse 1970, pp. 415-431; 1971. *E*

A. van Oostrom: A colour superposition technique in field ion microscopy.
Ned. T. Vacuümtechniek **9**, 13-15, 1971 (No. 1). *E*

C. van Opdorp: Anomalous diffusion of Al into SiC.
Solid-State Electronics **14**, 613-625, 1971 (No. 7). *E*

G. den Ouden: The electric arc.
Philips Welding Reporter 1971, No. 1, pp. 3-12. *E*

T. H. Peek: Dynamic polarization detection.
Optics Comm. **2**, 377-382, 1971 (No. 8). *E*

T. H. Peek: Use of Savart plates in grating interferometers.
Appl. Optics **10**, 1092-1096, 1971 (No. 5). *E*

A. Pelissier: Les propulseurs ioniques à ionisation du césium par contact.
La technologie spatiale française, Paris, I.P.F., 1971, tome 1, pp. 395-415. *L*

A. Rabenau, H. Rau & G. Rosenstein: Phase relationships in the gold-selenium system.
J. less-common Met. **24**, 291-299, 1971 (No. 3). *A*

O. Reifenschweiler: A high output sealed-off neutron tube with high reliability and long life.
Proc. Int. Conf. on Modern Trends in Activation Analysis, Washington 1968, Vol. 2, 905-910; 1971. *E*

J. G. Rensen, J. A. Schulkes & J. S. van Wieringen: Effet Mössbauer dans du $BaFe_{12}O_{19}$; substitutions et mouvement thermique anisotrope.
J. Physique **32**, C1/924-925, 1971 (Colloque No. 1, Vol. II). *E*

H. P. M. Rijpert: Microgolf-oscillatoren met lawinelooptijddiodes.
Polytechn. T. Elektr. **26**, 378-383, 1971 (No. 10). *E*

T. E. Rozzi, J. H. C. van Heuven & A. Meyer: Linear networks as Möbius transformations and their invariance properties.
Proc. IEEE **59**, 802-803, 1971 (No. 5). *E*

K. H. Sarges: Nuclear surface spin waves.
Physics Letters **36A**, 9-10, 1971 (No. 1). *H*

H. Schemmann: Theoretische und experimentelle Untersuchungen über das dynamische Verhalten eines Einphasen-Synchronmotors mit dauermagnetischem Läufer.
Thesis, Eindhoven 1971. *A*

C. Schiller: Polarité et orientation de composés semiconducteurs.
C.R. Acad. Sci. Paris **272B**, 764-766, 1971 (No. 12). *L*

H. J. Schmitt & K. H. Sarges: Wave propagation in microstrip.
Nachrichtentechn. Z. **24**, 260-264, 1971 (No. 5). *H*

J. F. Schouten (Institute for Perception Research, Eindhoven): The residue revisited.
Frequency analysis and periodicity detection in hearing, editors R. Plomp & G. F. Smoorenburg, publ. Sijthoff, Leiden 1970, pp. 41-58.

P. J. Severin: Measurement of the resistivity and thickness of a heterotype epitaxially grown silicon layer with the spreading-resistance method.
Philips Res. Repts. **26**, 359-372, 1971 (No. 5). *E*

J. G. Siekman: Electron beam drilling of curved holes into solids.
Electron and ion beam science and technology, Proc. 2nd Int. Conf., New York 1966, Vol. 1, pp. 155-161; 1969. *E*

J. G. Siekman: Boren met de elektronenbundel.
Ned. T. Natuurk. **37**, 21-25, 1971 (No. 2). *E*

J. G. Siekman: Materiaalbewerking met de CO_2-laser.
Elektrotechniek **49**, 299-304, 1971 (No. 7). *E*

M. Sintzoff: Modèles et déductions pour langages algorithmiques.
Mitt. Ges. Math. Datenverarb. Bonn **8**, 47-49, 1970. *B*

F. A. Staas & A. P. Severijns: Refrigeration by vortex motion in superfluid helium.
Proc. 3rd Int. Cryogenic Engng. Conf., Berlin 1970, pp. 320-324; 1971. *E*

R. P. van Stapele, J. S. van Wieringen & P. F. Bongers: Strong anisotropy in the cubic ferrimagnet $FeCr_2S_4$.
J. Physique **32**, C1/53-54, 1971 (Colloque No. 1, Vol. I). *E*

J. M. Stevels: Le verre dans l'électronique.
Silicates industr. **36**, 37-44, 1971 (No. 2). *E*

F. L. H. M. Stumpers: International co-operation in the suppression of radio interference — the work of C.I.S.P.R.
Proc. I.R.E.E. Australia **32**, 51-55, 1971 (No. 2). *E*

K. Teer: Elektronisch perspectief.
Ingenieur **83**, ET 31-38, 1971 (No. 11). *E*

J. P. Thiran & Ph. van Bastelaer: An accuracy study of filter synthesis methods.
IEEE Trans. **CT-18**, 203-205, 1971 (No. 1). *B*

G. W. Tichelaar, J. G. Verhagen & G. A. M. Willems: Narrow gap electro-gas welding.
Proc. Conf. on Advances in Welding Processes, 1970, pp. 220-225; publ. The Welding Institute, Cambridge 1971. *E*

H. van Tongeren: Radial density distribution measurements of neutral Cs in the positive column of a Cs-Ar dc discharge.
J. appl. Phys. **42**, 3011-3012, 1971 (No. 7). *E*

H. J. L. Trap: Elektronenleitung in Glas.
Nachrichtentechn. Z. **24**, 353-360, 1971 (No. 7). *E*

J. F. Verwey: Avalanche-injected current in MNOS structures.
Philips Res. Repts. **26**, 382-390, 1971 (No. 5). *E*

Q. H. F. Vrehen & A. Broese van Groenou: Ferromagnetic relaxation in porous polycrystalline ferrites.
J. Physique **32**, C1/156-158, 1971 (Colloque No. 1, Vol. I). *E*

W. L. Wanmaker (Philips Lighting Division, Eindhoven): Luminescerende verbindingen voor lampen.
Chem. Weekblad **67**, No. 27, 14-17, 2 juli 1971.

W. L. Wanmaker & D. Radielović (Philips Lighting Division, Eindhoven): De bereiding van fosforen.
Klei en Keramiek **21**, 146-153, 1971.

W. L. Wanmaker, J. G. Verriet & J. W. ter Vrugt (Philips Lighting Division, Eindhoven): Luminescence of phosphors based on the host lattice $ABGe_2O_6$ (A, B = Ca, Sr, Ba).
J. Solid State Chem. **3**, 194-196, 1971 (No. 2).

W. L. Wanmaker & J. W. ter Vrugt (Philips Lighting Division, Eindhoven): New phosphors for lamps.
Lighting Res. Technol. **3**, 147-151, 1971 (No. 2).

W. L. Wanmaker, J. W. ter Vrugt & J. G. Verlijsdonk (Philips Lighting Division, Eindhoven): Luminescence of alkaline earth yttrium and lanthanum phosphate-silicates with apatite structure.
J. Solid State Chem. **3**, 452-457, 1971 (No. 3).

W. L. Wanmaker, J. W. ter Vrugt & J. G. Verlijsdonk (Philips Lighting Division, Eindhoven): Synthesis of new compounds with apatite structure.
Philips Res. Repts. **26**, 373-381, 1971 (No. 5).

M. W. M. Wanninkhof & P. Zuidema (Philips Electronic Components and Materials Division (Elcoma), Eindhoven): Some experimental aspects concerning the thermogravimetric study of barium oxide condensation.
J. Physics E **4**, 321-323, 1971 (No. 4).

K. Weiss: Beweglichkeit von Elektronen und scheinbare Doppelschichtkapazität in AgBr.
Electrochim. Acta **16**, 201-221, 1971 (No. 2). *E*

K. Weiss: Überführungswärme und Überführungsvolumen von fehlgeordneten Atomen in einem anisothermen, elastisch verformten Kristall.
Z. phys. Chemie Neue Folge **74**, 81-92, 1971 (No. 3-6). *E*

K. Weiss: Zur Diffusion von Störstellen in den Gradienten des chemischen Potentials und der elastischen Spannung.
Z. angew. Physik **31**, 165-169, 1971 (No. 3). *E*

C. J. Wellekens: Generalisation of Vlach's method for the numerical inversion of the Laplace transform.
Electronics Letters **6**, 742-744, 1970 (No. 23). *B*

G. F. Weston: Glow discharge matrix displays.
Proc. Electro-Optics '71 Int. Conf., Brighton 1971, pp. 547-553. *M*

J. Wolter: Acoustoelectric amplification of Rayleigh waves in presence of a d.c. transverse magnetic field.
Physics Letters **34A**, 87-88, 1971 (No. 2). *E*

H. Zijlstra: Hysteresis measurements on RCo_5 microparticles.
J. Physique **32**, Cl/1039-1040, 1971 (Colloque No. 1, Vol. II). *E*

H. Zijlstra: Critical fields determining magnetic coercivity in microparticles of $SmCo_5$ and $LaCo_5$.
J. appl. Phys. **42**, 1510-1515, 1971 (No. 4). *E*

Contents of Electronic Applications **30**, No. 4, 1970:

M. J. Köppen: Three-stage 15 W power amplifier for the 470 MHz communication band (pp. 121-130).
E. B. G. Nijhof: Speed control of d.c. shunt motors (pp. 131-144).
D. R. Armstrong: TTL interfacing with GRL111 and GRL101 (pp. 145-154).
J. M. Bakker: Alphanumeric display using a MOS starburst character generator (pp. 156-161).

Contents of Mullard Technical Communications **12**, No. 112, 1971:

P. G. Boulton & D. Skelton: Transistor line timebase for 110° monochrome receivers (pp. 38-43).
P. G. Boulton, M. C. Gander & D. Gent: Transistor 90° colour line deflection and e.h.t. circuit (pp. 44-50).
H. F. Dittrich: Welding press coupling to a half-wave generator (pp. 51-57).
D. E. Nightingale: Induction heater operating at 2 MHz using magnetically beamed triode YD1352S (pp. 58-61).
P. J. Hart: High-power broadband linear amplifiers using transistors, type 810BLY/A (pp. 62-68).

Contents of Valvo Berichte **16**, No. 5, 1971:

J. Wölber: Aktive und semiaktive Modulation zur Korrektur der Ost-West-Verzeichnung von Farbfernsehbildern (pp. 127-134).
A. Petersen: Permanentmagnete zum Betätigen von Schutzgaskontakten (pp. 135-144).
W. Aschermann: Der Einsatz von monolithischen integrierten Schaltungen im Empfangsteil von Rundfunk- und Fernsehgeräten (pp. 145-158).

Volume 32, 1971, No. 3/4 pages 61-116 Published 11th February, 1972

PHILIPS TECHNICAL REVIEW

VOLUME 32, 1971, No. 5

Innovation in electronic devices

G. W. Rathenau

The article below gives the text of a talk which Prof. Rathenau gave on 12th October last in Washington at the International Electronic Device Meeting of the Institute of Electrical and Electronic Engineers (IEEE). Although various topics mentioned in the talk have been dealt with at greater length in earlier issues of our journal or will be the subject of forthcoming articles, and although some of the general aspects are quite widely known, the editors have nevertheless chosen this article for its special merits. Apart from treating device innovation in a systematic and lucid way, the paper concludes with some noteworthy comments on the social relevance of electronic systems, on the requirements that society of tomorrow will impose on them, and on the relationship between technological development and society in general.

Devices are elements of systems which are designed to fulfil some useful function in society at large. In order to design and construct devices that are increasingly better adapted to the functions demanded by society, it is necessary to exploit all the potentialities of science and technology. From this point of view the device designer depends both on the society in which he lives and on the scientific knowledge and technological skills upon which he can draw. In what ways will science and technology on the one hand and society on the other affect future developments?

I should like first of all to look at science and technology and attempt to analyse the present situation in order to discern the main trends (*Table I*). Having done that, I shall examine some of the social aspects.

The physical principles that govern the behaviour of electromagnetic waves and free electric charges such as electrons in present-day electronic devices rely on work done by Faraday and Maxwell more than a hundred years ago. To learn about the behaviour of materials in devices we can turn to the quantum theory of solids, which has developed into a mature and reliable instrument in the last forty years. Where radiation is involved in electronic devices, the theory is scarcely any younger. We may therefore be expected to know broadly which physical phenomena are available to serve in electronic devices. But it would be foolish to conclude from this that science and technology leave no room for innovation in electronic devices. Let me try to indicate where I believe such innovation is to be looked for. In doing so I shall draw for convenience mainly on examples from our own laboratories. I trust you will not misinterpret convenience in this matter as vanity.

There is a type of innovation in which established science and technology are used to meet newly dis-

Table I. Scientific, technological and social aspects and potential for innovation in electronic devices.

1. The underlying physical principles were established long ago.
2. Latent needs may be brought to light.
3. It may become apparent how to achieve long-desired functions with existing knowledge.
4. Technological innovations can be brought about in various ways, e.g.:
   4.1. Through the improvement of device performance, by
      4.1.1. Design improvements,
      4.1.2. Greater dimensional accuracy,
      4.1.3. Better materials technology.
   4.2. Through improvement of the production process.
5. New materials may be developed.
6. Extreme physical conditions might be used.
7. New physical effects may be found.
8. Not all newly discovered technological advances are economically feasible or of equal social relevance.

Prof. Dr. G. W. Rathenau, Director of Philips Research Laboratories, Eindhoven.

covered needs. The recognition of the need is the essence of this type of innovation, the invention is entirely of a social nature. I am not thinking now of needs that have arisen from the rapid changes in our society, but of needs that have existed for some time but have not been recognized as such. A familiar example is the electric dry shaver; a more recent one is the cassette recorder for sound or image [1] [*]. The cassette is a device that has contributed enormously to the popularity of sound and image recording. Anyone anywhere can insert a cassette in a tape recorder. This innovation did not call for much scientific knowledge, but rather for a proper appreciation of human clumsiness. Slightly adapting a literary quotation to his purpose, a friend and colleague of mine characterizes devices of this class with the words: "It is not so important to know what nobody has known before, but to do what nobody has done before with what everybody knew".

There are other device innovations that spring from the discovery of how lóng-desired functions can be realized with existing science and technology. This class includes the gyrator (*fig. 1*), whose function was clearly defined in mathematical terms at the beginning of the fifties [2]. Gyromagnetic resonance was also a well-established effect, but the connection between the effect and the function still had to be made. I include electronic shift registers in this same class of innovation, where primary importance attaches to the idea of how to realize the function in principle, even though a great deal of scientific and technological effort is needed later for the actual device. An extremely versatile electronic device is the bucket-brigade delay line with variable delay [3]. The essence of its simplicity of operation lies in the fact that instead of the electric charges representing the analog signal being shifted at a fixed frequency from input to output, a fixed reference voltage is introduced which shifts the charge deficits in the opposite direction. One might be tempted to include in the same category the magnetic-bubble memory devices of Bell Laboratories [4]. This is an interesting example for several reasons. The brilliant and simple idea is to shift information in the form of well-determined locally demonstrable amounts of reverse magnetization within a homogeneous crystal (*fig. 2*). This idea was carried into effect by making use of an effect that had been observed ten years earlier in our laboratories [5]. In this way new functions are fulfilled with what had long been known in the literature. However, the innovation embodied in the magnetic-bubble memory is perhaps not in all respects comparable with that of the gyrator. The magnetic-bubble memory entails too much new technology and comes up against too many unexpected effects.

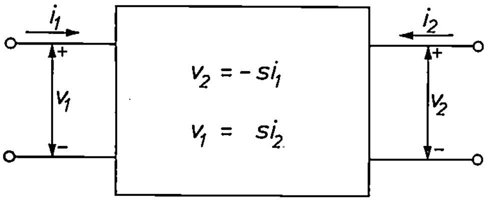


Fig. 1. The ideal gyrator is a network with two pairs of terminals for which the instantaneous values of the input and output currents and voltages, i_1 and v_1, i_2 and v_2, satisfy the equations shown. Since a network of this type does not obey the principle of reciprocity, it cannot be built up from resistors, inductors and capacitors.

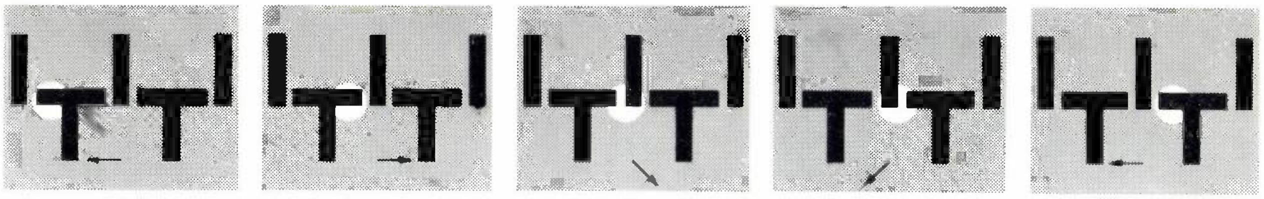
This brings us to what I think is at present the most important field in which we should look for device innovation: the field of technology.

It seems to me that there are two extremes in technological innovation. One extreme leads to an improvement in the performance of the product. The improvement is immediately apparent to the user; he gets a better device. The other, no less important, kind of innovation concerns the improvement of the production method. The improvement in terms of lower production costs or of low rejects will not always be apparent to the user.

In discussing the improvement of device performance we should start by considering how the design can be improved. In this connection we should recall the almost limitless capabilities brought by the computer. It calculates and forecasts the characteristics of highly complicated integrated circuits, for example, thus permitting optimization of the design. The computer also draws the photomasks (*fig. 3*) with better accuracy than can be achieved by hand [6]. In this way the computer gives rise to hardware in future computer generations with an ever-increasing degree of sophistication.

Another interesting application of computer-aided drawing with micron precision is in the manufacture of screens for our beam-indexing colour television tube [7]. To give good colour quality the control circuits must receive a highly constant indexing frequency, which means that the pattern of the index-phosphor — and consequently also the pattern of the three colour-phosphor strips — on the screen must be applied in very accurately determined positions. It is only due to the use of a computer-controlled drawing machine that we are able to generate the patterns with the required tolerances — patterns consisting of thousands of lines a few microns wide, whose position must be fixed to within an accuracy of one micron.

[*] *Literature references are listed at the end of the article.*



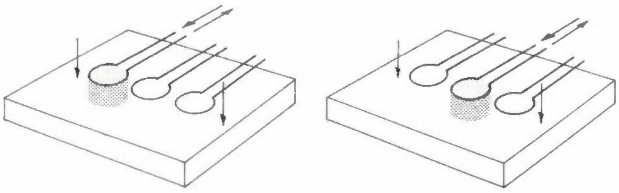


Fig. 2. Cylindrical magnetic domains ("bubbles") contained in a thin layer of a suitable magnetic material can be displaced in several ways. *On the left:* the successive energization of adjacent conductors on the layer; *above:* a relatively weak rotating magnetic field, which is locally concentrated by a system of T- and I-shaped strips of permalloy.

Closely related to the improvement of device performance by computer-aided design is the improvement brought about by the use of high-precision tools in conjunction with materials technology, which makes it possible to achieve extremely high geometrical precision.

In our laboratories lathes with pressurized bearings are in operation which are capable of machining surfaces of optical quality [8]. In this class of innovation the progress being made in electron-beam machining is promising [9]. Ion implantation is also leading to an appreciable improvement of the geometrical definition in semiconducting devices [10].

With regard to the role that materials can play in the improvement of geometrical precision I should like in the first place to mention the photographic materials now coming into use for making photomasks, for example, which have a grain size in the region of 10 nm [11]. The complete Encyclopædia Britannica has been produced on an area of 25 cm^2 of such material, and an easily readable text can be obtained on enlargement of the reproduction.

The recent invention in our laboratories of a method that uses the masking action of thin layers of silicon nitride to create an oxide pattern on silicon is leading to a substantial improvement in the definition of semiconducting devices [12]. We have given this technique the name LOCOS, which means "local oxidation of silicon". In contrast with the conventional planar technique, the oxide patterns are embedded in the silicon substrate (*fig. 4*). This is important not only for MOS circuits, where a thick oxide layer is needed to prevent parasitic effects, but it also makes it possible to replace the usual isolation diffusion in bipolar integrated circuits, in part at least, by oxide. As the thick oxide pattern, made with the aid of only one photomask, is also used to define the various device areas to be made later in the process, the photographic alignment steps are less critical. The use of the LOCOS technique gives a higher packing density and improves integrated-circuit performance. This technique may therefore be expected to come into large-scale use in the semiconductor industry.

It can also be claimed that the introduction of a physical effect, not previously used in this connection, can result in greater geometrical accuracy. This is shown by the example of Integrated Injection Logic, a method recently developed in our laboratories [13], in which power is supplied to integrated circuits by local injection of charge carriers. The power is generated by the incidence of light on built-in phototransistors. For a logic circuit this means that a few transistors have to be added, but no supply leads are used, since only "logic wiring" is needed. Phototransistors are particularly suitable for integrated injection logic because their output current has a natural upper limit. There is therefore no need for current-limiting series resistors, and the heat generation on the chip surface is very much lower. Very high packing densities can be achieved with this power supply method: it is not difficult to achieve 100 logic gates per mm^2 with clearances of 10 microns.

There is no doubt that better control of the chemical composition of materials is of decisive importance for improving the performance of electronic devices. Analytical and preparative chemical methods are steadily

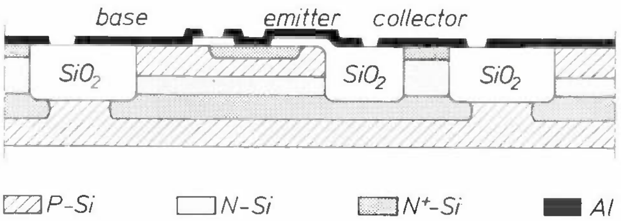
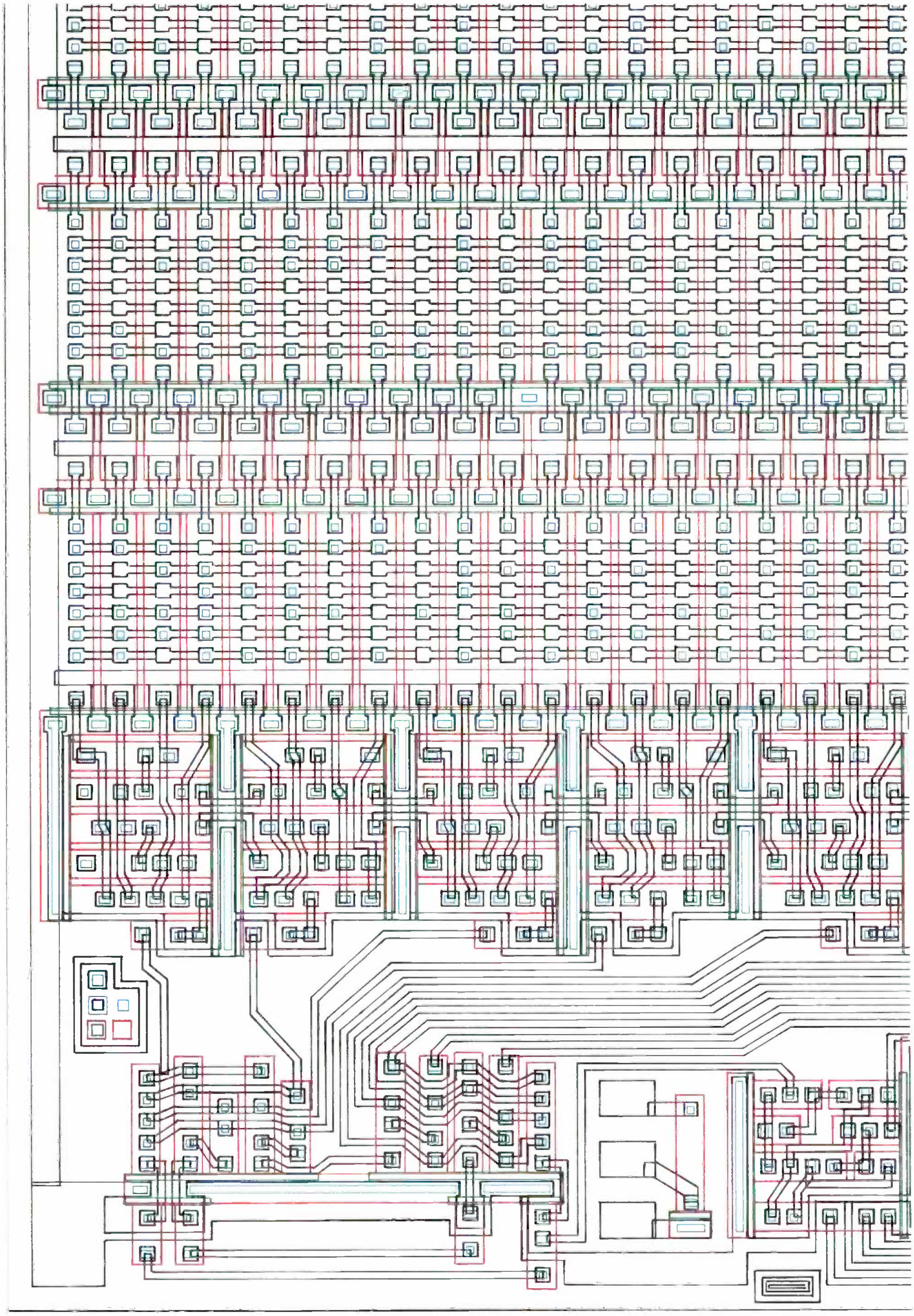


Fig. 4. Cross-section of an N^+PN transistor made in an integrated circuit by the LOCOS technique. The usual isolation diffusion is replaced here by a thick strip of SiO_2 immediately surrounding the P layer. The collector is connected to the contact zone by a buried layer that also passes underneath one of the thick SiO_2 layers.



progressing. In this context we are happy with the vacua of 10^{-11} torr that can now easily be produced in commercial equipment. They allow reproducible vapour deposition on clean surfaces. Important contributions to materials technology are also being made by the better control of the microstructure of materials [14]. By hot pressing, and notably continuous hot pressing of ceramics during sintering, porosity can be almost completely suppressed at a working temperature low enough to avoid decomposition of the material. The resultant material is at the same time homogeneous and of high quality.

After these examples of the impact of technological innovation on product improvement, I should now like to turn to the influence of technology on production improvement. It should be added that the two are often to some extent interdependent.

I have already mentioned the computer as a tool in device design. But the computer also regulates and controls production processes, resulting in ever better control of product quality. It would not surprise me if the refinement and extension of computer-aided production methods proved to be the most important production improvement in the electronic industry in the next decade.

I can never get away from science-fiction ideas about new, revolutionary developments in production techniques. Nature itself confronts us with examples of immensely complicated devices, which are duplicated simply by being placed in a suitable environment. Enzymatic specificity ensures that exact replicas of a given pattern are continuously produced. Incredible as it may seem, the failure rate is almost zero; 1 in 10^{10} would be a reasonable guess.

Although innovation of materials for electronic devices is now often due to better control of the composition, to slight alterations in composition, to better control of the microstructure, in short to better technology, we should not forget that new classes of materials are occasionally brought to light by fundamental research. Quite often representatives of such materials have led a secluded life in textbooks for many years.

In this connection I would point to the class of metallic materials of the type RT_5, where R stands for a rare-earth element and T for a transition metal of the iron group. These materials have been thoroughly studied in several laboratories during recent years. One member of this class, $SmCo_5$, can be made into the most effective permanent-magnet material yet known [15]; see *fig. 5*. This material — and some others related to it — can also absorb enormous quantities of hydrogen at low pressures [16]. These exceptional properties of magnetization and absorption are both bound up with the extraordinary crystal structure of the material.

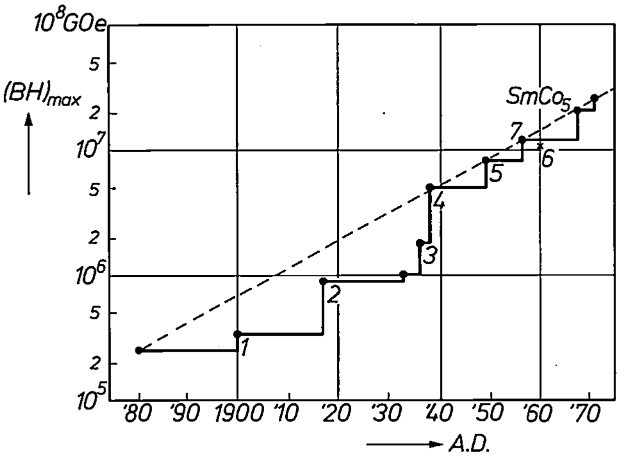


Fig. 5. During the last hundred years the value of $(BH)_{max}$ in materials for permanent magnets has increased by a factor of about 60. The latest upward steps (in 1968 and 1971) were due to the discovery of $SmCo_5$. *1* tungsten steel; *2* cobalt steel; *3* "Ticonal" II [*]; *4* "Ticonal" G; *5* "Ticonal" GG; *6* platinum-cobalt; *7* "Ticonal" XX.

[*] "Ticonal" is a registered trademark.

In this context I should also mention the families of liquid crystals, some of whose members are capable, because of their sensitivity to external voltages, of forming the basis for some modern passive displays [17]. It is interesting, and perhaps significant, that these materials bring liquids into the ranks of electronic materials, other than as glasses or liquid insulators.

An obvious source of innovation I should now like to discuss is to be found in the new device capabilities that arise when materials are used in extreme conditions, for example at very low temperatures, or in extremely high electric or magnetic fields, or when they are subjected to exceptionally short voltage or current pulses. And what do we find when we study the behaviour of materials with extremely sensitive methods of observation?

It is remarkable that there are as yet very few electronic devices whose operation depends on the relevant material being subjected to extreme conditions. This applies not only to consumer products but equally to professional devices. Does this indicate, perhaps, that not enough effort is being made to produce reliable designs that do not make undue demands on the operator's intelligence?

It is noticeable in particular that little use is being made of the remarkable quantum effects that occur in

Fig. 3. Part of a superposed set of masks for a read-only memory containing bipolar transistors. The drawing is automatically made by a drawing machine that receives its information from a computer via punched tape. This part of the circuit measures 1.3 × 1.7 mm.

materials at very low temperatures. Is the perfect diamagnetism of superconducting materials going to be used more extensively than hitherto in electron lenses? And what about the advantages offered by the Josephson effect for measuring changes as small as 10^{-10} gauss in a magnetic field at a level of 1 gauss? Applications for routine inspection work in industry and in medical diagnostics have been proposed. What about the routine use of Josephson detectors for radiation at levels corresponding to less than 10^{-15} watts?

It seems to me that not much use is yet being made in electronic devices of the high electric field-strength available in laser radiation, in spite of its promise of very unusual applications. Just to give an idea I should like to mention the semiconducting photographic plate which is a current topic of research in our laboratories [18]. With a laser beam in which an object is placed and a reference beam a holographic interference pattern is produced in a slice of silicon. If the energy of the light quanta is slightly greater than the energy required to transfer electrons from the valence band to the conduction band of silicon, the optical interference pattern is converted into a corresponding spatial pattern of holes and electrons. Owing to the dispersion of the free charge carriers the refractive index of the silicon slice is modulated spatially in accordance with the interference pattern. If now the reference beam is reflected upon itself by means of a mirror situated behind the silicon slice, it will reconstruct a real image of the object from the hologram (the spatial refractive-index pattern); see *fig. 6*. Because of the rapid diffusion of charge carriers in silicon the lifetime of the charge pattern is only about 20 ns (*fig. 7*).

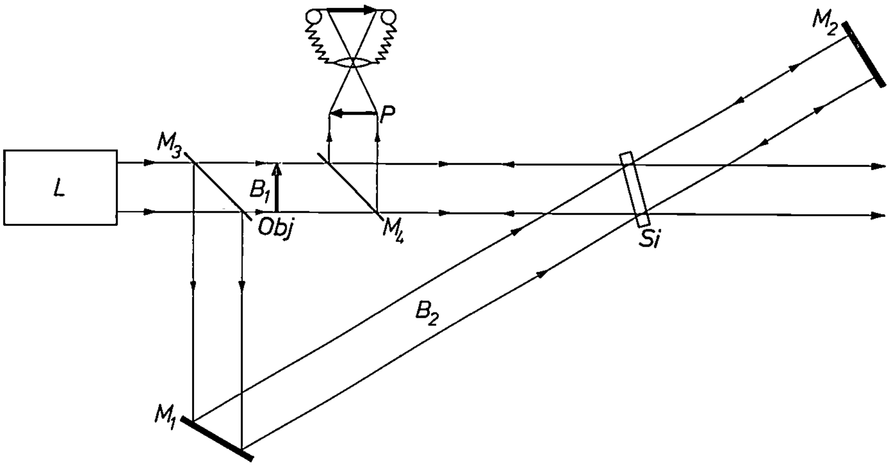


Fig. 6. Holography, in which the hologram consists of a distribution of free charge carriers in a silicon slice. *L* Nd-YAG laser. $M_{1,2}$ mirrors. $M_{3,4}$ semi-transparent mirrors. *Si* silicon slice. B_1 main beam. B_2 reference beam. *Obj* object. *P* real image, formed by part of the light returning from M_2 being diffracted by the hologram in the silicon.

If the arrangement is changed somewhat and if non-continuous laser light is used, it is possible to produce a difference image. Given a transparent object — a "phase object" — which changes rapidly with time, for example a shock wave in a gas, then with the aid of a short pulse of laser light we can obtain an image of the change that takes place in the object between the instant at which the light pulse first passes through the object — the state at this instant is holographically recorded in the silicon slice — and the instant at which the returning pulse passes through the object. If the object has not changed in that time, it then "neutralizes" the holographically produced image.

In the foregoing we have looked at several sources of electronic device innovation, with technology considered as a very important source. It should not be forgotten, however, that there are new physical effects that remain to be discovered. Now that solid-state physics is so far advanced, these new effects will generally be of higher order. In other words, they will usually be small and liable to be disturbed by others. But there are exceptions.

I should like to draw attention to a new type of effect on which our laboratories are working, the photomagnetic effect [19]. This is an effect in which light changes the magnetic properties of materials. When a magnetic material is irradiated with light of wavelength ranging from the infra-red to the X-ray region, the coercive field-strength can be increased by a factor of three. The squareness of the hysteresis loop can also be improved, and a high permeability can be reduced to almost 1. With these effects it is possible in principle to "print" a pattern of magnetic permeability on a substrate. As yet, unfortunately, the effects are only significant at low temperature.

Let me now summarize what has been said in the foregoing about the limitations and prospects of in-

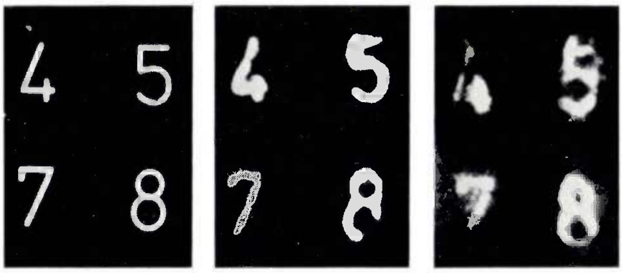


Fig. 7. Owing to the diffusion of the charge carriers a hologram in silicon has a lifetime of only about 20 ns. On the left is the object, and beside it the images reconstructed from the hologram after 3 and 22 ns.

novation in electronic devices. It must be realized that we are talking about a field that is a little past its prime. Even so, there is still room for the intelligent discovery of a forgotten need, for the ingenious realization of a desired function, for new types of materials and for new effects in materials under extreme or even, occasionally, under ambient conditions. And in the next decade there will certainly be room for the improvement of device performance and production by technological refinements.

It looks as if the scope for innovation in electronic devices today is limited not so much by science and technology as by economics, finance and the value which society attaches to the innovations. Production capacity in the world has of course been stretched considerably. It requires quite a bit of entrepreneurial courage to invest perhaps a hundred million dollars in the mass production of a new electronic device if the market is highly competitive and approaching saturation. This is particularly dangerous when interest rates are not far below 10%. It is also risky to develop small series of very advanced components. Unless the risk is removed by a firm order, e.g. a government contract, there is the risk that the prices are likely to be quickly eroded. Although the situation differs from one country to another, it is true to say that it requires an ever increasing knowledge of markets and growing economic and financial skills to bring out technological innovations, including innovations in electronic devices.

In these circumstances the electronic engineer and the scientist seeking outlets for their creativity will do well to remember that it is often difficult to gain acceptance for improvements of existing technical solutions, particularly if the existing solutions have proved satisfactory. "Unfortunately", we are tempted to say, many solutions in the field of electronics are nearly perfect already.

When I say that one should be careful in trying to improve satisfactory subsystems, I am not making a purely negative observation. Let us consider the aircraft that brought me from Amsterdam to Kennedy Airport. As far as I am concerned it does not have to fly faster or offer more comfort. What I would like to see improved is not the ease of transport between the Amsterdam and New York runways, but from my house near Eindhoven to this conference room in Washington. This involves so much additional time in changing from one means of transport to another, in queuing, walking and waiting, that suboptimization of the transatlantic flight is not very helpful. Many believe that it would be more appropriate to replan and reconstruct the whole transport system, town planning included. The moral of this story is that the device engineer should consider very carefully the functions the devices fulfil in the system as a whole.

In our densely populated world there are many new functions connected with communication, traffic, education and health that cry out to be fulfilled. They certainly require new devices, especially developed for the purpose. The only electronic example I would mention in this context is the SO_2 monitor developed in our laboratories. Monitors of this type have been set up at scattered points in the industrial regions of the Netherlands and pass continuous information on the state of air pollution to a central processor, and they do this job unattended and without maintenance for months on end [20]. This enables an alert to be put out to industrial plants during periods of serious air pollution, warning them to take appropriate action. Similar monitor/computer systems are now being installed in several countries. Continuously operating automated systems for monitoring other air pollutants than SO_2, and systems for monitoring water pollution, are under development in our laboratories.

I believe it is a good thing that the social relevance of systems in which electronic devices are to be used is now becoming a major issue, at least in the "rich parts" of the world. I would therefore like to say a few words about the most serious of the social problems confronting us today.

It has been pointed out that the social needs arising from overpopulation, from pollution of many kinds (including heat), from the exhaustion of mineral resources, etc., will enforce drastic cutbacks of economic growth in the foreseeable future. Studies made in various countries [21] indicate that if these matters are not taken into account humanity will be heading for universal disaster within the next hundred years or so. It is obvious that international agreements are necessary if effective measures are to be imposed. It will take a great deal of time, too much time perhaps, to arrive at such agreements. It is probable that well before then the requirements imposed on our technology, including the standards to be met by electronic devices, will become much more stringent. Greater reliability and longer life will be required, rules will be laid down governing the consumption and recovery of scarce materials, the economic use of energy in the production of devices and in their use, and the drastic abatement of noise.

In this context I should like in conclusion to touch on the anti-technical, anti-scientific, and perhaps counter-cultural or even anti-cultural attitude nowadays encountered among intelligent young people. It is of course true that technological progress has created the conditions for ever-increasing population growth and individual consumption. This has given rise to undesirable situations, which we are now having to cope with. Obviously this does not disqualify science and technology, but it does reflect our present lack of knowledge about the behaviour of "systems" that include people, and our political ability to give direction to them. It is of course difficult in the free world to fix priorities, but it seems to me, and I find it alarming, that there is not enough critical discussion either in the newspapers, in governments or in the universities of these current anti-technological and anti-scientific views. Could it be that the scientific community is paralysed by a bad conscience because of the actual or potential misuse of its inventions? In the years ahead we shall be faced with growing overcrowding, and there will be immense problems to be solved if we are to reach a state of equilibrium without major disasters. In this respect continuing scientific analysis and technological measures are the principal means at our disposal.

I believe that none of us should be left untouched by the immensity of the problems. Too much is at stake to allow any of us to hide behind isolated scientific and technological problems. We are all involved in the course events will take. I think we should all take to heart what Arnold Toynbee says in his "Study of History" [22]: "The nature of the breakdown of civilizations can be summed up in three points: a failure of creative power in the minority, an answering withdrawal of mimesis on the part of the majority and a consequent loss of social unity in the society as a whole" [*].

[*] Terms such as *creative minority*, *majority* and *mimesis* are used by Arnold Toynbee in a rather special way to indicate concepts that he has defined in his writings. *(Ed.)*

Bibliography

[1] See for example P. van der Lely and G. Missriegler, Audio tape cassettes, Philips tech. Rev. **31**, 77-92, 1970.

[2] B. D. H. Tellegen, Philips Res. Repts. **3**, 81, 1948. See also Philips tech. Rev. **18**, 120, 1956/57.

[3] F. L. J. Sangster, The "bucket-brigade delay line", a shift register for analogue signals, Philips tech. Rev. **31**, 97-110, 1970.

[4] A. H. Bobeck, R. F. Fischer, A. J. Perneski, J. P. Remeika and L. G. Van Uitert, IEEE Trans. **MAG-5**, 544, 1969.

[5] C. Kooy and U. Enz, Philips Res. Repts. **15**, 7, 1960. See also C. Kooy, Philips tech. Rev. **19**, 286, 1957/58.

[6] See for example C. Niessen and H. E. J. Wulms, Philips tech. Rev. **30**, 29, 1969.

[7] For a description of the beam-indexing tube, see E. F. de Haan and K. R. U. Weimer, Roy. Telev. Soc. J. **11**, 278, 1967/68, G. J. Lubben, Onde électr. **48**, 918, 1968 and P. M. van den Avoort, Onde électr. **48**, 921, 1968. The idea behind the tube is briefly described on pages 71 and 72 of Philips tech. Rev. **32**, 1971 (No. 3/4). A forthcoming article in this journal will be entirely devoted to the beam-indexing tube.

[8] H. J. J. Kraakman and J. G. C. de Gast, Philips tech. Rev. **30**, 117, 1969. See also Philips tech. Rev. **31**, 126, 1970.

[9] See for example C. van Osenbruggen, Philips tech. Rev. **30**, 195, 1969.

[10] See for example J. M. Shannon, Philips tech. Rev. **31**, 267, 1970.

[11] See for example H. Jonker, L. K. H. van Beek, H. J. Houtman, F. T. Klostermann and E. J. Spiertz, J. photogr. Sci. **19**, 187, 1971 (No. 6). A survey article about the investigations at Philips Research Laboratories on such unconventional materials is planned for the next volume of Philips Technical Review.

[12] J. A. Appels, H. Kalter and E. Kooi, Some problems of MOS technology, Philips tech. Rev. **31**, 225-236, 1970.
LOCOS technology, Philips tech. Rev. **31**, 276, 1970.
J. A. Appels, E. Kooi, M. M. Paffen, J. J. H. Schatorjé and W. H. C. G. Verkuylen, Local oxidation of silicon and its application in semiconductor-device technology, Philips Res. Repts. **25**, 118-132, 1970.
J. A. Appels and M. M. Paffen, Local oxidation of silicon; new technological aspects, Philips Res. Repts. **26**, 157-165, 1971 (No. 3).
E. Kooi, J. G. van Lierop, W. H. C. G. Verkuylen and R. de Werdt, LOCOS devices, Philips Res. Repts. **26**, 166-180, 1971 (No. 3).

[13] This subject will be dealt with in a forthcoming article in Philips Technical Review.

[14] For a general review see G. H. Jonker and A. L. Stuijts, Philips tech. Rev. **32**, 79, 1971 (No. 3/4).

[15] K. H. J. Buschow, W. Luiten, P. A. Naastepad and F. F. Westendorp, Philips tech. Rev. **29**, 336, 1968.

[16] J. H. N. van Vucht, F. A. Kuijpers and H. A. C. M. Bruning, Reversible room-temperature absorption of large quantities of hydrogen by intermetallic compounds, Philips Res. Repts. **25**, 133-140, 1970.

[17] A survey is given in: A. Saupe, Angew. Chemie **80**, 99, 1968, and also in: E. A. Kosterin and I. G. Chistyakov, Soviet Phys. Crystall. **13**, 229, 1968.

[18] J. P. Woerdman, Formation of a transient free carrier hologram in Si, Optics Comm. **2**, 212-214, 1970.

[19] See for example U. Enz and R. W. Teale, Photomagnetic effects, Philips tech. Rev. **31**, 33-39, 1970.

[20] H. J. Brouwer, S. M. de Veer and H. Zeedijk, The SO_2 monitoring network in the Rhine estuary region, Philips tech. Rev. **32**, 33-41, 1971 (No. 2).

[21] See for example Jay W. Forrester, World dynamics, Wright-Allen Press, Cambridge, Mass., U.S.A., 1971.

[22] Arnold J. Toynbee, A study of history, abridged version by D. C. Somervell, Oxford University Press, New York 1947, page 578.

Fast phosphors for colour television

A. Bril, G. Blasse, A. H. Gomes de Mesquita and J. A. de Pôorter

Two types of picture tube used in colour television, the cathode-ray tube for the flying-spot scanner and the beam-indexing tube, both require phosphors whose luminescence decays very rapidly in intensity when the excitation is removed. In this respect, the phosphors previously available were rather less than ideal. Investigations of compounds containing ions of rare-earth metals have led in recent years to the discovery of a number of very fast phosphors, all containing the Ce^{3+} ion as an activator, which considerably improve the performance of these tubes.

Introduction

Colour television is even more dependent than monochrome television on the availability of substances that luminesce when bombarded by electrons; these substances, known as phosphors, have to meet specific requirements for use in television. This applies not only to those phosphors with whose action the viewer is directly confronted — the colour phosphors of the screen, whose emission spectrum is of prime importance — but equally to phosphors whose action is not immediately apparent to the viewer. These latter phosphors, with which we are concerned in this article, have to meet a very exacting combination of requirements, in which the most important is that they should be "fast".

When the excitation of a phosphor is removed, the intensity I of the luminescent radiation in many cases decreases exponentially with time t:

$$I = I_0 \, e^{-t/\tau} .$$

In this expression τ is the decay time of the phosphor, i.e. the time it takes for the intensity of the luminescent radiation to decrease to 1/e of its initial value I_0 when the excitation is removed.

Different phosphors have very widely different values of τ. For example, Mn-activated ZnF_2 has a decay time of about 0.1 s, whereas some organic phosphors have decay times of the order of 10^{-8} s, i.e. as much as 10^7 times shorter.

An effect found with many phosphors, which is often a nuisance in television, is the *afterglow* or *persistence*. When the excitation of these phosphors is removed, the intensity of their luminescence decays at first in the manner just described, but when a particular intensity is reached — which may sometimes be only a few per cent of the original value — the decay proceeds much more slowly.

In our work on fast phosphors the measure used for the afterglow was the fraction δ of the initial luminescence intensity that is left 80 μs after the excitation has been removed. All the phosphors were excited with an electron beam, and the excitation time in all cases was 20 μs. We call the fraction δ the *afterglow level.*

Afterglow occurs because the crystal lattice contains other impurities or imperfections than the ones responsible for the luminescence. These can act as electron traps, and introduce additional energy levels into the energy-band diagram, in which excited electrons are able to remain for some time before returning to the ground state. These then give rise to luminescence long after the excitation has been removed.

The suitability of a phosphor for a particular application depends not only on the emission spectrum, the decay time and the afterglow level, but also of course on the radiant efficiency η, i.e. the ratio of the radiant power emitted to the power supplied by the electron beam.

In this article we shall discuss a number of recently discovered phosphors that are used in two types of cathode-ray tube for which it is absolutely essential to have a short decay time (less than 100 ns) combined with the highest possible radiant efficiency.

The first application relates to the luminescent screen

Dr. A. Bril and J. A. de Poorter are with Philips Research Laboratories, Eindhoven; Prof. Dr. G. Blasse, formerly with Philips Research Laboratories, Eindhoven, is now Professor of Solid-State Chemistry at the University of Utrecht; Dr. A. H. Gomes de Mesquita, formerly with Philips Research Laboratories, Eindhoven, is now with the Philips Information Systems and Automation Department (ISA).

in the cathode-ray tube of a flying-spot scanner for colour television, which generates television signals for the transmission of cine films and slides. For this purpose the spectrum of the luminescence must cover the whole visible range.

The second application relates to the phosphor that produces the synchronizing signal for the electron beam in a beam-indexing tube, again for colour television. This phosphor is referred to as the index phosphor. The luminescence here must lie in the ultraviolet, and the afterglow level must be as low as possible. In these two types of tube it is not possible to use fast organic phosphors, because they would decompose with the increase in temperature that occurs in the tube-manufacturing process.

The fast phosphors we have discovered belong to a group of yttrium compounds doped with cerium (1 to 2 atomic %) whose luminescence is caused by an electron transition in a Ce^{3+} ion that behaves as an activator (this is known as characteristic luminescence) [1]. It was found some twenty years ago that Ce^{3+}-activated phosphors are characterized by a short decay time and a high efficiency [2]. That was why a phosphor of this type was chosen for the tube of the flying-spot scanner for monochrome television. As the emission spectrum of that phosphor has its maximum in the ultra-violet and blue part of the spectrum, however, it was not possible to use it in flying-spot scanners for colour television. Nor could it be used in the beam-indexing tube, since its afterglow level was too high. For the flying-spot scanner two phosphors have now been found whose combined emission spectra cover almost the entire visible range when they are used together, and which also meet the other requirements extremely well.

Before discussing the new phosphors and their application in colour television, we shall first take a look at the mechanism of the luminescence in Ce^{3+}-activated phosphors.

Luminescence of Ce^{3+}-activated phosphors

The element Ce, used to activate the phosphors discussed here, appears in the periodic system as one of the rare-earth metals. The characteristic feature of these elements is that the 4f group of the N shell is not completely filled with electrons, whereas the 5s and 5p groups of the O shell are. In the Ce atom the 4f group contains two electrons, and in the trivalent positively charged Ce^{3+} ion the 4f group contains one. The two other electrons that are not present in the ion are the 6s electrons; see *fig. 1.*

	1s	2s	2p	3s	3p	3d	4s	4p	4d	4f	5s	5p	5d	5f	6s
Ce	2	2	6	2	6	10	2	6	10	2	2	6	—	—	2
Ce^{3+}	2	2	6	2	6	10	2	6	10	1	2	6	—	—	—

Fig. 1. Electron configuration of the Ce atom and the Ce^{3+} ion. In the rare-earth metals, which include Ce, the 4f levels are not all filled, but the outer-lying 5s and 5p levels are.

This single 4f electron gives rise to two energy states: the $^2F_{7/2}$ level (with the quantum numbers $S = 1/2$, $L = 3$, $J = 7/2$) and the $^2F_{5/2}$ level ($S = 1/2$, $L = 3$, $J = 5/2$). These two levels, which differ in energy by an amount corresponding to about 2000 cm^{-1}, relate to two distinct states: in the one the orbital and spin moments of the electron are parallel and in the other they are anti-parallel. The 4f electron is screened from the environment by the electrons in the O shell.

The Ce^{3+} phosphors emit radiation when an electron that has been raised by excitation from a 4f to a 5d state returns to the 4f state. Unlike the 4f orbit, the 5d orbit lies at the surface of the ion and is therefore much more exposed to the influence of the crystal lattice. The electric field of the surrounding ions, the "crystal field", has the effect of splitting the 5d level into a number of sublevels (the Stark effect). Moreover the emission and absorption bands are considerably broadened due to the interaction with the lattice, as discussed earlier [3]. Transitions from this level to the 4f level give rise to a band instead of discrete lines in the emission spectrum. The splitting of the 4f state gives the emission a doublet character, i.e. the band consists of two overlapping sub-bands.

The strongest emission corresponds to a transition from the lowest sublevel of the 5d level to the two 4f levels. In general the emission of Ce^{3+} phosphors lies in the ultra-violet and the blue. However, if the crystal-field splitting of the 5d level is very considerable, radiation of longer wavelength may be emitted. This is the case, for example, with one of the flying-spot-scanner phosphors which we shall presently discuss, and whose optical transitions are schematically represented in the energy-level diagram of *fig. 2.* The crystal-field splitting of the 5d level in this phosphor is so great that the colour of the emission from the lowest excited level is yellow-green [4]. Since emission can also occur from a higher sublevel emission, bands are found in the ultra-violet as well.

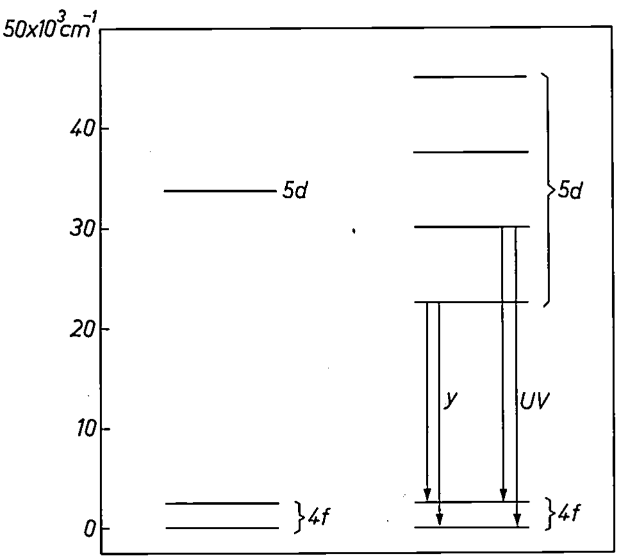


Fig. 2. Example of how the crystal field can affect the energy-level diagram and hence the emission of Ce^{3+} ions. *Left:* the two 4f levels and the "centre of gravity" of the 5d level of the unperturbed ion. Transitions 5d-4f correspond here to ultra-violet emission. *Right:* the Ce^{3+} levels in one of the phosphors used in the flying-spot scanner for colour television. The crystal field causes such strong splitting of the 5d level that the transitions from the lowest-lying sublevel to the 4f levels (arrows *y*) correspond to yellow-green emission. The UV radiation also emitted by this phosphor is due to the transitions indicated by the arrows *UV*.

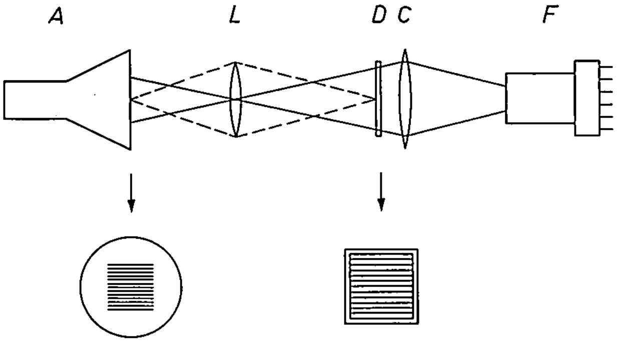


Fig. 3. Principle of a flying-spot scanner for monochrome television. On the screen of tube *A* the electron beam describes a line raster which is projected by means of the lens *L* on to the transparency *D* (slide or film). The condenser lens *C* collects the light transmitted by *D* and directs it on to the photocathode of the photomultiplier tube *F*. For colour television the light transmitted by *D* is split into red, green and blue components, each of which are picked up by a separate photomultiplier tube.

In the trivalent ions of the rare-earth metals Eu and Tb, also widely used in phosphors, the luminescence depends on forbidden transitions [3]. In the 5d-4f transition of the Ce^{3+} ion mentioned earlier, however, the situation is quite different: the transition is a permitted one (electric-dipole transition). It is this that accounts for the very short lifetimes of the electrons in the 5d states — 30 to 100 ns in the lowest 5d level — and the luminescence decays in a similar short time after the excitation has been removed.

Phosphors for the flying-spot scanner

We shall first discuss the Ce^{3+} phosphors for the tube in the flying-spot scanner. The operation of a flying-spot scanner for monochrome television is recalled in *fig. 3* [5]. A raster pattern of constant intensity is traced out by an electron beam on the phosphor-coated flat screen of a cathode-ray tube *A*. This raster pattern (or "field") is projected on to the transparent object *D* (cine film or slide), which is thus completely scanned. The transmitted light, modulated in intensity by the film or slide, is directed on to the photocathode of a photomultiplier tube *F*, which delivers an output current proportional to the intensity of the transmitted light. In the flying-spot scanner for colour television [6] the light transmitted by the film or slide is separated by means of dichroic mirrors and lenses into red, green and blue components, for each of which there is a separate photomultiplier tube.

To reduce blurring as much as possible the decay time τ of the phosphor [7] must be at most of the same order of magnitude as the time taken by the electron beam to scan a picture element of diameter equal to the cross-section of the beam itself. It can be deduced from this that τ should then not be much greater than about 50 ns.

In addition, the ratio $\sqrt{(\eta/\tau)}$ should be high to give a good signal-to-noise ratio. Hence, in addition to a small value of τ a high *radiant efficiency* η is required.

As we noted earlier, the *emission spectrum* of the phosphor in a flying-spot scanner for colour-television signals is also required to cover the whole visible range.

In the past, green-luminescent ZnO has generally been used. Since the emission band of this phosphor is very broad it can also be used to provide red and blue signals. The decay time of this phosphor, however, is

[1] A brief discussion of characteristic luminescence and a review of developments in this field during recent years will be found in: G. Blasse and A. Bril, Characteristic luminescence, Philips tech. Rev. **31**, 303-332, 1970.

[2] A. Bril and H. A. Klasens, Intrinsic efficiencies of phosphors under cathode-ray excitation, Philips Res. Repts. 7, 401-420, 1952.
A. Bril and H. A. Klasens, New phosphors for flying-spot cathode-ray tubes, Philips Res. Repts. 7, 421-431, 1952. (See also Philips tech. Rev. **15**, 63, 1953/54.)
A. L. Smith, Trans. Electrochem. Soc. **96**, 287, 1949.

[3] More about this will be found in the article of note [1] on pages 315 and 316.

[4] This is also the case with the recently discovered Ce^{3+}-activated phosphor (Ca,Ce)S. See W. Lehmann, J. Electrochem. Soc. **118**, 1164, 1971 (No. 7).

[5] A more detailed description is given in: F. H. J. van der Poel and J. J. P. Valeton, The flying-spot scanner, Philips tech. Rev. **15**, 221-232, 1953/54.

[6] See also H. van Ginkel, Flying-spot scanners for colour televisicn, Philips tech. Rev. **21**, 234-250, 1959/60.

[7] The decay time has been referred to as the afterglow time in earlier articles [5] [6]. We have chosen the term decay time to distinguish it from the term "afterglow" which relates to the weak radiation that often persists for quite a long time and decreascs very slowly.

about 1 µs, which is too long. Although the effect of an excessively long decay time can be compensated by means of an electronic circuit [5] [6], this can only be done at the expense of the signal-to-noise ratio.

The more important of the two new fast Ce^{3+} phosphors for the flying-spot scanner [8] is the garnet $(Y,Ce)_3Al_5O_{12}$.

The decay time of its luminescence is 70 ns, which is a considerable improvement on the ZnO phosphor. In addition it has a somewhat higher radiant efficiency. The part of the emission spectrum of the new phosphor that is of interest for the flying-spot scanner (*fig. 4*) covers a large part of the visible range; it extends from about 450 nm to beyond 700 nm, and has its maximum at 550 nm. The spectrum contains much more red radiation than the spectrum of the ZnO phosphor. This is also an improvement since the sensitivity of photomultiplier tubes generally shows a marked decrease towards longer wavelengths, as can also be seen from fig. 4. The complete spectrum is given in *fig. 5*.

As there is no blue in the emission spectrum of $(Y,Ce)_3Al_5O_{12}$, this phosphor must be combined with a blue-luminescent phosphor for use in the flying-spot scanner, and this phosphor should of course also have a short decay time. A compound from the Y_2O_3-SiO_2 system, activated with 2% Ce [9] [10], is particularly suitable. This new phosphor is $(Y,Ce)_2SiO_5$; its emission spectrum, shown in *fig. 6*, has a maximum at 415 nm. The decay time is only 30 ns and the efficiency is 6%.

Table I presents the radiant efficiency, the decay time, the afterglow level and the position of the maximum in the emission spectrum λ_{max} for the two Ce^{3+} phosphors, in comparison with the values for the ZnO phosphor previously used. Tubes with a luminescent screen formed from a combination of $(Y,Ce)_3Al_5O_{12}$ and $(Y,Ce)_2SiO_5$ have been on the market for some time (Philips Q13-110 GU) and are now in successful practical use. The electronic circuit needed to compensate the long decay time of the ZnO phosphor is not required with these tubes. A very much better signal-to-noise ratio can therefore be achieved than with the ZnO tubes. Electronic compensation of the slight afterglow of the $(Y,Ce)_3Al_5O_{12}$ phosphor can be applied if desired, but does not cause any perceptible deterioration of the signal-to-noise ratio.

Phosphors for the beam-indexing tube

Before discussing the fast Ce^{3+} phosphors that have been developed for the beam-indexing colour-television tube, we shall briefly describe the construction of this tube, which is part of a display system as yet in an experimental stage [11].

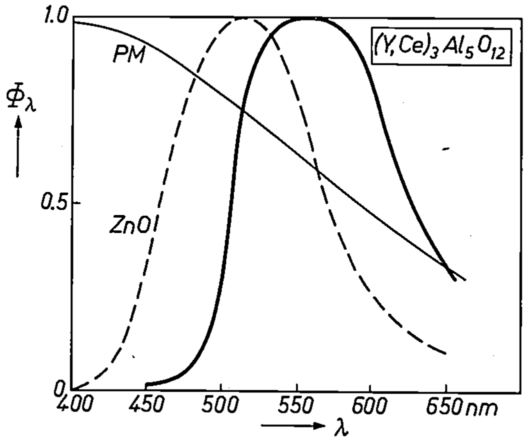


Fig. 4. The useful part of the emission spectrum of the new, fast, yellow-green emitting phosphor $(Y,Ce)_3Al_5O_{12}$ for application in the cathode-ray tube of a flying-spot scanner for colour television. The relative spectral radiant power Φ_λ, as found on excitation with an electron beam, is shown as a function of wavelength λ. The weak UV emission is not included in this figure (see fig. 5). The emission spectrum of the green-luminescent ZnO formerly used is also shown for comparison (dashed curve). The straight line *PM* gives the spectral sensitivity of a type S20 photomultiplier tube (in arbitrary units). The maxima in the emission spectra are made equal to unity.

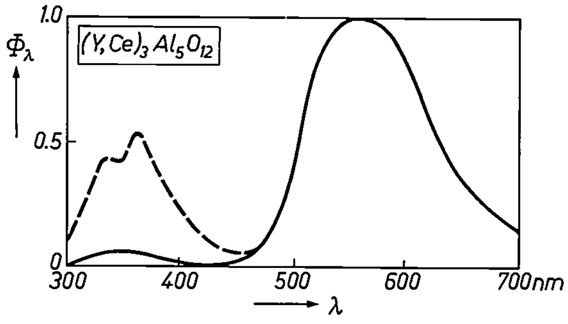


Fig. 5. The complete emission spectrum of $(Y,Ce)_3Al_5O_{12}$. The solid curve represents the emission on excitation with an electron beam. The dashed curve is the spectrum on excitation with ultraviolet radiation of 253.7 nm wavelength. The dashed curve and the solid curve coincide in the visible region. Two emission bands can clearly be distinguished, corresponding to the energy-level diagram of fig. 2. In addition to the yellow-green emission (fig. 4) there is also an ultra-violet emission band. The splitting of the UV band is a consequence of the doublet character of the 4f level. In the yellow-green band this also occurs if the temperature is lower than about 100 K. The maximum in the spectrum is again made equal to unity.

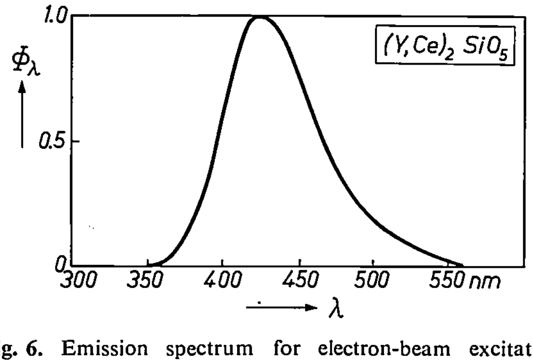


Fig. 6. Emission spectrum for electron-beam excitation of $Y_{1.98}Ce_{0.02}SiO_5$. This blue-luminescent phosphor is used in combination with $(Y,Ce)_3Al_5O_{12}$ in the cathode-ray tube of the flying-spot scanner for colour television, since $(Y,Ce)_3Al_5O_{12}$ emits relatively little blue.

Table I. Radiant efficiency η, decay time τ, afterglow level δ (80 µs after excitation is removed), and the wavelength $\lambda_{\max}$ at which the emission is at a maximum of new fast phosphors for use in the cathode-ray tube of the flying-spot scanner for colour television. For comparison values are also quoted for the ZnO phosphor previously used. All values were measured for excitation with an electron beam.

Phosphor	η (%)	τ (ns)	δ (%)	$\lambda_{\max}$ (nm)
$(Y,Ce)_3Al_5O_{12}$	4	70	6	550
$(Y,Ce)_2SiO_5$	6	30	0.1	415
ZnO	2.5	~1000	—	505

The main difference from the conventional shadow-mask tube is that the red, green and blue phosphor elements in the indexing tube are *consecutively* excited by an electron beam from a *single* electron gun, whereas in the shadow-mask tube the phosphor dots are excited *simultaneously* by three separate electron guns.

The screen of the beam-indexing tube consists of 400 "triplets" of vertical, finely-spaced phosphor strips. In each triplet one strip luminesces red, another green and the third blue. All the strips are coated on the inside with a thin aluminium foil. The screen is scanned by the electron beam in a horizontal line raster. To ensure that the electron gun is controlled at every instant by the signal for the appropriate colour, a control signal, the beam-indexing signal, is generated from the "index strips" and applied to the electron gun via an electronic circuit. The index strips are phosphor strips that emit ultra-violet radiation and are arranged at a regular distance apart in the space between the colour strips at the back of the aluminium foil; they thus radiate backwards. Every time the electron beam passes an index strip an ultra-violet flash is emitted. This is transmitted through a window in the tube wall to a photomultiplier tube which is situated outside the display tube and produces the indexing signal. To make sure that this signal is not too weak in the very dark parts of the picture, the display system is designed so that the beam current can never fall below a certain minimum, called the black-level current.

As we have said, the Ce^{3+} phosphors we have developed are intended for the index strips. Index phosphors have to be extremely fast because clearly separated beam-indexing signals can only be obtained if the decay time is less than the time taken by the electron beam to pass an index strip. This time is 80 ns for the maximum permissible diameter of the spot and 20 ns for an infinitely sharp spot. A typical value would be 50 ns.

An index phosphor must not have too long an *afterglow*, because if the afterglow level is too high the photomultiplier tube continuously receives radiation from the whole screen, giving a fairly strong, constant photocurrent. The noise associated with this may be so strong as to swamp the beam-indexing signal. For this reason it is not possible to compensate for this background.

The phosphor must also have a high efficiency, because an indexing signal of sufficient strength must also be present in the dark parts of the picture, where the electron beam exciting the phosphors has its minimum value. If the efficiency of the index phosphor were low, it would be necessary to use so high a black-level current that the colour phosphors would be too strongly excited, to the detriment of the picture contrast.

The *emission spectrum* of the index phosphor must lie in the wavelength region between about 340 and 400 nm, for the following reasons. In the first place the radiation must be able to pass through the glass of the display tube in order to reach the external photomultiplier tube. The wavelength of the radiation must therefore be longer than the ultra-violet absorption edge of the glass, which lies at about 340 nm. On the other hand it is desirable that the wavelength of the radiation should be shorter than about 400 nm — i.e. not visible — because the thin aluminium layer between the colour phosphors and the index phosphor is not entirely light-tight. In addition there is an optical filter in front of the photomultiplier tube, which only passes radiation with a wavelength below 400 nm; this filter ensures that any radiation from the colour phosphors passing through the Al to the photomultiplier tube does not give rise to a photocurrent.

Until recently the phosphors used for the beam-indexing signal were the Ce^{3+}-activated compounds $Ca_2Al_2SiO_7$ (gehlenite) and $Ca_2MgSi_2O_7$ (akermanite). These phosphors have decay times of 50 and 80 ns and efficiencies of 4.5 and 4% respectively. The emission spectrum of $(Ca,Ce)_2Al_2SiO_7$ lies in the ultra-violet and the blue, with a maximum at 405 nm, while the spec-

[8] G. Blasse and 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.
G. Blasse and A. Bril, Investigation of some Ce^{3+}-activated phosphors, J. chem. Phys. **47**, 5139-5145, 1967.
G. Blasse, Influence of crystal structure on luminescence, Mat. Res. Bull. 3, 807-815, 1968.

[9] A. H. Gomes de Mesquita and A. Bril, Preparation and cathodoluminescence of Ce^{3+}-activated yttrium silicates and some isostructural compounds, Mat. Res. Bull. **4**, 643-650, 1969.

[10] A. H. Gomes de Mesquita and A. Bril, The afterglow of some old and new Ce^{3+}-activated phosphors, J. Electrochem. Soc. **116**, 871, 1969.

[11] For details see:
E. F. de Haan and K. R. U. Weimer, The beam-indexing colour television display tube, Roy. Telev. Soc. J. **11**, 278-282, 1967/68;
G. J. Lubben, Le tube index, Onde électr. **48**, 918-920, 1968;
P. M. van den Avoort, Traitement des signaux dans le tube index, Onde électr. **48**, 921-924, 1968.

trum of $(Ca,Ce)_2MgSi_2O_7$ lies at shorter wavelengths, with a maximum at 370 nm.

The afterglow of both phosphors, however, is too long. The afterglow level δ of gehlenite is between 5 and 10%, and of akermanite about 3%.

The Ce^{3+} phosphors that we have discovered have a much shorter afterglow [10] [12]. These phosphors, which are found in the Y_2O_3-SiO_2 system mentioned above, are the Ce^{3+}-activated compounds β-$Y_2Si_2O_7$ and γ-$Y_2Si_2O_7$ [13].

The emission spectra of these phosphors, which differ very little from each other, are shown in *fig. 7a*. They extend from 350 to 500 nm. The blue light present in this spectrum is so weak that it has very little effect on the picture if light should pass through the aluminium foil. The decay time τ of both phosphors is 40 ns, the value of δ is only 0.1% and the efficiencies are 8% and 6.5% respectively. Thus, these two phosphors not only have a much lower afterglow level than those previously used but have in addition a shorter decay time and a higher efficiency.

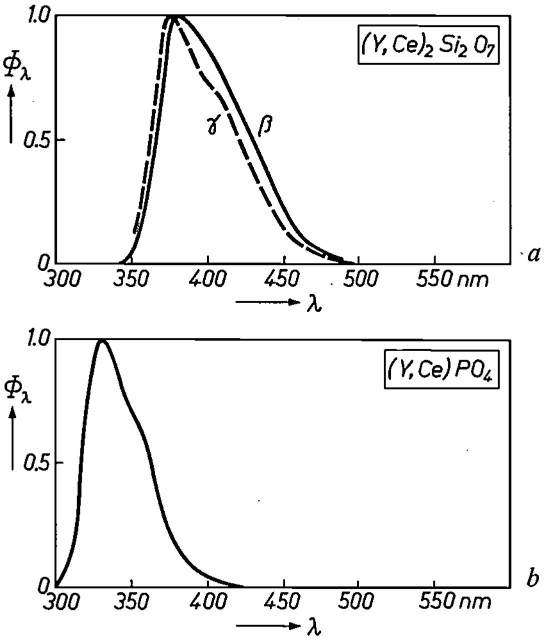


Fig. 7. Emission spectra for cathode-ray excitation of ultraviolet-luminescent phosphors for use as index phosphors in the beam-indexing tube. *a*) The spectra of β- and of γ-$(Y,Ce)_2Si_2O_7$ [13]. *b*) The spectrum of $(Y,Ce)PO_4$ [14].

The principal data relating to these two phosphors are presented in *Table II*. For comparison the table also gives the data for the two index phosphors previously used, together with data for a fast Ce^{3+} phosphor, $(Y,Ce)PO_4$, recently investigated elsewhere [14]. This phosphor also has a lower afterglow level than the phosphors previously used, though it is not so low as that of the other two new phosphors. Its decay time is very short, but its efficiency is lower than that of the other phosphors. Moreover the wavelength of its emission is shorter (fig. 7*b*) and therefore part of the emitted radiation is absorbed by the glass tube. On the other hand there is not much blue radiation in the spectrum, which again is an advantage.

Table II. As Table I, but now for the new phosphors for the beam-indexing tube, β- and γ-$(Y,Ce)_2Si_2O_7$ [13]. For comparison values are given for the phosphors previously used and also for the phosphor $(Y,Ce)PO_4$ [14].

Phosphor	η (%)	τ (ns)	δ (%)	λ_{max} (nm)
β-$(Y,Ce)_2Si_2O_7$	8	40	0.1	380
γ-$(Y,Ce)_2Si_2O_7$	6.5	40	0.1	375
$(Y,Ce)PO_4$	2.5	25	1.5	330
$(Ca,Ce)_2Al_2SiO_7$	4.5	50	5-10	405
$(Ca,Ce)_2MgSi_2O_7$	4	80	3	370

Since only time will show how much importance should be attached to each property — decay time, efficiency, afterglow level and spectrum — it is too early yet to say which of the phosphors is to be preferred for the beam-indexing tube.

The afterglow level and decay time, however, of the new index phosphors are much better than for the phosphors that have previously been used.

[12] A. Bril, G. Blasse and J. A. de Poorter, Fast-decay phosphors, J. Electrochem. Soc. **117**, 346-348, 1970.

[13] The existence of an earlier reported compound α-$Y_2Si_2O_7$ is doubtful. It may possibly be identical with Y_2SiO_5. The structure of β-$Y_2Si_2O_7$ is the same as that of thortveitite $Sc_2Si_2O_7$, the structure of γ-$Y_2Si_2O_7$ is unknown. There is also a δ-modification of $Y_2Si_2O_7$, which is isomorphous with orthorhombic $Gd_2Si_2O_7$. A discussion of the system Y_2O_3-SiO_2 will be found in the article of note [9].

[14] R. C. Ropp, Phosphors based on rare earth phosphates, fast decay phosphors, J. Electrochem. Soc. **115**, 531-535, 1968.
R. C. Ropp, Preparation and spectra of YPO_4:Ce phosphors, J. Luminescence 3, 152-154, 1970.
See also G. Blasse and A. Bril, J. chem. Phys. **47**, 5139, 1967.

Summary. Ce^{3+}-activated phosphors have been found that have such a short decay time τ, good efficiency η and low afterglow level δ that they represent a great improvement on the phosphors previously used in the cathode-ray tube of the flying-spot scanner for colour television and in the beam-indexing colour tube. The luminescence of these phosphors arises from an electron transition between the 5d and 4f levels. The flying-spot scanner requires a continuous spectrum that covers the whole of the visible range. This is obtained with the new phosphor $Y_3Al_5O_{12}$-Ce^{3+} (τ = 70 ns, η = 4%, δ = 6%, λ_{max} = 550 nm) in combination with Y_2SiO_5-Ce^{3+} (τ = 30 ns, η = 6%, δ = 0.1%, λ_{max} = 415 nm). The decay time of the ZnO phosphor previously used is 1000 ns. The index phosphor for the beam-indexing tube should have emission in the near ultra-violet, a decay time of less than about 50 ns, negligible afterglow and a high efficiency. Two new phosphors have been found that meet all these requirements better than those previously used: they are β-$(Y,Ce)_2Si_2O_7$ (τ = 40 ns, η = 8%, δ = 0.1%, δ_{max} = 380 nm) and γ-$(Y,Ce)_2Si_2O_7$ (40 ns, 6.5%, 0.1%, 375 nm). A Ce^{3+} phosphor investigated elsewhere is somewhat faster but has a lower efficiency and a higher afterglow level (25 ns, 2.5%, 1.5%, 330 nm). Practice will show which of these three is to be preferred.

Investigation of the chemical behaviour of clean silicon and germanium surfaces

F. Meijer and G. A. Bootsma

Perfectly clean surfaces are not found in nature. Every surface has foreign molecules adsorbed on it. In the laboratory clean surfaces can be obtained under special conditions, for example by cleaving a crystal in a vacuum. The study of controlled adsorption on these surfaces is of both scientific and technological importance. As an example the authors describe their investigation of the adsorption of methyl alcohol and methyl mercaptan on clean silicon and germanium surfaces. To describe the binding of these molecules to the surface the authors propose certain models. With the aid of these models they explain how it is that other kinds of molecule may be released from the surface than the ones that were adsorbed on it and they show that the crystal orientation of the surface determines just how the molecules will be adsorbed and in what form they will be released from the surface — a fascinating play of forces that reveals the essentials of catalysis.

Introduction

When a crystal is cleaved a new surface is created, and bonds are broken which, in the case of silicon and germanium crystals, are covalent. Each of these uncompensated or "dangling" bonds consists of a single electron which has a strong tendency to form a new covalent bond with another electron. It is difficult for the dangling bonds to form a new bond with one another, because to do so they would have to be deflected too far from their original direction. The consequence of this is that the surface readily reacts with molecules that collide with it. Reactions of this type are the subject of the present article.

The probability that a gas molecule will be bound to the surface upon such a collision is called the "sticking probability". This depends both on the nature of the surface and on the nature of the molecule, and its value is 1 if each collision results in adsorption. The time taken for a newly formed clean surface to be completely covered again depends, of course, not only on the sticking probability but also on the number of collisions, i.e. on the gas pressure. At the low pressure of 10^{-4} N/m^2 (10^{-9} atmosphere) there are still so many collisions that a monomolecular layer (monolayer) is formed on the surface in one second at a sticking probability of 1. Even if the sticking probability is much smaller, e.g. 10^{-2} — which is about the sticking probability for oxygen in contact with silicon or germanium — the surface still becomes so quickly covered with a monolayer that it cannot be investigated. With present-day vacuum technology pressures from 10^{-7} to 10^{-9} N/m^2 can readily be reached, and under these conditions it is possible to keep a surface clean long enough to be able to study it.

Apart from by cleavage *in vacuo* a clean surface can also be obtained by heating a crystal or powder in a vacuum to a very high temperature, causing the evaporation (desorption) of all contaminants from the surface.

If a clean surface is brought into contact with molecules of gases that do not react chemically with the surface, e.g. krypton and methane, a reversible adsorption may occur in which weak Van der Waals bonds are formed between the adsorbed molecules and the substrate atoms; this is known as physical adsorption (physisorption). In adsorption of a chemical nature (chemisorption) irreversible adsorptions occur, which may be the result of the formation of covalent bonds.

Information on the bonds at the surface is present in the region where there is direct interaction between the adsorbed molecules and the surface atoms of the clean surface — i.e. in a layer only one or two molecules thick. Compared with the bulk of the substance a layer as thin as this contains extremely little material, and therefore special techniques are used to study such layers.

In the following section we shall briefly discuss the three methods used in the present investigation. They

Dr. F. Meijer is with Philips Research Laboratories, Eindhoven. Dr. G. A. Bootsma, formerly with Philips Research Laboratories, is now Reader in Physical Chemistry at the University of Utrecht.

are based on the determination of the adsorbed fraction of a quantity of gas admitted to a clean surface (gas-volumetric measurements), or on measurements of the adsorbed layer itself (Auger electron spectroscopy, ellipsometry). In the first case it is desirable to work with large surfaces, such as powders with a total surface area of about 1 m^2; in the second case the surface to be studied may be a crystal face of a few mm^2. A fourth method which we shall not discuss here, is the low-energy electron diffraction method [1], usually abbreviated to "LEED".

The information obtained with these methods relates in the first place to the number of molecules adhering to the surface. Starting from this information a model can be set up to describe the form taken by the molecules upon adsorption at the surface. In the case of chemisorption, bonds in the molecule are broken and at the same time new bonds are formed with the surface. The resultant reaction product between an adsorbed molecule and the surface atoms involved in the reaction is referred to as the adsorption complex. Some simple examples are given in *fig. 1*.

At the end of this article we shall discuss the adsorption of methyl alcohol (CH_3OH) and methyl mercaptan (CH_3SH) on clean silicon and germanium. This study illustrates the results that can be obtained with the three methods, particularly when used in combination.

The adsorption models constitute a first step towards the understanding of catalytic reactions at the surface, and they may also give some insight into epitaxial growth and into crystal growth in general.

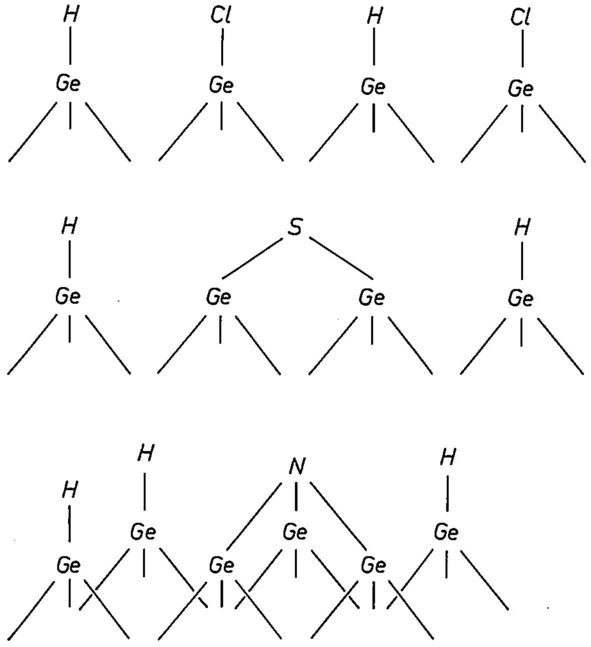


Fig. 1. Schematic representation of the way in which hydrogen chloride (HCl), hydrogen sulphide (H_2S) and ammonia (NH_3) can be adsorbed on (111) faces of clean germanium (Ge).

Some methods of surface investigation

Gas-volumetric measurements

In gas-volumetric measurements a powder is placed under vacuum in a vessel of volume V_2. A quantity of gas is then introduced in an inlet system of volume V_1, and a pressure gauge is used to determine the pressure of the gas before and after admission to the powder (P_1 before, P_2 after). The adsorbed quantity is equal to the difference in quantities of gas before and after adsorption: $P_1V_1 - P_2(V_2 + V_1)$.

The area of the surface on which adsorption takes place can be determined by a method developed by S. Brunauer, P. H. Emmett and E. Teller, called the BET method [2]. This method uses physically adsorbed gases for which the dependence of the adsorbed quantity of gas on the gas pressure is determined. Gases are chosen which occupy a known surface area per molecule on adsorption at the densest possible packing.

The method is based on the assumption that the interaction between the adsorbed molecules and the substrate atoms is stronger than that between the molecules themselves. If the adsorbed quantity of gas is plotted as a function of the pressure at which the adsorption takes place, there will be a knee in the curve at the pressure at which the surface is covered with a monolayer (approximately 1/10 of the saturation pressure). From the quantity adsorbed at this pressure and the area per adsorbed molecule the total area of the powder can be calculated.

As the adsorbed gases are desorbed again when the pressure in the system is lowered, the area of the surface can be determined without the surface remaining permanently contaminated.

A detailed interpretation of the chemical adsorption of powders is complicated because they consist of crystallites that have been cleaved along crystal planes of differing orientations. Since the mode of adsorption may differ appreciably depending on the orientation of the crystal face, what is found is a kind of average adsorption. The exact number of adsorbed molecules on a face with one particular orientation cannot be deduced from this average without furthcr information.

This difficulty is not of course encountered when the chemical adsorption is determined on faces of single crystals, as it is in the other methods.

Auger electron spectroscopy

Auger electron spectroscopy is a relatively new method of investigating surfaces [3]. A beam of electrons with energies of 1-2 keV is used to ionize deep-lying energy levels in an atom. In *fig. 2* level *3*, which lies a few hundred volts below the vacuum level *0*, is ionized. An electron from level *2* can occupy the hole in level *3*, and the energy E_{2-3} thus released can be transferred to an electron in the adjacent level *1*. This electron will then leave the atom with an energy of $E_{2-3} - E_{0-1}$. In this process, called the Auger effect, electrons are thus released with an energy that depends on the energy-level diagram of the atom, but is independent of the energy of the incoming primary electron beam. The energy spectrum of the emergent electrons is measured and peaks are found at energies that are specific to particular kinds of atom. The height of the peaks gives information on the numbers of these atoms. The absolute relation between peak height and number cannot be deduced theoretically and therefore has to be found by calibration [4].

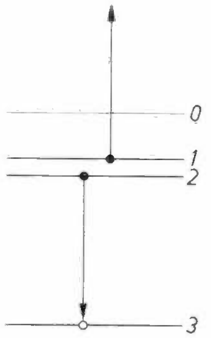


Fig. 2. Illustrating Auger electron spectroscopy.

Ellipsometry

Ellipsometry is an optical method of surface investigation which makes use of the change in the state of polarization of polarized monochromatic light that occurs on reflection at a surface [5]. The component of the incident light that is parallel to the plane of incidence (||) generally undergoes a change different from that of the component perpendicular to the plane of incidence (⊥). The components of the electric light vector parallel and perpendicular to the plane of incidence are given by:

$$E_{\parallel} = A_{\parallel}\, e^{i(\delta_{\parallel}+\omega t)},$$
$$E_{\perp} = A_{\perp}\, e^{i(\delta_{\perp}+\omega t)}.$$

At arbitrary values of the amplitudes A and phases δ the tip of the total electric vector will describe an ellipse. The light is then said to be elliptically polarized. On reflection the shape of the ellipse will change, hence the name ellipsometry. The changes in the amplitudes and phases can be described by two angles, Δ and ψ, which are directly measured by the ellipsometer, and are given by:

$$\Delta = (\delta_{\parallel} - \delta_{\perp})_{(r)} - (\delta_{\parallel} - \delta_{\perp})_{(i)},$$
$$\tan\psi = \frac{(A_{\parallel}/A_{\perp})_{(r)}}{(A_{\parallel}/A_{\perp})_{(i)}}.$$

The subscripts (r) and (i) indicate the reflected and incident light beams.

The reflecting system we are concerned with in this article is illustrated schematically in *fig. 3*. It consists of a substrate with a refractive index n_2 and an extinction coefficient k_2, on which there is a single layer with optical constants n_1 and k_1 and thickness d_1, while the refractive index of the ambient medium is n_0. At the first boundary surface the incident light will be partly reflected and partly transmitted. The same will happen at the other interface, and there will be a multiple reflection between the boundary surfaces of the layer. This causes a phase difference between the directly reflected light and the light that has undergone one or more internal reflections.

The ellipsometric quantities Δ and ψ that describe the reflection are a function of the optical constants of the substrate, layer and ambient medium, and of the layer thickness. If the surface of the substrate is clean, the optical constants n_2 and k_2 of the substrate can be determined directly from Δ and ψ. If a layer forms on this substrate, however, Δ and ψ will change by the amounts $\delta\Delta$ and $\delta\psi$, providing information on the thickness d_1 and on the optical constants n_1 and k_1 of the layer. The accuracy of the ellipsometric method, expressed in terms of measurable layer thickness, is of

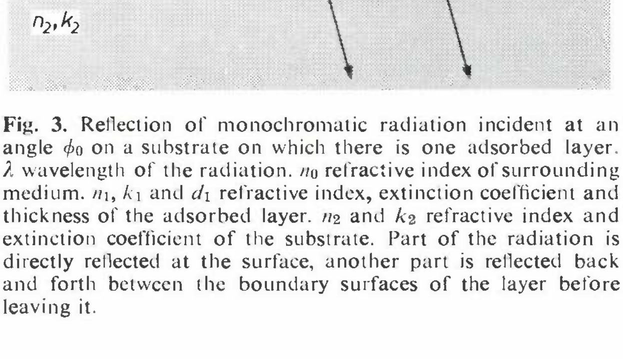


Fig. 3. Reflection of monochromatic radiation incident at an angle ϕ_0 on a substrate on which there is one adsorbed layer. λ wavelength of the radiation. n_0 refractive index of surrounding medium. n_1, k_1 and d_1 refractive index, extinction coefficient and thickness of the adsorbed layer. n_2 and k_2 refractive index and extinction coefficient of the substrate. Part of the radiation is directly reflected at the surface, another part is reflected back and forth between the boundary surfaces of the layer before leaving it.

[1] A description is given in: P. J. Estrup and E. G. McRae, Surface Sci. 25, 1, 1971 (No. 1).

[2] S. Brunauer, P. H. Emmett and E. Teller, J. Amer. Chem. Soc. 60, 309, 1938.

[3] See for example C. C. Chang, Surface Sci. 25, 53, 1971.

[4] Ellipsometry is a useful method here; see for example J. J. Vrakking and F. Meyer, Appl. Phys. Letters 18, 226, 1971.

[5] The principle of ellipsometry has been discussed in detail in this journal: see K. H. Beckmann, Philips tech. Rev. 29, 135, 1968. See also G. A. Bootsma and F. Meyer, Surface Sci. 14, 52, 1969 and F. Meyer, E. E. de Kluizenaar and G. A. Bootsma, Surface Sci. 27, 88, 1971 (No. 1).

the order of magnitude of 0.1 Å (the diameter of an atom is 3-4 Å).

In this article we are concerned with the investigation of monolayers, and most of our investigations have used the ellipsometric method. Since no satisfactory theory has yet been derived for this field of application, we have tested our results with this method by gas-volumetric measurements.

Before discussing these investigations, we ought first to define a few concepts relating to these kinds of layers. In physical adsorption, where there is no localized interaction with individual surface atoms, we define a monolayer as the layer with the densest possible packing; in this case we express the coverage as a fraction θ^* of the monolayer. In chemical adsorption localized bonds are formed with the surface atoms. Here a monolayer is defined as the coverage at which all the dangling bonds of the surface atoms have been compensated; the coverage θ is then the number of adsorbed molecules per surface atom.

Ellipsometric investigation of adsorbed monolayers

If we confine the ellipsometric measurements to very thin adsorbed layers ($\leqslant$ 10 Å) and extrapolate from the macroscopic theory that describes the system with optical constants and thicknesses, we find that $\delta\psi$ is determined almost entirely by the extinction coefficient k_1 of the layer and $\delta\Delta$ by the refractive index n_1 and the thickness d_1. The adsorbed layers discussed in this article consist of molecules of simple organic and inorganic substances which, in the liquid or solid phase, give no significant optical absorption in the range of wavelengths at which the ellipsometric measurements are carried out (340-1800 nm). We have therefore assumed that k_1 is zero, which should give $\delta\psi = 0$.

In the case of "layers" one or two atoms thick it is better to translate the macroscopic quantities n_1 and d_1 into microscopic quantities, i.e. into the number of adsorbed molecules per surface atom — the coverage θ — and the polarizability α of the atoms. Assuming that the relation between the macroscopic and microscopic quantities is given by the Lorentz-Lorenz equation, then the approximate dependence of $\delta\Delta$ on θ and α can be found from:

$$\delta\Delta = Cd_1 \frac{(n_1^2 - 1)}{(n_1^2 + 2)} = C \frac{4\pi}{3} \alpha\theta N .$$

Here N is the number of surface atoms per cm^2, and C is a constant that can be derived from the macroscopic theory. The constant C includes the optical constants of the substrate and the ambient medium. The polarizability α is a quantity that to a good approximation is independent of the environment (unless the chemical bonds differ very considerably, e.g. give rise to ionization). We have therefore used polarizabilities derived from measurements on liquids and gases. This enables us to calculate the coverage directly from $\delta\Delta$ and thus to determine ellipsometrically the number of atoms or molecules on a surface in a quantitative way.

In the next section experiments are described that were carried out to check the validity of the assumptions made, such as the relation between macroscopic and microscopic quantities and the use of "fixed" polarizability values. The experiments consisted in ellipsometric and gas-volumetric measurements, carried out in conditions that were as far as possible identical. Moreover the experimental conditions had to be chosen so as to obtain the same coverage in measurements on a single-crystal surface and on a powder surface. We did this first for an example of physical adsorption (the actual test) and subsequently for an example of chemisorption.

Comparative tests with gas-volumetric measurements

Physical adsorption was chosen for comparing the gas-volumetric and the ellipsometric method because this adsorption is reversible and pressure-dependent, and does not depend very specifically on the type of adsorbing surface. In the experiment performed a suitable gas is admitted into a vacuum system containing both a crystal for ellipsometric measurements and a powder for gas-volumetric measurements. It is assumed that in the equilibrium which is then established the coverage on the powder surface is the same as that on the single-crystal surface. When the pressure of the inlet gas is varied a different coverage is attained, and this enables the ellipsometric effects to be determined as a function of the coverage. The experiment is performed at liquid-nitrogen temperature to obtain the advantage of a low saturation pressure.

In *fig. 4* the values of $\delta\Delta$ found for various gas pressures at a wavelength of 546 nm are plotted as a function of the coverage θ^* found by gas-volumetric measurements for krypton and methane adsorbed on oxidized germanium. Also shown is the relation between $\delta\Delta$ and θ^* calculated with the aid of the polarizability α as reported in the literature. It can be seen that there is good agreement between the measured points and the calculated curves. The value of $\delta\psi$, not given here, was extremely small, as expected for a layer with $k_1 = 0$.

Fig. 5 relates to a layer of krypton adsorbed on clean silicon, for which the coverage, determined volumetrically, corresponds to that of a monolayer. The curves show how $\delta\Delta$ and $\delta\psi$ vary with the wavelength of the light used for the measurement. The curves calculated for a monolayer again show good agreement with the measured points. Here again, the

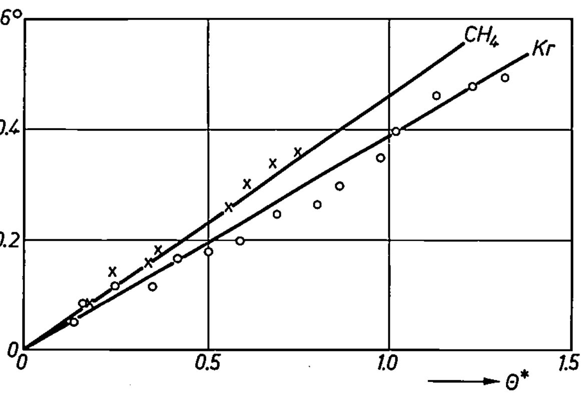


Fig. 4. The ellipsometric quantity Δ changes as a result of physical adsorption of krypton (circles) and methane (crosses) on oxidized germanium by amounts $\delta\Delta$. The graph shows how $\delta\Delta$ varies with the number of adsorbed molecules at the surface. This number is determined by gas-volumetric measurements and expressed as a fraction θ^* of a monolayer. The measurements were made at a wavelength of 546 nm and an angle of incidence of 71.5°. The solid curves show the relationship between $\delta\Delta$ and θ^* as calculated with the aid of values for the polarizability of krypton and methane molecules reported in the literature.

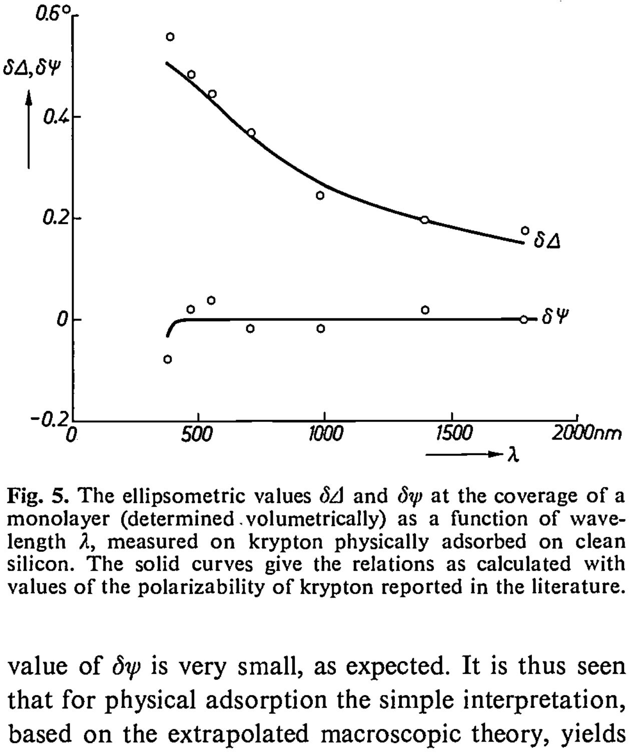


Fig. 5. The ellipsometric values $\delta\Delta$ and $\delta\psi$ at the coverage of a monolayer (determined volumetrically) as a function of wavelength λ, measured on krypton physically adsorbed on clean silicon. The solid curves give the relations as calculated with values of the polarizability of krypton reported in the literature.

value of $\delta\psi$ is very small, as expected. It is thus seen that for physical adsorption the simple interpretation, based on the extrapolated macroscopic theory, yields good results.

Chemisorption

Comparative tests were subsequently made for a number of chemisorptions at room temperature. We chose gases with the general formula H_xA ($x = 1, 2, 3$, etc.) such as HCl, H_2S and NH_3.

These gases have also been investigated by A. H. Boonstra and J. van Ruler [6], who concluded from volumetric measurements that the adsorption of these gases on clean silicon and germanium surfaces shows rather interesting behaviour. A fast, irreversible reaction occurs until the coverage reaches the value $1/2x$, and then the adsorption continues much more slowly. They assumed that a monolayer is present at the coverage $1/2x$.

Since the powder surface considered here mainly consists of (111) faces, which have one "dangling" bond per surface atom, the value $1/2x$ gives rise to the assumption that adsorption complexes arise as illustrated in fig. 1. The dangling or uncompensated bonds are apparently so reactive that the molecules are dissociated on adsorption.

For our purpose the adsorption of the gases H_xA is particularly suitable because the end of the fast reaction can easily be reached in a reproducible manner; because of the marked decrease in the reaction velocity, too long a reaction time gives hardly any further reaction. The same behaviour was found in the ellipsometric measurements: a rapid change in Δ and ψ, followed by a slow further change. It is assumed that the end of the fast reaction on the (111) face can be compared with that on the powder because, as stated, the surface of the powder consists mainly of (111) faces.

It is interesting to note that these measurements on chemically adsorbed layers revealed distinct changes in the value of ψ. This effect was unexpected, since we had assumed that there was *no* optical absorption in the layers investigated (see page 134). It was also found that the $\delta\Delta$ values could not in general be interpreted in the same way as for physical adsorption. Using the volumetrically determined coverage we calculated the optical constants n_1 and k_1 of the layer and found highly improbable values: the refractive index was less than 1 and the extinction coefficient greater than 0.5. This would imply that the adsorbed layers have metallic optical properties. Now solid H_2S, for example, certainly cannot be described as metallic, nor can silicon sulphide or germanium chloride.

It appears that the change in the ellipsometric parameters may be regarded as the sum of a change which is independent of the type of gas and a change which does depend on the adsorbed gas and is in agreement with the values calculated from the volumetrically determined value of θ and the value of α reported in the literature. For the slow adsorption virtually the only effect measured is the "normal" gas-dependent effect. The gas-independent effect is not the same for silicon and germanium surfaces, and also differs slightly for different surface orientations.

We shall now show that these observations can be explained by assuming that the substrate changes at the surface as a result of chemisorption.

[6] A. H. Boonstra and J. van Ruler, Surface Sci. 4, 141, 1966. A. H. Boonstra, Thesis, Eindhoven 1967 (also published as Philips Res. Repts. Suppl. 1968, No. 3).

Model for chemisorption on silicon and germanium

Fig. 6 shows a model that explains the ellipsometric measurements. The diagram at the centre represents the clean crystal, the one on the left the crystal with a physically adsorbed layer, and the one on the right the crystal with a chemically adsorbed layer. The clean crystal surface consists of atoms with uncompensated bonds, and it is assumed that the optical behaviour of this surface layer of atoms differs from that of the rest of the crystal (the thickness of this surface layer is about 5 Å, or 1 to 2 atomic layers). In physical adsorption the uncompensated bonds are not affected and this "transition layer" will continue to exist. All that ellipsometry can show here is the existence of the adsorbed layer. In chemisorption, however, the dangling bonds are compensated, and the optical constants of the transition layer will therefore return to the values applicable to the normal bulk material. The transition layer has then disappeared. Ellipsometry now shows both the normal effect ($\delta\Delta_{\mathrm{ads}}$ and $\delta\psi_{\mathrm{ads}}$) of the adsorbed layer and the effect of the disappearance of the transition layer ($\delta\Delta_{\mathrm{t}}$ and $\delta\psi_{\mathrm{t}}$) (this latter effect is the same for each adsorbed type of molecule that compensates the free bonds):

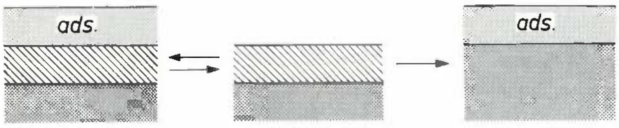


Fig. 6. Schematic representation of the change of clean silicon or germanium (*centre*) for physical adsorption (*on the left*) and for chemisorption (*on the right*). The upper layer (hatched) of the clean substrate, consisting of atoms with uncompensated bonds, is assumed to show a different optical behaviour from that of the rest of the crystal (dark grey). In physical adsorption an adsorbed layer (light grey) forms on the top of the "transition layer" without affecting it; in chemisorption the uncompensated bonds are compensated, and therefore the transition layer is no longer distinguishable from the rest of the crystal.

$$\delta\Delta = \delta\Delta_{\mathrm{t}} + \delta\Delta_{\mathrm{ads}},$$

$$\delta\psi = \delta\psi_{\mathrm{t}} + \delta\psi_{\mathrm{ads}}.$$

In *figs. 7* and *8* the measured values $\delta\Delta$ and $\delta\psi$ and the values of $\delta\Delta_{\mathrm{ads}}$ and $\delta\psi_{\mathrm{ads}}$ calculated by the normal procedure are plotted for various gases as functions of the wavelength of the light used in the measurements. Graphs *b* and *c* in both figures show the effect we have just described: the difference between these measured and calculated values is in each case the same for all gases ($\delta\Delta_{\mathrm{t}}$ and $\delta\psi_{\mathrm{t}}$).

Using the assumed value for the thickness of the transition layer (5 Å) we calculated the effective optical constants of the transition layer from $\delta\Delta_{\mathrm{t}}$ and $\delta\psi_{\mathrm{t}}$. The effective refractive index n_{t} and extinction coefficient k_{t} correspond fairly well to the optical constants of amorphous silicon and germanium. This is not unexpected, since an amorphous substance will contain a great many uncompensated or distorted bonds which, as compared with the crystalline material, will show the same optical deviations as the uncompensated bonds of the transition layer.

The measurements for the further investigation were made at wavelengths for which $\delta\Delta_{\mathrm{t}} = 0$, i.e. for silicon at 546 nm and for germanium at 800 nm (see figs. 7 and 8). At these wavelengths the above equations change to:

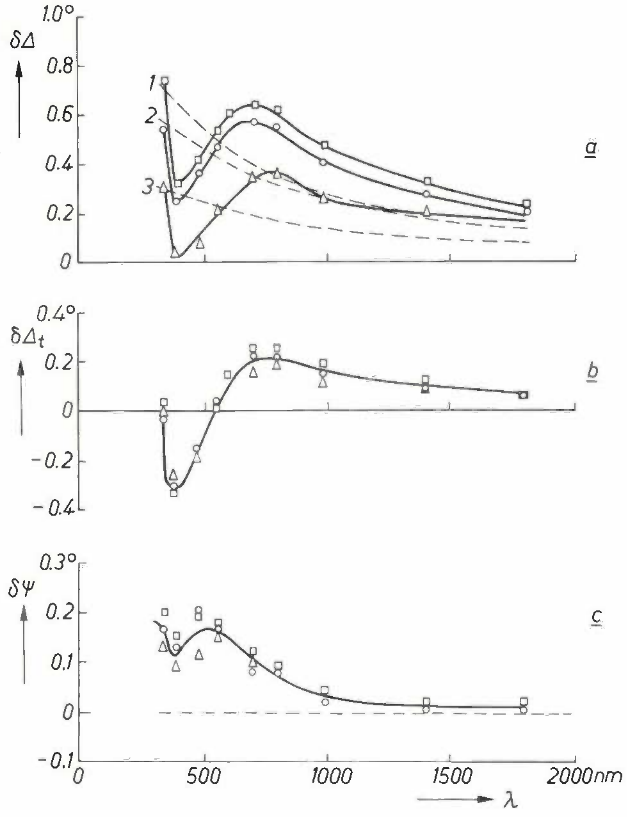


Fig. 7. *a*) The quantity $\delta\Delta$ for the chemisorption of O_2 (squares), Cl_2 (circles) and H_2S (triangles) on the (111) face of silicon, and the calculated $\delta\Delta_{\mathrm{ads}}$ for these substances (dashed curves) as a function of wavelength λ.
b) The difference $\delta\Delta_{\mathrm{t}}$ between $\delta\Delta$ and $\delta\Delta_{\mathrm{ads}}$ is found to have the same values for the three substances at each wavelength.
c) The quantity $\delta\psi$ for chemisorption of the above three substances on (111) faces of silicon, and the calculated $\delta\psi_{\mathrm{ads}}$ (dashed curve) as a function of wavelength λ. It can be seen from the figure that $\delta\psi_{\mathrm{ads}}$ is zero, and, as in the case (*b*), the conclusion is that $\delta\psi_{\mathrm{t}}$ is the same for the three substances.

$$\delta\Delta = \delta\Delta_{\mathrm{ads}} \qquad \delta\Delta_{\mathrm{t}} = 0,$$

$$\delta\psi = \delta\psi_{\mathrm{t}} \qquad \delta\psi_{\mathrm{ads}} = 0.$$

The effect of the adsorbed layer is then fully given by $\delta\Delta$, and the effect of the compensation of the free bonds by $\delta\psi$. The curves in which $\delta\psi$ is plotted against of $\delta\Delta$ therefore have a knee at the point where the coverage is reached at which all free bonds are compensated. A typical curve is shown in *fig. 9*.

Ellipsometric measurements are now sufficient for determining both the coverage and the number of compensated free bonds at the surface. This new development makes ellipsometry a very useful method for studying the surfaces of silicon and germanium. It is very likely that the results of measurements on other substances can be interpreted in a similar way.

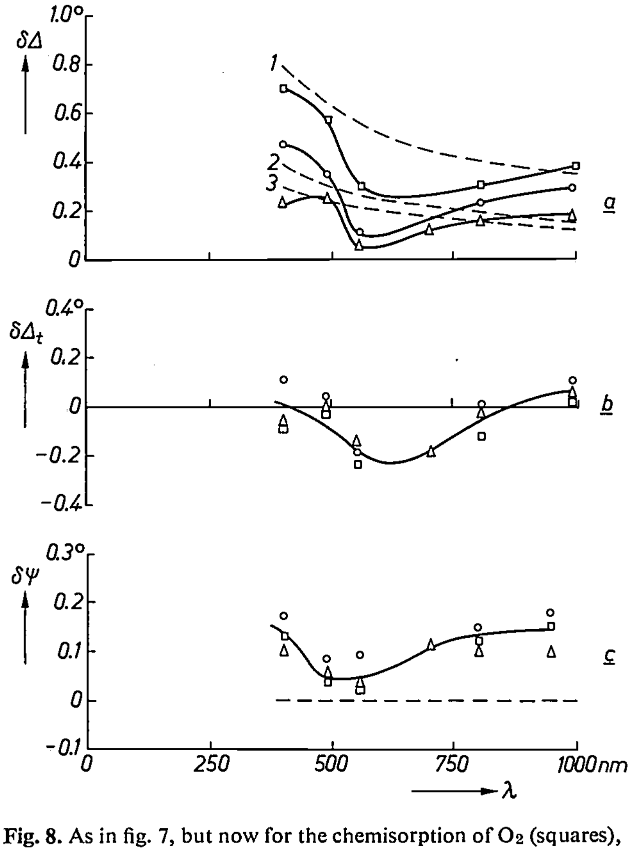


Fig. 8. As in fig. 7, but now for the chemisorption of O_2 (squares), HCl (circles) and H_2S (triangles) on the (111) face of germanium.

The adsorption of methyl alcohol and methyl mercaptan

As an example we shall now discuss a study we have made of the adsorption of methyl alcohol [7] (CH_3OH) and of methyl mercaptan [8] (CH_3SH) on silicon and germanium. The adsorption of these substances was studied on germanium powders by means of gas-volumetric measurements and on (111) and (100) faces of silicon and germanium with the aid of ellipsometry. An initial conclusion is that there is no difference in coverage between silicon and germanium, which agrees with A. H. Boonstra's results for the adsorption of H_xA gases [6].

In *figs. 10* and *11* values of $\delta\psi$ are plotted against values of $\delta\Delta$, and *Table I* gives the coverages calculated from these curves for the completed monolayer. The

Table I. Ellipsometrically determined coverages θ (decimal fractions) for monomolecular coverage with methyl alcohol (CH_3OH) and with methyl mercaptan (CH_3SH) of (111) and (100) faces of silicon and germanium. The fractions are the values that correspond to the model.

	Si(111)	Si(100)	Ge(111)	Ge(100)
CH_3OH	0.25 $\frac{1}{4}$	0.53 $\frac{1}{2}$	0.23 $\frac{1}{4}$	0.45 $\frac{1}{2}$
CH_3SH	0.15 $\frac{1}{6}$	0.55 $\frac{1}{2}$	0.14 $\frac{1}{6}$	0.42 $\frac{1}{2}$

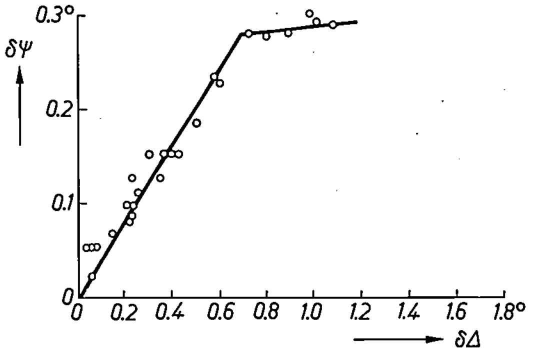


Fig. 9. The quantity $\delta\psi$ as a function of $\delta\Delta$ for the chemisorption of O_2 on the (111) face of silicon. The knee in the curve corresponds to the coverage at which all bonds are just compensated.

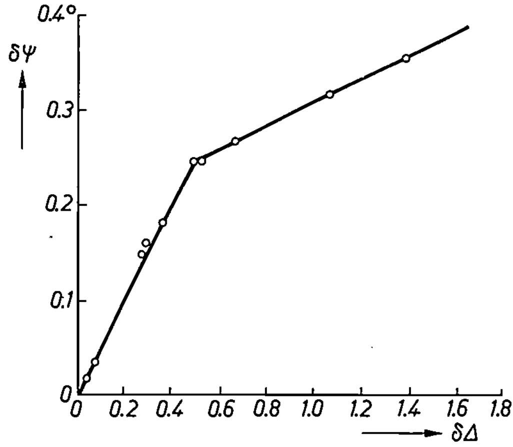


Fig. 10. The quantity $\delta\psi$ as a function of $\delta\Delta$ for the chemisorption of methyl alcohol (CH_3OH) on the (111) face of silicon.

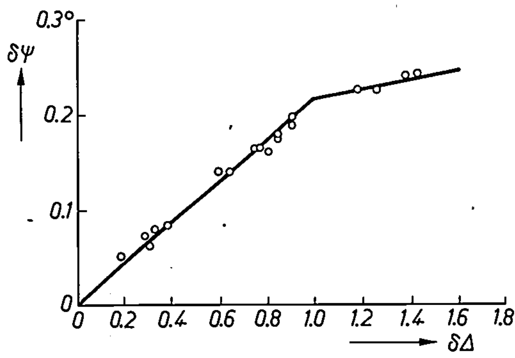


Fig. 11. The quantity $\delta\psi$ as a function of $\delta\Delta$ for the chemisorption of methyl mercaptan (CH_3SH) on the (111) face of germanium.

[7] F. Meyer, J. phys. Chem. 73, 3844, 1969.

[8] F. Meyer and J. M. Morabito, J. phys. Chem. 75, 2922, 1971 (No. 19).

error in these calculated values can be fairly large (about 20%), mainly because the polarizability values of the adsorbed molecules are not known with sufficient accuracy. In the further calculations we therefore used an "idealized" coverage, based on the following assumptions.

In the first place it was assumed that the molecules under consideration are dissociated upon chemisorption, and that free bonds of the surface atoms are compensated in the process. This takes place with one bond per surface atom of a (111) face and with two bonds per surface atom of a (100) face; see fig. 12. It was further assumed that the normal covalent valencies are maintained in the adsorption complex, and that no common bonds are formed between the adsorbed molecules. Since each dissociation gives two free bonds, an adsorbed molecule can only compensate an even number of free bonds of the substrate. The possible coverages are thus one molecule per 2, 4, 6, etc. surface atoms for a (111) face and one molecule per 1, 2, 3, etc. surface atoms for a (100) face. Using the data in Table I it is not difficult to make the choice from these surface coverages for each of the adsorptions under consideration.

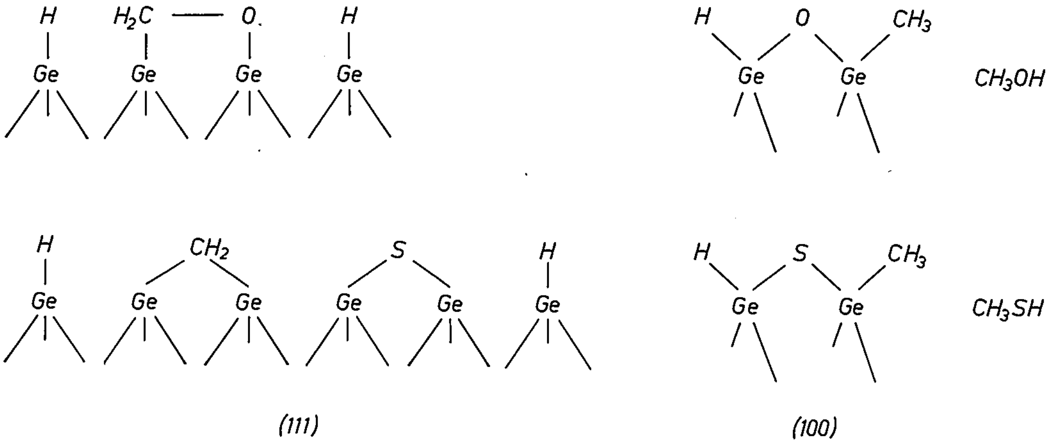


Fig. 12. Schematic representation of the adsorption complexes of CH_3OH (*above*) and CH_3SH (*below*) on the (111) face (*on the left*) and on the (100) face (*on the right*) of germanium.

For the adsorption of CH_3OH and CH_3SH on (111) faces we find coverages of 1/4 and 1/6 respectively, whereas a coverage of 1/2 is found for the adsorption of these substances on (100) faces.

In the volumetric measurements of the adsorption of CH_3OH and CH_3SH on germanium powders average coverages were found of 0.30 and 0.23 respectively. From the combined results of the ellipsometric and volumetric measurements it follows that the (111) and (100) faces occur in the powder in a ratio of 80:20. This agrees fairly well with the ratio of 70:30 calculated on the basis of estimated bonding energies.

Fig. 12 shows our models of the adsorption complexes formed by the molecules CH_3OH and CH_3SH on (111) and (100) faces of germanium. These structures agree both with the coverages determined and with the results of desorption measurements.

The desorption experiments were performed as follows. Methyl alcohol or methyl mercaptan was adsorbed on a clean germanium-powder surface. The unadsorbed gas was pumped out, and the powder with the adsorbed layer was then heated in a closed system. The resultant desorption products were analysed with a mass spectrometer. *Figs. 13* and *14* give the amounts of desorbed gas as a function of temperature. With methyl alcohol carbon monoxide (CO) and hydrogen (H_2) form at relatively low temperature, and methane (CH_4) starts to form at about 400 °C. With methyl mercaptan, hydrogen is desorbed in two steps, and here again methane starts to form at about 400 °C.

These results lead to the assumption that different adsorption-desorption reactions take place on faces of different crystallographic orientation. The total quantity of methane desorbed at the higher temperature corresponds exactly to the quantity of CH_3OH or CH_3SH that was adsorbed on the (100) faces of the powder (assuming that the coverage is 1/2 and that the (100) faces constitute 20% of the surface area of the powder). We were able to verify that the methane is indeed formed via these faces by an experiment in which a stack of germanium platelets with a (100) orientation was used instead of powder. The total surface area was too small for accurate volumetric measurements, but it was nevertheless possible to demon-

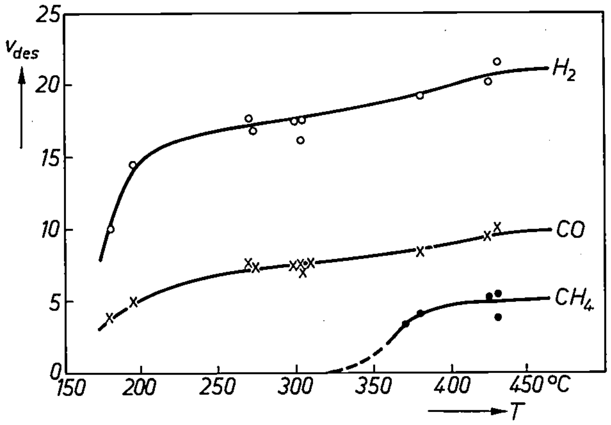


Fig. 13. When the temperature of the germanium-powder surfaces on which CH_3OH is adsorbed is gradually increased, hydrogen, carbon monoxide and methane are formed. The curve shows how the quantity V_{des} (in arbitrary units) of the released gas depends on the temperature.

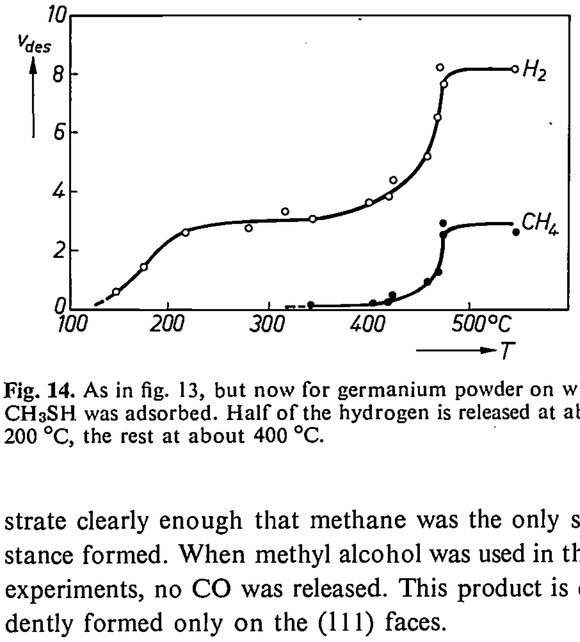


Fig. 14. As in fig. 13, but now for germanium powder on which CH_3SH was adsorbed. Half of the hydrogen is released at about 200 °C, the rest at about 400 °C.

strate clearly enough that methane was the only substance formed. When methyl alcohol was used in these experiments, no CO was released. This product is evidently formed only on the (111) faces.

In both cases the formation of methane was also found to depend on the hydrogen pressure in the reaction vessel, and in fact it stopped altogether when the hydrogen pressure fell below a particular value. The hydrogen is released by desorption at the lower temperatures before there is any desorption of methane. If, however, this desorbed hydrogen is pumped off before the desorption of methane begins (at 400 °C), then no methane is formed but the equivalent quantity of hydrogen is produced. The effect of the hydrogen pressure on the desorption of methane can be explained with a reaction diagram as given in *fig. 15.*

To conclude this article we shall deal at somewhat greater length with the principal considerations that prompted the choice of the structural formulae of fig. 12, then discuss a few supplementary Auger measurements, and finally we shall consider possible refinements of the models.

The basic assumption of our structural models was that all the free bonds of the surface atoms are compensated and that all the atoms possess their normal valencies. For the adsorption complex of methyl alcohol on the (111) face, however, there are five different alternatives for producing a coverage of 1/4 in this way. One of the reasons for choosing the structure given in fig. 12 was that methane is known not to be chemically adsorbed, and this makes it improbable that the first adsorption step will be a dissociation of a C-H bond. The fact that CO is desorbed at such low temperatures again indicates that CO is already present as a unit in the complex. The first adsorption step is therefore presumed to be the dissociation of an O-H bond. The model of fig. 12 is then the only one that satisfies the various conditions.

As we have seen, no CO is released via adsorption on (100) faces, only CH_4; the CO does not therefore have to be present here as a unit.

For the CH_3SH adsorption on the (111) face a coverage of 1/6 was found. The desorption of hydrogen from the (111) face in two steps corresponds to the occurrence of hydrogen bonds of two kinds, one on germanium and one on carbon. The first kind of bond is weaker, and it is in fact known that hydrogen adsorbed on germanium is desorbed at about 200 °C. In fig. 14 it can be seen that about the same amount of hydrogen is released in both desorption steps. This is also in agreement with the model in fig. 12.

Fig. 16 shows that the adsorption process can be followed with the aid of Auger electron spectroscopy. This figure relates to the adsorption of CH_3SH on (100) faces of silicon. The graph shows the linear increase of the carbon and sulphur peaks with increasing coverage, and the associated decrease of the silicon peak as a result of the coverage by the adsorbed atoms.

The Auger measurements give information in the first place about the initial state of the crystal. The indications are that a clean surface, as used for the

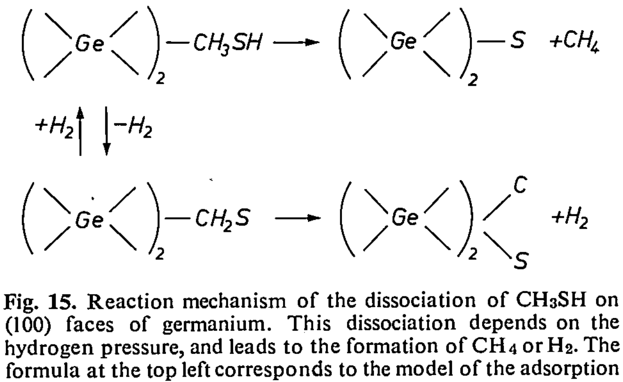


Fig. 15. Reaction mechanism of the dissociation of CH_3SH on (100) faces of germanium. This dissociation depends on the hydrogen pressure, and leads to the formation of CH_4 or H_2. The formula at the top left corresponds to the model of the adsorption complex on a (100) face of germanium, given in fig. 12. The other structures occur at higher temperature; it is not possible to give detailed models of these structures.

ellipsometric measurements, contains less than 1% of contaminant (1% is the detection limit). The measurements also show that under the experimental conditions (at room temperature) only the admitted gas is adsorbed and not, for example, gas that may have been released from the wall of the vacuum vessel.

A refinement of the models can be made by calculating the extent to which the surface atoms of the substrate can be displaced to form an oxygen or sulphur bridge. A simple calculation, based on the lengths and possible bending of the bonds, shows that on the (100) face the distance between the surface atoms can decrease by 10 to 20% in order to form a bridge. In general it is easier to displace surface atoms on the (100) face, where there are only two bonds with the underlying layer, than on the (111) face, where a surface atom is held by three bonds. A consequence of this is that certain adsorptions take place more readily on the (100) face than on the (111) face. It is possible for example to cover the (100) faces of a germanium powder completely with methyl chloride, without there being any reaction with the (111) faces.

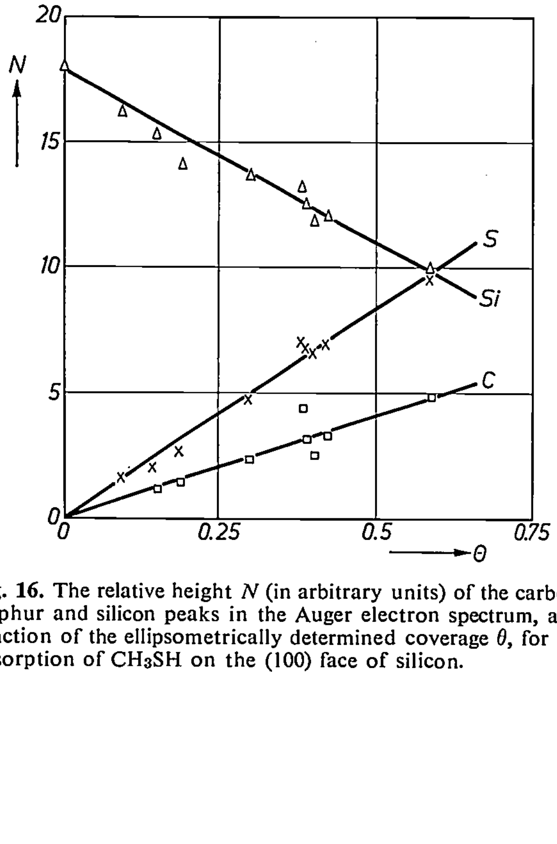


Fig. 16. The relative height N (in arbitrary units) of the carbon, sulphur and silicon peaks in the Auger electron spectrum, as a function of the ellipsometrically determined coverage θ, for the adsorption of CH_3SH on the (100) face of silicon.

With the aid of various examples we have tried to show the extent to which surface studies can provide a description of adsorptions on an atomic scale. The most outstanding result of the investigations on silicon and germanium is the observation that differently oriented crystal faces have quite different adsorption behaviour. This behaviour determines the nature of the catalytic reactions at the surface, and may possibly also influence the way in which epitaxial growth takes place.

Summary. After a short discussion of three methods of surface study — gas-volumetric measurements, Auger electron spectroscopy and ellipsometry — it is explained how ellipsometry can be applied for investigating monomolecular layers on clean silicon and germanium surfaces. Since no satisfactory theory yet exists for this application of ellipsometry, the results obtained are checked against gas-volumetric measurements. It is demonstrated that for physically adsorbed gases all the assumptions made were correct. The study of chemical adsorption revealed that the upper layer on the surface shows an optical behaviour different from that of the rest of the crystal, owing to the presence of uncompensated bonds. The article concludes with a description of an investigation of the adsorption of CH_3OH and CH_3SH on (111) and (100) crystal faces. It is found that CH_3OH is bound to (111) and (100) faces with coverages of 1/4 and 1/2, respectively, and CH_3SH with coverages of 1/6 and 1/2. The adsorption models postulated are in agreement with these coverages and also with the results of a number of desorption experiments. A striking aspect of the adsorptions studied is that they differ considerably for different crystallographic orientations of the surfaces.

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 Forschungslaboratorium Aachen GmbH, Weißhausstraße, 51 Aachen, Germany	*A*
Philips Forschungslaboratorium Hamburg GmbH, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (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.

J. Adams & I. C. P. Millar: Detection efficiency of channel plates to X-rays.
Acta Electronica **14**, 237-244, 1971 (No. 2). (*Also in French.*) *M*

C. L. Alting: Het vervaardigen van kleine glazen prisma's.
Glastechn. Meded. **9**, 92-97, 1971 (No. 3). *E*

A. C. Aten, J. C. Duran, J. J. Geurts & A. J. Griffioen (Philips Electronic Components and Materials Division (Elcoma), Eindhoven): Chemical transport in oxide cathodes.
Philips Res. Repts. **26**, 519-531, 1971 (No. 6).

M. Bannink & H. D. M. Ribberink (Philips Radio, Television and Record-playing Equipment Division, Eindhoven): De temperatuurverdeling tijdens het opwarmen en de opwarmtijd van voorwerpen in moffelovens.
Verfkroniek **44**, 276-286, 1971 (No. 9).

L. K. H. van Beek, J. R. G. C. M. van Beek, J. Boven & C. J. Schoot: Syntheses and cis-trans isomerization of light-sensitive benzenediazo sulfides.
J. org. Chem. **36**, 2194-2196, 1971 (No. 15). *E*

V. Belevitch & Y. Genin: Implicit interpolation, trigradients and continued fractions.
Philips Res. Repts. **26**, 453-470, 1971 (No. 6). *B*

H. Bex & E. Schwartz: Performance limitations of lossy circulators.
IEEE Trans. MTT-**19**, 493-494, 1971 (No. 5). *A*

P. Blood & R. J. Tree: The scattering factor for geometrical magnetoresistance in GaAs.
J. Physics D **4**, L 29-31, 1971 (No. 9). *M*

J. A. den Boer, A. M. E. Hoeberechts, W. K. Hofker, D. P. Oosthoek (all with Philips Research Laboratories, Dept. Amsterdam), **K. Mulder, L. A. Ch. Koerts, R. van Dantzig, J. E. J. Oberski, J. H. Dieperink, E. Kok & R. F. Rumphorst** (all with Institute for Nuclear Physics Research (IKO), Amsterdam): Position sensitive detector telescopes for charged particles used in the BOL-system.
Nucl. Instr. Meth. **92**, 173-176, 1971 (No. 2).

J.-P. Boutot: Degassing of microchannel plates.
Acta Electronica **14**, 245-262, 1971 (No. 2). (*Also in French.*) *L*

C. J. Bouwkamp: A new solid pentomino problem.
J. recreat. Math. **4**, 179-186, 1971 (No. 3). *E*

J. J. van den Broek & H. Zijlstra: Calculation of intrinsic coercivity of magnetic domain walls in perfect crystals.
IEEE Trans. **MAG-7**, 226-230, 1971 (No. 2). *E*

K. Bulthuis & G. J. Ponsen: Vibrational relaxation of the CO_2 lower laser level by H_2O.
Physics Letters **36A**, 123-124, 1971 (No. 2). *E*

K. H. J. Buschow: Structural and magnetic characteristics of Th-Co and Th-Fe compounds.
J. appl. Phys. **42**, 3433-3437, 1971 (No. 9). *E*

K. H. J. Buschow: The samarium-iron system.
J. less-common Met. **25**, 131-134, 1971 (No. 2). *E*

H. B. G. Casimir: 50 jaar theoretische natuurkunde.
Ned. T. Natuurk. **37**, 135-146, 1971 (No. 7). *E*

V. Chalmeton: Detection of high-energy electrons.
Acta Electronica **14**, 225-236, 1971 (No. 2). (*Also in French.*) *L*

V. Chalmeton & P. Chevalier: Pulse height distribution for single electron input.
Acta Electronica **14**, 99-106, 1971 (No. 1). (*Also in French.*) *L*

T. D. Clark: Interaction of microwaves with point contact Josephson junction arrays.
Proc. 12th Int. Conf. on Low Temperature Physics, Kyoto 1970, pp. 449-451. *M*

P. J. Courtois & J. Georges: On a single-server finite queuing model with state-dependent arrival and service processes.
Operations Res. **19**, 424-435, 1971 (No. 2). *B*

H. Dammann: Phase holograms of diffuse objects.
Applications de l'Holographie, C.R. Symp. Int., Besançon 1970, paper 11.14, 4 pp. *H*

R. Davies & B. H. Newton: Design trends for Gunn oscillators.
Electronic Equipment News **13**, No. 5, 18-23, Sept. 1971. *M*

P. Delsarte: Majority logic decodable codes derived from finite inversive planes.
Information and Control **18**, 319-325, 1971 (No. 4). *B*

P. Dewilde (University of California, Berkeley, Cal.), **V. Belevitch & R. W. Newcomb** (University of Maryland, College Park, Md.): On the problem of degree reduction of a scattering matrix by factorization.
J. Franklin Inst. **291**, 387-401, 1971 (No. 5). *B*

U. Dibbern: Rechnergesteuerte Sprachausgabe — ein Vergleich verschiedener Verfahren.
Int. elektron. Rdsch. **25**, 244-246, 1971 (No. 10). *H*

J. Dieleman: Comment on "p-type conduction in Li-doped ZnSe".
Appl. Phys. Letters **19**, 84-85, 1971 (No. 4). *E*

J. Dieleman & A. R. C. Engelfriet: The phase diagram of the system $Ga_{1-x}Se_x$ for $0.5 \leqslant x \leqslant 0.6$ and $300\,K \leqslant T \leqslant 1500\,K$.
J. less-common Met. **25**, 231-233, 1971 (No. 2). *E*

A. Downey, C. E. Fuller & R. W. Lindop: Reduction of the base/collector capacitance in high speed logic IC transistors by collector-profiling.
Microelectronics and Reliability **10**, 381-385, 1971 (No. 5). *M*

G. Engelsma & J. M. H. van Bruggen: Ethylene production and enzyme induction in excised plant tissues.
Plant Physiol. **48**, 94-96, 1971 (No. 1). *E*

G. Eschard: Neutron generators.
Acta Electronica **13**, 293-303 (*in French*), 305-315 (*in English*), 1970 (No. 4). *L*

G. Eschard & B. W. Manley: Principle and characteristics of channel electron multipliers.
Acta Electronica **14**, 19-39, 1971 (No. 1). (*Also in French.*) *L, M*

S. R. Fletcher, A. C. Skapski (both with Imperial College, London) **& E. T. Keve:** Crystallographic studies of irradiation/field treated triglycine sulphate: a new structure form.
J. Physics C 4, L 255-258, 1971 (No. 13). *M*

B. B. van der Genugten (Philips Information Systems and Automation Division, Eindhoven): The distribution of random variables reduced modulo *a*.
Philips Res. Repts. **26**, 471-485, 1971 (No. 6).

C. J. Gerritsma & P. van Zanten: Electric-field-induced texture transformation and pitch contraction in a cholesteric liquid crystal.
Mol. Cryst. liq. Cryst. **15**, 257-268, 1971 (No. 3). *E*

J. J. Goedbloed: F.m. noise of low-level-operating IMPATT-diode oscillators.
Electronics Letters **7**, 445-446, 1971 (No. 16). *E*

J. J. Goedbloed: Over het opwekken van microgolven met behulp van avalanche-diodes.
Ingenieur **83**, ET 63-74, 1971 (No. 22). *E*

P. Groenveld: Radiostoringen, veroorzaakt door thyristorregelingen.
Ingenieur **83**, ET 74-81, 1971 (No. 22). *E*

G. Groh & H. Weiss: A simple extended source for high resolution holography.
Optics Comm. **4**, 63-65, 1971 (No. 1). *H*

A. J. Guest: A computer model of channel multiplier plate performance.
Acta Electronica **14**, 79-97, 1971 (No. 1). (*Also in French.*) *M*

P. Guétin & G. Schréder: Tunneling in Pb/n-GaAs junctions under hydrostatic pressure.
Solid State Comm. **9**, 591-593, 1971 (No. 9). *L*

W. van Haeringen & H.-G. Junginger: Model calculation of the electron phonon coupling parameter λ.
Z. Physik **246**, 281-294, 1971 (No. 4). *E, A*

S. H. Hagen: The conduction mechanism in silicon carbide voltage-dependent resistors.
Philips Res. Repts. **26**, 486-518, 1971 (No. 6). *E*

P. Hansen & U. Krey (I. Inst. Theor. Physik Universität Hamburg): Das Spinwellenspektrum dünner magnetischer Schichten.
Z. angew. Physik **32**, 104-108, 1971 (No. 2). *H*

B. M. A. Hijdra: Ein kompatibler Rundfunkempfänger mit kontinuierlicher Abstimmung für den Empfang von Einseitenbandsendungen mit teilweise unterdrücktem Träger.
Rundfunktechn. Mitt. **15**, 101-107, 1971 (No. 3). *E*

L. Hollan, J. Hallais & C. Schiller: Interface study in epitaxial vapour grown GaAs.
J. Crystal Growth **9**, 165-170, 1971. *L*

F. A. de Jonge & W. F. Druyvesteyn: Roosterfouten in magnetische bubble-roosters.
Ned. T. Natuurk. **37**, 393-394, 1971 (No. 16). *E*

H. Jonker, L. K. H. van Beek, C. J. Dippel, C. J. G. F. Janssen, A. Molenaar & E. J. Spiertz: Principles of PD recording systems and their use in photofabrication.
J. photogr. Sci. **19**, 96-105, 1971 (No. 4). *E*

Y. Kamp: Realization of multivariable functions by a cascade of lossless two-ports separated by noncommensurate stubs.
Philips Res. Repts. **26**, 443-452, 1971 (No. 6). *B*

Y. Kamp & V. Belevitch: A class of multivariable positive real functions realizable by the Bott-Duffin method.
Philips Res. Repts. **26**, 433-442, 1971 (No. 6). *B*

A. Klopfer: Adsorption und Elektronenstoßdesorption von Sauerstoff an Nickel.
Ber. Bunsen-Ges. phys. Chemie **75**, 1070-1073, 1971 (No. 10). *A*

E. Krätzig: Ultrasonic attenuation of gapless superconducting films.
Physics Letters **37A**, 21-22, 1971 (No. 1). *H*

E. Lagendijk, W. J. Huiskamp (both with Kamerlingh Onnes Laboratory, Leiden) **& P. F. Bongers:** Heat capacity measurements on the diluted Ising system $Cs_3Co_pZn_{1-p}Cl_5$.
J. Physique **32**, C1/1008-1009, 1971 (Colloque No. 1, Vol. II). *E*

H. de Lang: Flow lines on the Poincaré sphere as an aid to the study of mode polarization in lasers.
IEEE J. **QE-7**, 441-444, 1971 (No. 9). *E*

R. Lorenz: Drucker mit direktem elektrostatischem Farbübertrag.
Feinwerktechnik **75**, 368-373, 1971 (No. 9). *H*

F. K. Lotgering & G. H. A. M. van der Steen: Ferromagnetic $Cu_{1+y}Cr_2Te_4$ and $CuAg_yCr_2Te_4$ with metal-excessive spinel structure.
Solid State Comm. **9**, 1741-1744, 1971 (No. 20). *E*

C. Loty: Saturation effects in channel electron multipliers.
Acta Electronica **14**, 107-119, 1971 (No. 1). (*Also in French.*) *L*

F. Meyer: Ellipsometric study of adsorption complexes on silicon.
Surface Sci. **27**, 107-116, 1971 (No. 1). *E*

F. Meyer, E. E. de Kluizenaar & G. A. Bootsma: Ellipsometry and the clean surfaces of silicon and germanium.
Surface Sci. **27**, 88-106, 1971 (No. 1). *E*

I. C. P. Millar: Detection efficiency of channel electron multipliers to electromagnetic radiation and positive ions.
Acta Electronica **14**, 145-150, 1971 (No. 2). (*Also in French.*) *M*

J. Monin, G. Hincelin (both with Conservatoire National des Arts et Métiers, Paris) **& G.-A. Boutry:** Détermination des constantes optiques du potassium dans le visible et le proche ultraviolet.
C.R. Acad. Sci. Paris **272B**, 761-763, 1971 (No. 12). *L*

M. Monneraye: La dévitrification contrôlée du verre.
Céramiques industr. No. 8, 45-49, 1971. *L*

J. Mulder: The calculation of counterflow recuperators with non-ideal gas and heat conduction in longitudinal direction.
Bull. Inst. Int. du Froid, Annexe 1970-1, pp. 231-241. *E*

A. Netten & A. H. Boonstra: Three-step adsorption of hydrogen sulphide at about — 120 °C on vapour-deposited lead monoxide layers.
Surface Sci. **27**, 77-87, 1971 (No. 1). *E*

S. G. Nooteboom (Institute for Perception Research, Eindhoven): Enkele opmerkingen over de relatie tussen generatieve fonologie en experimentele fonetiek.
Studia Neerlandica No. 6, 169-178, 1971.

S. G. Nooteboom (Institute for Perception Research, Eindhoven): Over de lengte van korte klinkers, lange klinkers en tweeklanken in het Nederlands.
Nieuwe Taalgids **64**, 396-402, 1971 (No. 5).

A. E. Pannenborg (Philips Board of Management, Eindhoven): Industriële innovatie en de stimulering ervan.
Ingenieur **82**, A 978-981, 1970 (No. 50).

C. H. Petley (Mullard Ltd., Mitcham, Surrey, England) **& R. Pook:** Some applications of single channel multipliers.
Acta Electronica **14**, 151-157, 1971 (No. 2). (*Also in French.*) *M*

U. Pick: A simple etching method for thin film patterns.
J. Physics E **4**, 925-926, 1971 (No. 11). *M*

P. Piret: Some optimal type B_1 convolutional codes.
IEEE Trans. **IT-17**, 355-356, 1971 (No. 3). *B*

R. Plumier (Centre d'Etudes Nucléaires de Saclay, Gif-sur-Yvette, France), **F. K. Lotgering & R. P. van Stapele:** Magnetic properties of $Cu_{1/2}In_{1/2}Cr_2S_4$ and some related compounds.
J. Physique **32**, C1/324-325, 1971 (Colloque No. 1, Vol. I). *E*

J. Polman: Relaxation of the electron velocity distribution in a time-dependent weakly ionized plasma.
Physica **54**, 305-317, 1971 (No. 2). *E*

R. Pook: The manufacture and performances of single channel electron multipliers.
Acta Electronica **14**, 135-143, 1971 (No. 2). (*Also in French.*) *M*

A. Rabenau, H. Rau & G. Rosenstein: Goldselenidhalogenide.
Monatsh. Chemie **102**, 1425-1428, 1971 (No. 5). *A*

H. I. Ralph & F. D. Hughes: Capture cross section of trapping centres in polar semiconductors.
Solid State Comm. **9**, 1477-1480, 1971 (No. 17). *M*

G. W. Rathenau: Natuurkunde in de Nederlandse industrie.
Ned. T. Natuurk. **37**, 189-191, 1971 (No. 7). *E*

J. A. J. Roufs (Institute for Perception Research, Eindhoven): Threshold perception of flashes in relation to flicker.
The perception and application of flashing lights, Proc. int. Symp., London 1971, pp. 29-42.

P. Schagen & G. Piétri: Channel electron multipliers.
Acta Electronica **14**, 13-18, 1971 (No. 1). (*Also in French.*) *M*, *L*

E. Schwartz: Über die Existenz oberer Grenzen der Nullstellen-Empfindlichkeit bei beliebigen Impedanzen.
Arch. Elektronik & Übertr.technik (A.E.Ü.) **25**, 379-386, 1971 (No. 8). *A*

J. M. Shannon: D.c. measurement of the space charge capacitance and impurity profile beneath the gate of an MOST.
Solid-State Electronics **14**, 1099-1106, 1971 (No. 11). *M*

J. G. Siekman: Materiaalbewerking met de CO_2-laser.
Glastechn. Meded. **9**, 38-48, 1971 (No. 2). *E*

I. H. Slis (Institute for Perception Research, Eindhoven): Articulatory effort and its durational and electromyographic correlates.
Phonetica **23**, 171-188, 1971 (No. 3).

C. G. Sluyter: Audiovisuele media bij onderzoek in de industriële research.
Polytechn. T. Elektr. **26**, 334-336, 1971 (No. 9). *E*

H. J. L. Trap: Electronic conductivity in oxide glasses.
Acta Electronica **14**, 41-77, 1971 (No. 1). (*Also in French.*) *E*

H. J. L. Trap & J. M. Stevels: Les verres à conductibilité électronique, leurs propriétés et quelques applications en électronique.
Verres Réfract. **25**, 176-196, 1971 (No. 4/5). *E*

C. van Trigt: Analytically solvable problems in radiative transfer, III.
Phys. Rev. A **4**, 1303-1316, 1971 (No. 3). *E*

J. van der Veen & A. H. Grobben: The conformation of aromatic Schiff bases in connection with liquid crystalline properties.
Mol. Cryst. liq. Cryst. **15**, 239-245, 1971 (No. 3). *E*

M. A. Verschuren & H. F. Wilbrink: Vormgeving aan glas.
Glastechn. Meded. **9**, 58-60, 1971 (No. 2). *E*

J. F. Verwey: On the emitter degradation by avalanche breakdown in planar transistors.
Solid-State Electronics **14**, 775-782, 1971 (No. 9). *E*

J. Volger: Natuurkunde en Maatschappij.
Ned. T. Natuurk. **37**, 200-210, 1971 (No. 7). *E*

K. Walther & K. H. Sarges: Sound velocity and magnetoelastic coupling in MnTe.
Physics Letters **36A**, 309-310, 1971 (No. 4). *H*

D. Washington, V. Duchenois, R. Polaert & R. M. Beasley: Technology of channel plate manufacture.
Acta Electronica **14**, 201-224, 1971 (No. 2). (*Also in French.*) *M*, *L*

G. Winkler: Substituted polycrystalline YIG with very-low ferrimagnetic resonance linewidth and optical transparency.
IEEE Trans. MAG-**7**, 773-776, 1971 (No. 3). *H*

J. P. Woerdman: Some optical and electrical properties of a laser-generated free-carrier plasma in Si.
Thesis, Amsterdam 1971. *E*

A. W. Woodhead & G. Eschard: Microchannel plates and their applications.
Acta Electronica **14**, 181-200, 1971 (No. 2). (*Also in French.*) *M*, *L*

Contents of Philips Telecommunication Review **30**, No. 1, 1971:

H. da Silva, E. Goldstern & J. A. Kok: Simplex TOR STB 75 (pp. 1-14).
W. Hermes & A. Potuit: Level regulation, a decisive feature of coaxial line equipment (pp. 15-28).
L. R. Bourgonjon: Progress in telemetry applications in the ELDO launcher programme (pp. 29-35).
W. J. A. Prentice: The DS 714 as a message switcher for the overseas telegraph traffic in the United Kingdom (pp. 38-41).

Contents of Mullard Technical Communications **12**, No. 113, 1972:

J. M. Lavallee: Thyristor trigger circuit for a reversible d.c. motor drive with regenerative braking (pp. 70-79).
M. C. Gander, P. W. Bacon, D. J. Beakhust & R. P. Gant: Colour decoder using integrated circuits (pp. 80-93).
R. S. Babbs, N. F. Howarth & B. Cunningham: Transient voltage suppression using transient suppressor diodes (pp. 94-111).

Volume 32, 1971, No. 5 pages 117-144 Published 5th may, 1972

PHILIPS TECHNICAL REVIEW

VOLUME 32, 1971, No. 6/7/8

On 15th June 1972 Prof. Dr. H. B. G. Casimir retired from the Board of Management of N.V. Philips' Gloeilampenfabrieken, in which he was the member responsible for all scientific research in the Philips Group of Companies. Although the nature of our journal rules out a direct account of such an event, we feel that the retirement of Prof. Casimir — for many years the man to whom the editorial staff were responsible and a reader who has always taken an active interest in Philips Technical Review — is an occasion that merits special attention. We have therefore, unbeknown to him, prepared the present issue as a parting gift with his sphere of interests in mind and also to reflect in some way the nature and scope of his function. Some of the articles in this issue are consequently of a type that are more the exception than the rule in our pages, others on the contrary are perfectly at home. In addition to three contributions from members of the Eindhoven Research Laboratories, there are articles from each of the research laboratories abroad: Philips Forschungslaboratorium Hamburg, Philips Forschungslaboratorium Aachen, the Laboratoires d'Electronique et de Physique Appliquée at Limeil-Brévannes, the MBLE Laboratoire de Recherches at Brussels, and Mullard Research Laboratories at Redhill. There is also an article from Philips Laboratories at Briarcliff Manor, New York.

Apart from a small editorial diversion, the issue opens with some thoughts from Prof. Casimir himself, the editors having considered that it was proper in this case to let the reader whose retirement prompted the compiling of the contents contribute towards his own parting gift.

Knowledge and learned people in early times

Under this heading we present a collection of extracts from Gulliver's Travels by Jonathan Swift (London 1826). They are all taken from the chapters in part III describing Gulliver's sojourn on the islands of Laputa and Balnibarbi.

The island of Laputa floats in the air near Balnibarbi. It is suspended by the force of a giant magnet or "load-stone", which is supported by an "axle of adamant" passing through its middle. By means of the magnet the island is made to rise and fall and to move from one place to another. The magnet is under the care of astronomers, who adjust its position from time to time to suit the direction in which the king wishes the island to travel. The male inhabitants of Laputa, including the king, are so engrossed in mathematics and music that they can scarcely attend to anything else.

Balnibarbi, an ordinary island, is under the dominion of the king of Laputa; the capital of the island is Lagado, which is the seat of a Grand Academy.

Gulliver on Laputa and its inhabitants...

The Knowledge I had in Mathematicks gave me great Assistance in acquiring their Phraseology, which depended much upon that Science and Musick; and in the latter I was not unskilled. Their Ideas are perpetually conversant in Lines and Figures.

⁂

Their Houses are very ill built, the Walls bevil, without one right Angle in any Apartment; and this Defect ariseth from the Contempt they bear for practical Geometry; which they despise as vulgar and mechanick, those Instructions they give being too refined for the Intellectuals of their Workmen; which occasions perpetual Mistakes. And although they are dextrous enough upon a Piece of Paper in the Management of the Rule, the Pencil, and the Divider, yet in the common Actions and Behaviour of Life, I have not seen a more clumsy, awkward, and unhandy People, nor so slow and perplexed in their Conceptions upon all other Subjects, except those of Mathematicks and Musick. They are very bad Reasoners, and vehemently given to Opposition, unless when they happen to be of the right Opinion, which is seldom their Case.

⁂

Although I cannot say that I was ill treated in this Island, yet I must confess I thought my self too much neglected, not without some Degree of Contempt. For neither Prince nor People appeared to be curious in any Part of Knowledge, except Mathematicks and Musick, wherein I was far their inferior, and upon that Account very little regarded.

On the other Side, after having seen all the Curiosities of the Island, I was very desirous to leave it, being heartily weary of those People. They were indeed excellent in two Sciences for which I have great Esteem, and wherein I am not unversed; but at the same time so abstracted and involved in Speculation, that I never met with such disagreeable Companions. I conversed only with Women, Tradesmen, Flappers, and Court-Pages, during two Months of my Abode there; by which at last I rendered my self extremely contemptible; yet these were the only People from whom I could ever receive a reasonable Answer.

Gulliver visits the Grand Academy of Lagado...

I was received very kindly by the Warden, and went for many Days to the Academy. Every Room hath in it one or more Projectors; and I believe I could not be in fewer than five Hundred Rooms.

The first Man I saw was of a meagre Aspect, with sooty Hands and Face, his Hair and Beard long, ragged and singed in several Places. His Clothes, Shirt, and Skin were all of the same Colour. He had been Eight Years upon a Project for extracting Sun-Beams out of Cucumbers, which were to be put into Vials hermetically sealed, and let out to warm the Air in raw inclement Summers. He told me, he did not doubt in Eight Years more, that he should be able to supply the Governors Gardens with Sun-shine at a reasonable Rate; but he complained that his Stock was low, and intreated me to give him something as an Encouragement to Ingenuity, especially since this had been a very dear Season for Cucumbers. I made him a small Present, for my Lord had furnished me with Money on purpose, because he knew their Practice of begging from all who go to see them.

* * *

I had hitherto seen only one Side of the Academy, the other being appropriated to the Advancers of speculative Learning; of whom I shall say something when I have mentioned one illustrious Person more, who is called among them *the universal Artist.* He told us, he had been Thirty Years employing his Thoughts for the Improvement of human Life.

* * *

The whole Discourse was written with great Acuteness, containing many Observations both curious and useful for Politicians, but as I conceived not altogether compleat. This I ventured to tell the Author, and offered if he pleased to supply him with some Additions. He received my Proposition with more Compliance than is usual among Writers, especially those of the Projecting Species; professing he would be glad to receive farther Information.

* * *

The Professor made me great Acknowledgments for communicating these Observations, and promised to make honourable mention of me in his Treatise.

* * *

I saw nothing in this Country that could invite me to a longer Continuance; and began to think of returning home to England.

The Waalre complex of Philips Research Laboratories at Eindhoven.

Theoretical physics and industrial research

H. B. G. Casimir

Introduction

"If you see Casimir, give him my regards, and address him as Herr Direktor; I know that will annoy him." With these or similar words my great teacher Wolfgang Pauli — I was his assistant at Zürich during the academic year 1932-33 — would occasionally instruct physicists going to visit the Netherlands. And he expressed his feelings even more clearly during one of his last visits to this country. We were in a small circle of friends and someone asked whether I had not had a very difficult time at Zürich. I replied somewhat as follows: "No, not really. Pauli was on the whole rather mild. He had with considerable difficulty managed to pass the driving test and since he was then unmarried he often drove me out in the evening to have a quiet meal at some restaurant in the countryside. It was a tacit understanding that as long as I refrained from disparaging remarks about his driving he would not attack me about my physics. Now I do not want to brag about my physics, but I do think it was definitely better than Pauli's driving and that put me into a strong position." Pauli thought this over and then said: "Yes, that may have been the situation. Today I do not drive any more. And you are not doing physics any more. Die Sache stimmt noch immer. (Things still match up.)" When the editor of this journal suggested that I should write something about the experience of a theoretical physicist in an industrial laboratory this little anecdote came back to mind.

I suppose that in reality Pauli's feelings about my becoming a Herr Direktor in an industrial research laboratory were a mixture of astonishment, amusement and regret. Astonishment that a man of some understanding of theory and some mathematical ability, who had taken refuge in an industrial laboratory during the war years, should choose to stay there once the war was over, instead of returning to academic life. Amusement because of the fact that apparently his former assistant was able to hold his own in such an incongruous situation. Regret because I might have been able to do work far more valuable in his scale of values. And let us admit: an industrial laboratory is not the place for Fundamental work with a large capital F, as I said on a former occasion. The theory of relativity and cosmology, quantum mechanics, quantum electrodynamics, the detailed theory of phase transitions, including order-disorder transitions, elementary-particle and field theories — none of these things have come out of an industrial laboratory. Nor is industry particularly quick at absorbing such theories; many of the most profound and most beautiful results have so far not found any application.

Prof. Dr. W. Pauli (1900-1958)

Yet I have never felt that a theoretical physicist in an industrial laboratory is a displaced person. There are three aspects to this. First of all, basic theory provides a simple and satisfactory framework for surveying the vast wealth of experimental and theoretical problems and achievements of modern technology. This is particularly important for someone who has to deal with a variety of problems: without the scaffolding provided by theory one would soon become entangled in cumbersome detail. In my opinion a sound knowledge of basic theory is a first requirement for every research manager. To put it differently: theory is an essential tool of research management.

Secondly, although the problems presented by industry may not be very profound, they are sometimes

Prof. Dr. H. B. G. Casimir was a member of the Board of Management of N.V. Philips' Gloeilampenfabrieken until 15th June 1972 and was responsible for all research activities in the Philips Group of Companies.

quite difficult and challenging. There is a special side to this challenge. In so-called pure research we may be inclined to pick those problems we believe we can solve and avoid those that look unpromising. In industry there is less opportunity to indulge this cowardly inclination.

Thirdly, occasionally apparently mundane questions may lead back to ideas that are unexpectedly fundamental — though perhaps not Fundamental.

In the following I shall try to illustrate this situation by a number of examples. They refer to electromagnetic theory and to applications of the notions of group theory.

Electromagnetism

Knowledge of Maxwell's theory and its further development by Lorentz does not make one an electrical engineer, but it makes it possible to understand the essential features of any electric or electronic device. I should like to present a version of the theory that has served me well and that I prefer to any other formulation. I use the Gaussian system of units. In a vacuum we have:

$$\operatorname{div} \boldsymbol{E} = 4\pi\varrho , \qquad \operatorname{div} \boldsymbol{B} = 0 ,$$

$$\operatorname{curl} \boldsymbol{B} = \frac{4\pi}{c}\boldsymbol{i} + \frac{1}{c}\dot{\boldsymbol{E}} , \quad \operatorname{curl} \boldsymbol{E} = -\frac{1}{c}\dot{\boldsymbol{B}} .$$

The same equations hold in matter, but then the charge density also contains a contribution due to polarization and the current density contains atomic currents that correspond to changing polarization and to magnetism. We can write:

$$\varrho_{\text{total}} = \varrho_{\text{external}} + \varrho_{\text{polarization}},$$

$$\boldsymbol{i}_{\text{total}} = \boldsymbol{i}_{\text{conduction}} + \boldsymbol{i}_{\text{polarization}} + \boldsymbol{i}_{\text{magnetic}}.$$

This leads to:

$$\operatorname{div} \boldsymbol{E} = 4\pi(\varrho_{\text{ext}} + \varrho_{\text{pol}}) ,$$

$$\operatorname{curl} \boldsymbol{B} = \frac{4\pi}{c}(\boldsymbol{i}_{\text{cond}} + \boldsymbol{i}_{\text{pol}} + \boldsymbol{i}_{\text{magn}}) + \frac{1}{c}\dot{\boldsymbol{E}} .$$

The separation of total charge density and total current density into "external" and "internal" (or atomic) components is not always unambiguous but often there is only one obvious choice. We can write:

$$\varrho_{\text{pol}} = -\operatorname{div} \boldsymbol{P},$$

and further analysis shows that this vector $\boldsymbol{P}$ is often to a very good approximation the average value of the atomic dipole moments. Then:

$$\boldsymbol{i}_{\text{pol}} = \dot{\boldsymbol{P}}, \qquad \boldsymbol{i}_{\text{magn}} = c \operatorname{curl} \boldsymbol{M}.$$

Sometimes, especially when the choice of ϱ_{pol} and $\boldsymbol{i}_{\text{magn}}$ is unambiguous, it is useful to introduce auxiliary quantities:

$$\boldsymbol{D} = \boldsymbol{E} + 4\pi\boldsymbol{P}, \qquad \boldsymbol{H} = \boldsymbol{B} - 4\pi\boldsymbol{M}.$$

Then:

$$\operatorname{div} \boldsymbol{D} = 4\pi\varrho_{\text{ext}} , \qquad \operatorname{curl} \boldsymbol{H} = \frac{4\pi}{c}\boldsymbol{i}_{\text{cond}} .$$

Of course electromagnetic theory can also be formulated in the M.K.S. system, the only disadvantage being that then the vectors $\boldsymbol{H}$ and $\boldsymbol{D}$ already appear in the theory for empty space. This makes it necessary to put even more emphasis on the fact that these vectors are auxiliary quantities whose physical meaning is often ambiguous.

A few examples of the ambiguity:

a) In a superconducting sphere $\boldsymbol{B} = 0$. How about $\boldsymbol{H}$? We can regard the screening current as a true conduction current. Then $\boldsymbol{H}$ is also zero inside the sphere. Alternatively we can write $\boldsymbol{i} = c \operatorname{curl} \boldsymbol{M}$. Then for a sphere in a homogeneous field $\boldsymbol{B}_0$ we have $\boldsymbol{H} = \frac{3}{2}\boldsymbol{B}_0$.

b) A dielectric consists of a plastic in which a large number of small spheres of radius r are embedded. We can regard the charges on the sphere as external charges. But we can also write:

$$\boldsymbol{P} = n \,.\, r^3\boldsymbol{E},$$

where n is the number of spheres per cm^3.

c) A wave of circular frequency ω traverses a dilute plasma with n electrons per cm^3. For the electrons we have

$$m\ddot{x} = e\boldsymbol{E} ,$$

and this leads to a current

$$\boldsymbol{i} = n\,e\,\dot{\boldsymbol{x}} = -\frac{n\,e^2}{m\,\omega^2}\dot{\boldsymbol{E}} .$$

Now we can either write:

$$\boldsymbol{i} = \dot{\boldsymbol{P}}, \quad \text{with} \quad \boldsymbol{P} = -\frac{n\,e^2}{m\,\omega^2}\boldsymbol{E} ,$$

or, instead, writing $\dot{\boldsymbol{E}} = c \operatorname{curl} \boldsymbol{B}$,

$$\boldsymbol{i} = c \operatorname{curl} \boldsymbol{M}, \quad \text{with} \quad \boldsymbol{M} = -\frac{n\,e^2}{m\,\omega^2}\boldsymbol{B} .$$

The moral of these examples: in case of doubt one should analyse the atomic charges and currents. This is especially true when we are dealing with boundary conditions [1].

Notwithstanding the foregoing remarks, $\boldsymbol{D}$ and $\boldsymbol{H}$ are often very *useful* auxiliary quantities. Take the case of a permanent magnet. No conduction currents, hence curl $\boldsymbol{H} = 0$. Therefore (elementary theorem of vector analysis)

$$\int \boldsymbol{H}\cdot\boldsymbol{B}\,\mathrm{d}V = 0\,,$$

or

$$\left|\int \boldsymbol{H}\cdot\boldsymbol{B}\,\mathrm{d}V\right|_{\text{inside magnet}} = \left(\int \boldsymbol{H}\cdot\boldsymbol{B}\,\mathrm{d}V\right)_{\text{outside magnet}}.$$

Try to arrange things in such a way that $\boldsymbol{H}\cdot\boldsymbol{B}$ is at its maximum value everywhere inside the magnet. Then

$$\left(\int \boldsymbol{H}\cdot\boldsymbol{B}\,\mathrm{d}V\right)_{\text{outside magnet}} = (\boldsymbol{B}\cdot\boldsymbol{H})_{\max}\,V_{\text{magnet}}.$$

Further we can write:

$$\left(\int \boldsymbol{H}\cdot\boldsymbol{B}\,\mathrm{d}V\right)_{\text{outside}} = \left(\int \boldsymbol{B}\cdot\boldsymbol{B}\,\mathrm{d}V\right)_{\text{air gap}} + \\ + \left(\int \boldsymbol{B}\cdot\boldsymbol{B}\,\mathrm{d}V\right)_{\text{stray field}} + \left(\int \boldsymbol{H}\cdot\boldsymbol{B}\,\mathrm{d}V\right)_{\text{soft iron}}.$$

Of course this simple mathematical description does not tell us how to calculate stray fields. It does not tell us how to design a magnetic circuit in such a way that $\boldsymbol{B}\cdot\boldsymbol{H}$ is really at its maximum value. It says nothing at all about the way in which $(\boldsymbol{B}\cdot\boldsymbol{H})_{\max}$ depends on the structure of matter. But it does give a simple and straightforward classification of these problems.

In practical applications of Maxwell's equations it is often convenient to use these equations in integral form. This is well known: I have only to mention transformers and electromagnets. Other examples are the electric field in a betatron, in the accelerating sections of cosmotrons or along the beam of a klystron: one can easily determine the magnetic flux linked by an appropriately chosen contour of integration. However, many people have some difficulty in translating the differential equations into tangible reality. Consider an electromagnet with a symmetry plane where $H_x = H_y = 0$ but where H_z varies in this plane. We have:

$$\frac{\partial H_z}{\partial x} = \frac{\partial H_x}{\partial z}\,.$$

If H_z decreases in the direction of positive x then if we go outside the symmetry plane in the direction of H_z we shall find a negative H_x. The normals to the lines of force point in the direction of increasing H_z. This is essential in the theory of vertical stability of orbits in cyclotrons and betatrons.

A useful theorem says that if a quantity satisfying Laplace's equation is averaged over the surface of a sphere one obtains the value of that quantity at the centre of the sphere. (Proof: expand in spherical harmonics; all spherical harmonics of order $\geqslant 1$ have average 0; only the constant remains.) This can be used in some numerical procedures. It also plays a role in calibration of magnetic fields: a uniformly wound spherical test coil measures exactly the value at its centre, however inhomogeneous the field.

Incidentally, the fact that the average of the second spherical harmonic is zero, $\overline{(1 - 3\cos^2\phi)} = 0$, has an amusing physical interpretation. If two dipoles at arbitrary points are oriented by a homogeneous magnetic field and this field rotates through all directions then the average interaction energy is zero. I have suggested that this might be used for dispersing magnetic powders but I do not think it is a very practical proposal.

I could ramble on like this for page after page — and perhaps some day I will, in the context of a book — but that is not my purpose on this occasion. I only wanted to show that in an industrial laboratory like Philips Research Laboratories one meets striking illustrations of electromagnetic theory at every step.

And not only simple cases. There are many problems that tax the mathematical abilities of the physicist and sometimes also the capability of a large computer and the skill of the programmer. In the course of the years Philips Research Laboratories have made many contributions in this field. I might mention the work of H. Bremmer and B. van der Pol on the diffraction of radiowaves around the Earth or the work of C. Bouwkamp on antennae and on diffraction through holes. I should also like to pay tribute to the many ingenious calculations that G. de Vries generously made available to his colleagues without ever bothering about publishing them.

In this connection I should like to mention briefly the theory of eddy currents in ferromagnetic materials. The late J. L. Snoek had pointed out that there is a limiting case where the skin depth is small compared with the dimensions d of the sample although μd may have any value. I found [2] that in this case the field can be derived from a solution of the Laplace equation with the curious boundary condition

$$\frac{\partial\Phi}{\partial n} - \frac{\mu d}{1+\mathrm{j}}\left[\frac{\partial^2\Phi}{\partial n^2} + \left(\frac{1}{R_1} + \frac{1}{R_2}\right)\frac{\partial\Phi}{\partial n}\right] = 0\,,$$

where R_1, R_2 are the main radii of curvature.

As to the third aspect of theoretical work in an industrial research laboratory I hope I will be forgiven if I draw upon my personal experience. Studying standing waves in cavities [3] in connection with centimetric waves taught me to regard such modes as very real physical entities, much more so than I did in formal radiation theory. Also, this study provided me with simple formulae for dealing with small perturbations. This gave me the courage to study the zero-point energy of these modes, $\Sigma\frac{1}{2}h\nu$. The sum is wildly divergent but it is possible to assign a uniquely defined value

[1] See H. B. G. Casimir, in: Polarisation, Matière et Rayonnement, Volume jubilaire en l'honneur d'Alfred Kastler, Presses Universitaires de France 1969, p. 185, and H. B. G. Casimir, Philips Res. Repts. 21, 417, 1966.

[2] H. B. G. Casimir, Philips Res. Repts. 2, 42, 1947.

[3] H. B. G. Casimir, Philips Res. Repts. 6, 162, 1951.

to the difference between the (undefined) values of this sum in different conditions. This leads to the prediction [4] of a universal attractive interaction between two metal plates at distance a and with area l^2:

$$U = -\frac{\pi^2}{720}\hbar c\frac{l^2}{a^3}.$$

A prediction that has been confirmed by experiment. It also led to a simplified deduction [5] of the formulae for retarded Van der Waals forces, derived by D. Polder and myself [6] in answer to a question from J. Th. G. Overbeek arising out of the latter's work on the stability of colloidal suspensions.

Of course this one example might be supplemented by many others.

Group theory

My second series of examples is held together by the notions of group theory, that is the theory of the structure and the representations of groups of transformations. The formal theory has played an important part in the development of quantum mechanics; the general ideas provide an excellent unifying pattern of thought. Unfortunately this mathematical discipline may be unfamiliar to many of my readers and this is not the place to explain it in any detail. I hope that my sketchy indications will suffice to show the general tenor.

Units and Abelian groups

In physics we have to deal with two classes of equations. On the one hand those with mathematical identities, such as

$$\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x\prod_{n=1}^{\infty}\left(1-\frac{x^2}{n^2\pi^2}\right),$$

which holds for any (real or complex) x, on the other hand those with equations that describe properties of nature. For instance the equation

$$c^2 = a^2 + b^2 \tag{1}$$

may be said to express a property of a plane rectangular triangle, an interesting fact about the square of the hypothenuse. In this equation a, b and c express length measured in a certain unit. If we change the unit by a factor α, then a, b and c will undergo a transformation:

$$a' = \alpha a, \qquad b' = \alpha b, \qquad c' = \alpha c.$$

All possible transformations of this kind form an Abelian group and eq. (1) is invariant under this group.

If we state that the volume of a sphere is given by

$$V = (4\pi/3)R^3, \tag{2}$$

this is true if volume is measured in cubic units of length. Equation (2) is invariant under a transformation

$$V' = \alpha^3 V, \qquad R' = \alpha R.$$

But we may also introduce an independent unit of volume like a gallon or a barrel or a whiba. Then we have to write

$$V = v_0\frac{4\pi}{3}R^3,$$

where v_0 is the volume of a cube with an edge of one unit of length. This equation admits a wider group of transformations:

$$V' = \beta V, \qquad R' = \alpha R, \qquad v_0' = (\beta/\alpha^3)v_0.$$

This example shows that we have a certain measure of freedom: we can define our quantities in different ways and depending on such definitions our equations become invariant under a wider or a more restricted group of transformations. On the whole the choice is a matter of convenience. Electromagnetic units [7] provide an interesting example.

Abelian groups; linear displacement

A crystal lattice with lattice vectors u_1, u_2, u_3 permits a displacement operation of the general form

$$D = u_1^{n_1}u_2^{n_2}u_3^{n_3},$$

where the u_i commute. The Ds form an infinite (but denumerable) Abelian group. It follows (Bloch's theorem) that the wave function of an electron moving in such a lattice has the form

$$\psi = e^{i(kx)}p(x), \tag{3}$$

where $p(x)$ is periodic; $p(Dx) = p(x)$.

A multicavity magnetron permits rotations R, R^2, ... If there are N cavities then $R^N = E$. It follows that the field must be a superposition of modes in which the field in cavity m is given by

$$F_m = e^{2\pi iqm/N}f_q(x), \text{ with } q = 0, 1, \ldots N{-}1, \tag{4}$$

where $f_q(x)$ depends on q but is the same in all cavities.

It is obvious that these two examples are very analogous: in both cases the situation is invariant under a certain Abelian group, the elements of which are translations through a lattice vector $n_1u_1 + n_2u_2 + n_3u_3$ in the first case, rotations over an angle $m.2\pi/N$ in the second one; this determines the general form of the field.

Similar considerations hold for linear chains, elastic vibrations in a crystal or in a benzene molecule, etc., etc.

Rotation group

The theory of irreducible representations of the three-dimensional rotation group applied to atomic wave functions provides an elegant mathematical explanation of the vector model of the atom. In electromagnetic theory it is useful to know that spherical harmonics are a base for these irreducible representations. This is also true for multipole radiation fields [8] but there is a difference from the scalar case. In the case of a scalar potential a $(2l+1)$-dimensional representation has necessarily an inversion factor $(-1)^l$. In the case of vector fields the inversion factor for a $(2l+1)$ dimensional representation can be either (+) or (—). In electrical 2^l-pole fields, the E have an inversion factor $(-1)^l$, the H an inversion factor $(-1)^{l+1}$; for magnetic multipole fields it is just the other way round.

This type of group theoretical considerations does not obviate the need for explicit calculations. However, group theory predicts many general properties of the explicit solutions. This leads to considerable simplification and provides also a useful check on the results.

Symmetry and integrals of the equations of motion

If a Hamiltonian permits a continuous group of transformations then the quantities generating the infinitesimal canonical transformations of this group are integrals of the equations of motion. This profound and useful theorem holds in quantum mechanics as well as in classical mechanics but our first example is decidedly classical. Watch a figure skater on the ice. The field of force is symmetrical around the vertical axis. So is the ice, so is the whole situation. Therefore the quantity generating the infinitesimal transformation corresponding to a rotation around the vertical is an integral of the motion: the angular momentum. Since angular momentum is the product of moment of inertia and angular velocity, angular velocity can be controlled by changing the moment of inertia. A figure skater spinning round in an ever faster pirouette is not only giving a show of elegance but also a demonstration of the isotropy of space. Similar considerations may be applied to the Einstein-De Haas effect [*]; there also we have conservation of total angular momentum.

There exists an analogous connection between linear displacement and conservation of momentum; not much needs to be said about this. But a much more interesting situation arises if we study phenomena in a crystal at rest. Are the Hamiltonians we use invariant under an infinitesimal translation with respect to the crystal? If so, than there must exist a corresponding "pseudo-momentum" that is conserved, if not, then no such conclusion can be drawn. And the answer is affirmative only when the crystal may be regarded as a continuum. Now consider elastic waves. In a continuum, whatever the non-linearity that leads to interaction between waves, there exists a constant pseudo-momentum. But a crystal is *not* exactly a continuum, therefore this pseudo-momentum is not exactly conserved. This is the gist of Peierls's theory of conduction of heat; his "Umklapprozesse" express the difference between a true continuum and a crystal lattice.

Time reversal

The equations of electrodynamics and mechanics are invariant under a substitution $t \rightarrow -t$. This is true in classical theory and in quantum theory. But the behaviour of macroscopic matter is in no way reversible although matter consists of atoms for which invariance under time reversal should hold. The answer to this apparent contradiction is given by statistical mechanics. J. Onsager has shown that, although the paradox can be resolved, the reversibility of the atomic processes is reflected in certain macroscopic symmetry properties.

Onsager's theory provided an unexpected link between B. D. H. Tellegen's considerations on gyrators, the Thomson relations between thermoelectric quantities, the symmetry of conductivity tensors and many other phenomena.

Studying the connection between network theory and Onsager's ideas gave rise to a slight extension of Onsager's formalism [9], one more example of how practical applications may lead back to fundamentals.

Conclusion

This has become a very egotistic article. In trying to explain how a theoretician reacts to the problems of an industrial laboratory I have overstressed my particular fads and fancies, paid too much attention to my own contributions, and done scant justice to the ideas and achievements of others.

Perhaps this is understandable. But before I finish I should emphasize that in industrial research the type of understanding that is provided by fundamental theory is only a first step, not the final one. It is a long chain of assiduous efforts that leads to a new product or a new process. Industrial innovation calls for inventiveness and technological skill, for accuracy, perseverance and organizational ability and for all of these I want to express my grateful admiration to my co-workers at Philips Research Laboratories.

[4] H. B. G. Casimir, Proc. Kon. Ned. Akad. Wetensch. **51**, 793, 1948.

[5] H. B. G. Casimir, J. Chimie phys. **46**, 407, 1949.

[6] H. B. G. Casimir and D. Polder, Phys. Rev. **73**, 360, 1948.

[7] H. B. G. Casimir, On electromagnetic units, Helv. Phys. Acta **41**, 741-742, 1968.

[8] H. B. G. Casimir, Helv. Phys. Acta **33**, 849, 1960.

[9] H. B. G. Casimir, Rev. mod. Phys. **17**, 343, 1945, or Philips Res. Repts. **1**, 185, 1946.

[*] The Einstein-De Haas effect is the phenomenon whereby an iron rod which is being magnetized (or demagnetized) is subject to a small torque. This is due to the spin of the electrons, which gives them an angular momentum as well as a magnetic moment *(Ed.)*

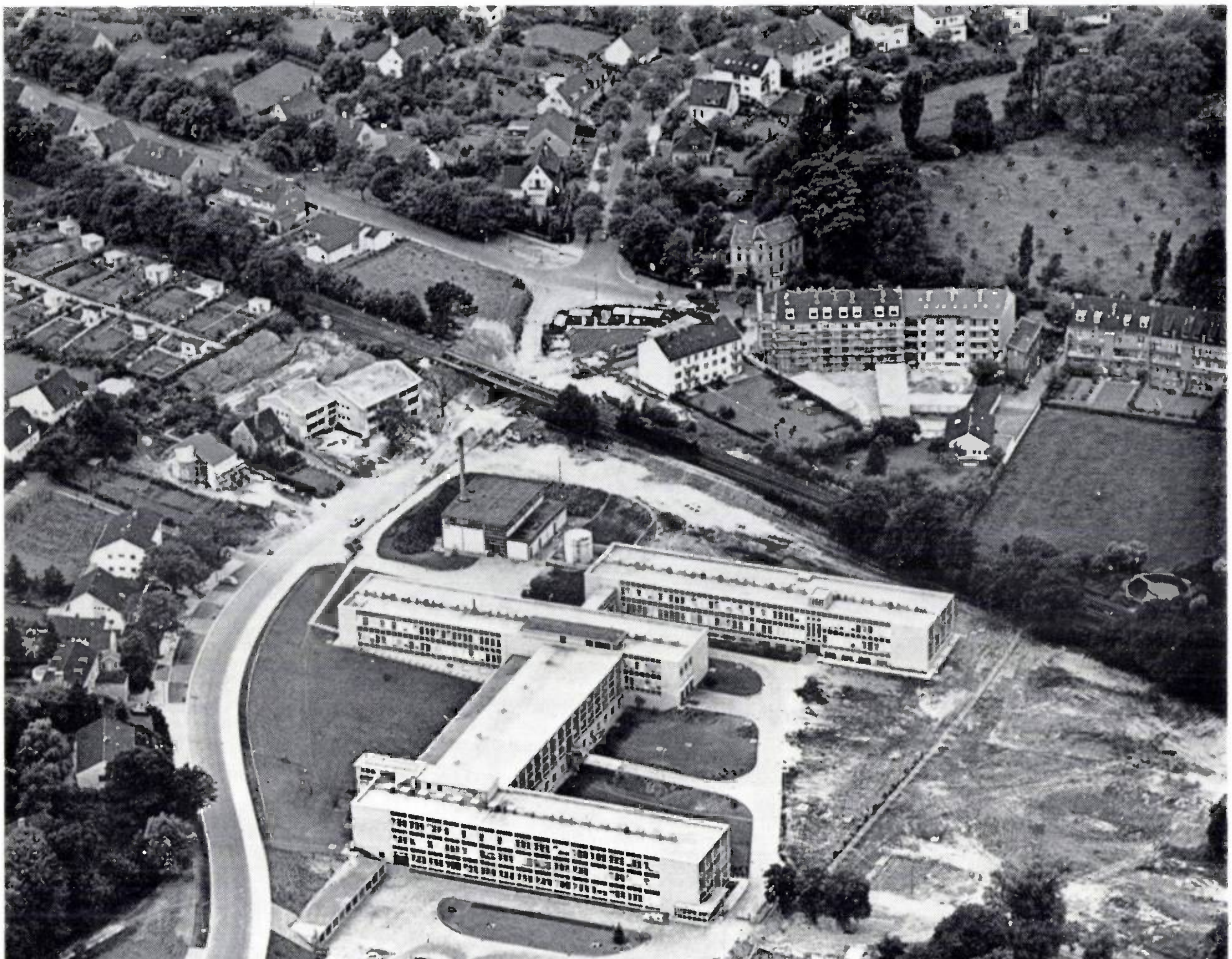


Photo Aero-Foto: Freigabe-Nr. Reg. Präs. Düsseldorf 831/06.

Philips Forschungslaboratorium Aachen.

The burn-out mechanism of incandescent lamps

H. Hörster, E. Kauer and W. Lechner

Introduction

Even after 60 years of mass production there still remains a great deal of uncertainty about physical processes in the incandescent light bulb. One question that remains unanswered is that of which process controls the life of the lamp. It is generally known that the burn-out mechanism is caused by the evaporation of tungsten from the wire, but this explanation is insufficient. This can be shown by considering a homogeneous incandescent wire, burning at constant voltage with an exactly uniform temperature distribution and hence a uniform rate of evaporation along its axis. Under these conditions the wire diameter would decrease and its resistivity would increase to cause a continuous decrease in temperature. This would finally result in an infinite life. This is not observed. Instead, it is found that the wire burns through at a point in a finite time, during which evaporation losses from the rest of the wire are small.

This can be explained by the presence of defects, the so called *hot spots*, which influence the local energy balance of the wire to produce a local increase in temperature and, hence, in evaporation. This results in an intensified decrease in diameter because the current is controlled by the defect-free part of the wire. In this way the local increase in temperature of the spot is amplified as a consequence of the cycle: evaporation → decrease in diameter → increase in temperature → additional evaporation losses etc. This concept was used in 1925 by R. Becker [1] to give a qualitative interpretation of the life of an incandescent wire.

In the first part of this article the spot model is developed to obtain quantitative expressions for the life and other fundamental properties of an incandescent wire *in vacuo* by an analysis of the nature, the magnitude and the subsequent development of the defects [2].

In the second part the model is applied to the gas-filled lamp. Here the mass transport is controlled by diffusion of tungsten atoms through the inert gas. For a straight wire the model remains valid. In the case of a coil, however, there are pronounced changes such as the reduction of the quantity known as the "deadly weight loss". This requires the introduction of a new model which takes into account both radial and axial tungsten diffusion [3]. These two quantities are compared and the condition for which the axial mass flow overrides the radial mass flow is discussed. In the case of a low-voltage lamp the axial mass flow proves to be the life-determining process [4].

Burn-out mechanism of a wire *in vacuo*

Formulation of the spot model

The burn-out mechanism of an incandescent wire can be understood by assuming the existence of a local disturbance of the energy balance of the wire. This process can be studied in detail by considering the behaviour of a long straight wire. In the model (*fig. 1*) the wire has a uniform temperature T_1 with the exception of a local disturbance, usually called a spot, where the temperature is T_2. This temperature differs only slightly from T_1. In fig. 1 it is assumed that this disturbance is caused by a decrease of diameter.

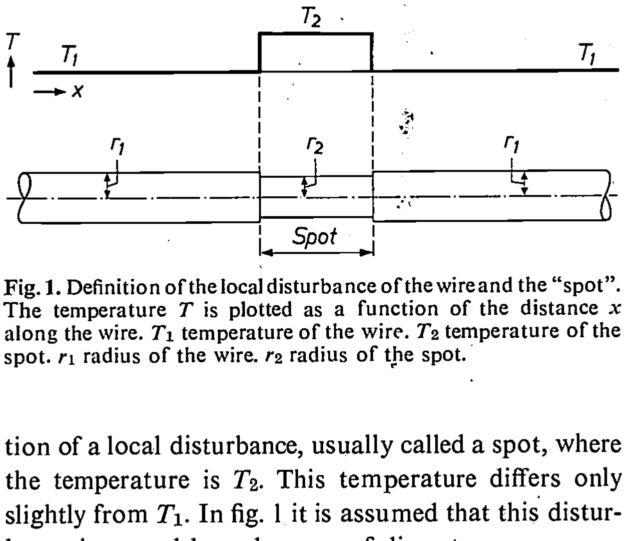


Fig. 1. Definition of the local disturbance of the wire and the "spot". The temperature T is plotted as a function of the distance x along the wire. T_1 temperature of the wire. T_2 temperature of the spot. r_1 radius of the wire. r_2 radius of the spot.

To describe the life and other important physical quantities it is necessary to evaluate the development of the wire temperature T_1 and the spot temperature T_2 with time. For this purpose it is assumed that all interesting quantities can be described by the temporal decrease of the wire diameter only. This is determined by evaporation losses, which are only a function of

Prof. Dr. E. Kauer is Director of Philips Forschungslaboratorium Aachen GmbH (PFA), Aachen, Germany; Dr. H. Hörster and W. Lechner are also with PFA.

[1] R. Becker, Z. techn. Physik **6**, 309, 1925.

[2] An article by the authors on this subject will be published in the July issue of J. of IES 1972. A paper was presented at the Annual IES Conference, August 17-20, 1971, Chicago, Ill., U.S.A.

[3] The same concept has been briefly mentioned by F. Blau; it is quoted by F. Koref and H. C. Plaut, Z. techn. Physik **11**, 515, 1930.

[4] See also B. A. Moys, Trans. Illum. Engng. Soc. 29, 13, 1964.

temperature. Starting from the energy balance of the defect-free wire the dependence on time t of the temperature $T_1(t)$ and of the current $I(t)$ is calculated with the assumption that the resistance of the spot is negligible in comparison with the resistance of the whole wire. In other words, for an incandescent wire subjected to a constant voltage the current $I(t)$ is only a function of the electrical resistance of the unaffected portion of the wire, and, consequently, of its temporal development. Hence with a knowledge of the current as a function of time $I(t)$, the time dependence of the spot temperature can be calculated. The life τ is defined as the time at which the temperature $T_2(t)$ has reached the melting point of tungsten. It is further assumed that the heat conduction between the spot and the unaffected part of the wire has no influence on the development of the spot temperature, i.e. it is supposed that the length of the spot is large in comparison to the diameter of the wire.

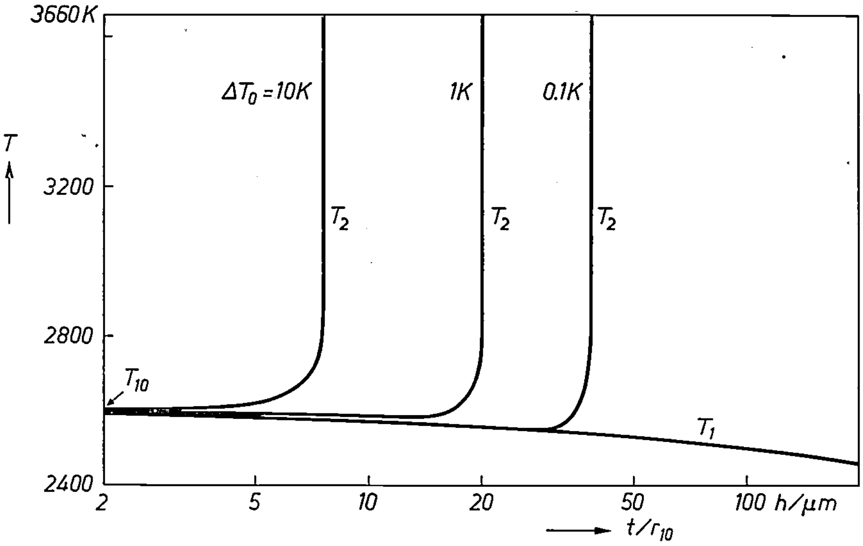


Fig. 2. Time dependence of the temperature T_1 of the defect-free wire and T_2 of the spot for constant voltage and for a temperature T_{10} of the defect-free wire of 2600 K at the beginning of the life. The temperatures are ploted against the ratio of the time t to the radius r_{10} of the wire at the beginning of the life. The three curves T_2 apply for different values of the difference ΔT_0 between the spot temperature and the wire temperature T_{10} at the beginning of the life.

The time dependence of the temperatures is obtained as a solution of two differential equations in T_1 and T_2. The solutions have a rather complex form. Essentially they depend on the evaporation rate $m(T)$, the wire radius r and the initial temperature difference between spot and defect-free wire, ΔT_0. They can be expressed in the following form:

$$T_1(t) = \mathrm{f}\{m(T_{10}), r_{10}, T_{10}\}, \qquad (1)$$

$$T_2(t) = \mathrm{f}\{m(T_{10}), r_{10}, T_{10}, \Delta T_0\}, \qquad (2)$$

where r_{10} and T_{10} are the initial values of the radius and the temperature of the defect-free wire at the beginning of the life.

The solutions for $T_1(t)$ and $T_2(t)$ at a starting temperature of 2600 K are given as a logarithmic function of a ratio of the burning time t to r_{10} in *fig. 2.* With this form of representation the curves shown in fig. 2 can be applied to a wire of any given radius. The parameter ΔT_0, the difference between the initial temperatures of the spot and the defect-free wire is a measure of the initial defect. $T_1(t)$ takes the form which is to be expected for a defect-free wire, i.e. evaporation losses cause a decrease in diameter, which at constant voltage leads to a reduction in temperature. The temperature-time dependence $T_2(t)$ is controlled by the temperature difference ΔT_0. For a given ΔT_0 (e.g. 0.1 K), $T_2(t)$ deviates only slightly from $T_1(t)$ over a large period. This is then followed by an extremely steep increase of the spot temperature $T_2(t)$, which reaches the melting point of tungsten at 3655 K. The life τ of the wire is then attained. Fig. 2 shows that τ depends very strongly on the initial temperature difference ΔT_0.

By substituting the melting point of tungsten in the equation for the spot temperature $T_2(t)$ an analytical expression for the life τ is obtained. It is found that the life is proportional to the wire radius, and inversely proportional to the evaporation rate of the defect-free wire and a function of the initial temperature difference ΔT_0. The exact relationship is given by:

$$\tau = C \frac{r_{10}}{m(T_{10})} \left[\left\{ 1 - \left(1 + \frac{\Delta T_0}{T_{10}}\right)^{-a} \right\}^{-b} - 1 \right], \qquad (3)$$

where $m(T_{10}) = m_0 T_{10}^{\gamma}$ is the evaporation rate, expressed as a power function. This power function is an approximation of the exponential temperature-dependent evaporation losses $m = m_0' \exp(-Q/RT)$, with Q the heat of evaporation equal to 205 kcal/mol[5]. The approximation is obtained by making the derivatives of both equations (power function and exponential function) equal at a given temperature. Thus from $\gamma = Q/RT$ and $T = T_{10} = 2600$ K we obtain $\gamma = 40$, $a = \frac{1}{3}(4 + 3\gamma) = 41.3$ and $b = (\gamma - 4)/(4 + 3\gamma) = 0.29$. C is a constant. Thus, if the model is correct, the life depends only on the initial value of the temperature T_{10}, the wire radius r_{10}, and the defect ratio $\Delta T_0/T_{10}$. Experiments show that relation (3) describes the dependence of life on temperature and wire radius.

Having found an expression for the life it is now possible to obtain an analytical expression for the "deadly weight loss" of a filament. This is defined by the ratio of the evaporation losses ΔW of a filament during its life to its initial mass W_0. This is given by:

$$\frac{\Delta W}{W_0} = \frac{2\pi l}{W_0} \int_0^{\tau} r_1(t)\, m\{T_1(t)\}\, \mathrm{d}t\,, \qquad (4)$$

in which l is the length of the wire.

Previously the deadly weight loss could only be determined experimentally. However, because of its independence of temperature and wire radius it is an interesting physical quantity. For a given type of lamp (e.g. vacuum lamps) it is constant within a limited range. By using (3) one obtains from (4):

$$\frac{\Delta W}{W_0} = 1 - \left\{1 - \left(1 + \frac{\Delta T_0}{T_{10}}\right)^{-a}\right\}^{c}, \qquad (5)$$

where $c = 8/(4 + 3\gamma) \approx 0.065$ for $\gamma = 40$. Surprisingly $\Delta W/W_0$ only depends on the initial relative defect $\Delta T_0/T_{10}$. For the burning-temperature range of interest (2500-3000 K) this is independent of T_{10}. Thus equation (5) is in agreement with the experimentally determined independence of wire radius and burning temperature, i.e. the deadly weight loss is independent of the temporal development of the current and voltage. Hence, it is an ideal measure of quality which is moreover superior to life determination. The life method depends strongly on the burning temperature T_{10} and hence on the applied voltage. The effect of this is that small changes in the burning temperature conceal the influence of ΔT_0.

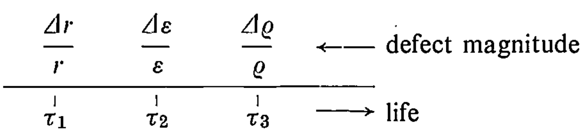
Magnitude of the initial defects

The deadly weight loss is a quantity which can be easily measured. Experiments performed on 50 vacuum lamps (110 V, 75 W) yielded a mean value for $\Delta W/W_0$ of 15%. By substituting this value in equation (5) an initial temperature difference ΔT_0 of 5 K was obtained. This is a surprisingly small value. This small temperature difference can be caused by a number of independent defects which give rise to a local disturbance of the energy balance.

From the energy balance it is found that

$$\frac{\Delta T_0}{T_{10}} = -\frac{3}{4}\left(\frac{\Delta r}{r}\right)_{\varrho,\varepsilon} + \frac{1}{4}\left(\frac{\Delta\varrho}{\varrho}\right)_{r,\varepsilon} - \frac{1}{4}\left(\frac{\Delta\varepsilon}{\varepsilon}\right)_{r,\varrho} + \dots \quad (6)$$

Thus, the relative temperature difference $\Delta T_0/T_{10}$, which is equal to 5 K/2600 K or 0.2%, can be caused by a local decrease in the wire diameter $\Delta r/r$ of only 0.26%, a local increase in the electrical resistivity $\Delta\varrho/\varrho$ of 0.8%, or a local decrease of the emissivity $\Delta\varepsilon/\varepsilon$ of 0.8%. This is by no means a complete list of possible defects. For instance in the case of a coil there can be errors in the pitch. The electrical resistivity can be influenced by grain boundaries, small holes, doping etc. In fact quite a number of different types of defects can control the above-mentioned quantities. In general it is not possible to determine the influence of all these different types of defects on the local energy balance. The situation can be depicted by a life axis on which the various life-determining defects are related to their equivalent life.


$\frac{\Delta r}{r}$
$\frac{\Delta\varepsilon}{\varepsilon}$
$\frac{\Delta\varrho}{\varrho}$
defect magnitude
τ_1
τ_2
τ_3
life

The largest defect is responsible for the observed life, and characterizes the technical standard. When this is eliminated the next largest defect determines the life and the life increases. One question of interest is the difference in magnitude between neighbouring defects, in other words: is the life determined by a single type of defect? In the deadly-weight-loss experiments a spread of ± 2% was observed. This corresponds to a $\Delta(\Delta T_0)$ of ± 1.5 K. From this small deviation it is concluded that only one type of defect is responsible for the observed life.

In the derivation of the spot model it was assumed that axial heat conduction along the wire had no influence on the spot temperature T_2. Hence, the values discussed above of the defects are the smallest. A more exact analysis of the problem shows that with a burning temperature T_{10} of 2600 K and a wire radius of 50 μm the approximation of the simple model is only valid for defects greater than $60 r_1 = 3$ mm. If the defect length is smaller the defect itself must be larger to maintain the temperature difference ΔT_0. For example, a defect

[5] O. Kubaschewski, E. Ll. Evans and C. B. Allcock, Metallurgical Thermochemistry, 4th edn., Pergamon Press, Oxford 1967.

length of 5 μm requires a decrease in diameter of 24% to maintain the temperature increase of 5 K. Other defects must have the same value. Defects of this size are never observed experimentally. Thus individual defects of such small length obviously do not determine the life. However, the situation is changed if a large number of these defects are present along the axis of the wire. Then, if the deviation of their average magnitude along a wire length of about $60r_1$ exceeds the value corresponding to the case of no heat conduction, they can determine the life.

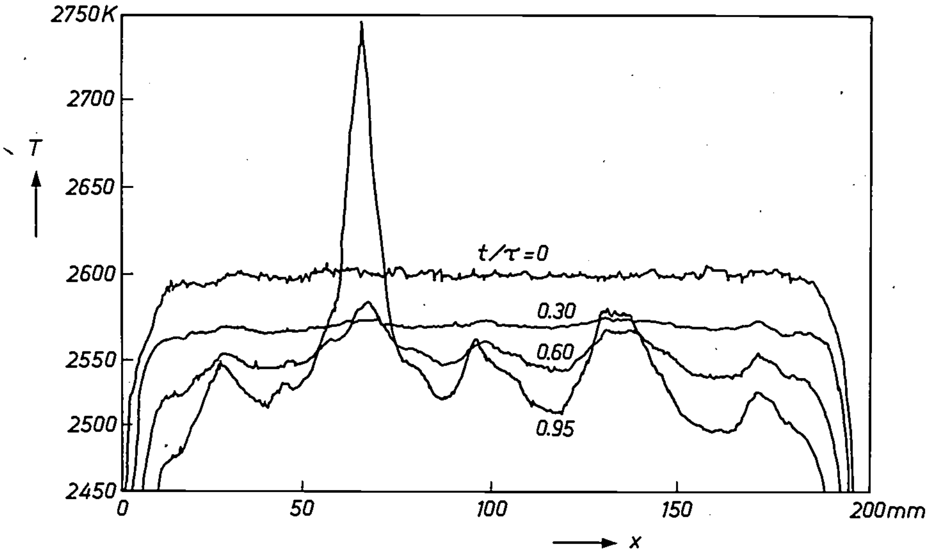


Fig. 3. Temperature distribution along a conventionally produced tungsten wire at different stages of the life, the time t for which the lamp has burnt being expressed as a fraction of the life τ.

Comparison with experimental results

The spot model is based on the assumption that the critical defect is present at the beginning of the life. This assumption was investigated experimentally. Furthermore it was attempted to identify the life-determining defect.

An electronic micropyrometer was developed which measured the surface brightness of a circular area of diameter 10 μm. Assuming constant local emissivity the signal is proportional to the true temperature of the wire. The sensitivity is about 0.5 K. By means of a specially designed device differences in diameter of 0.1 μm could be registered. In *fig. 3* the temperature distribution along a conventionally produced wire at different stages of its life is shown. Shortly before burn-through a strong local temperature increase is apparent. In *fig. 4* the temperature distribution of the same tungsten wire is compared to the local variation of the wire diameter. The striking resemblance of the shape of the two curves confirms the assumption in the spot model that the spot temperature at the end of the life is only controlled by the temperature-dependent evaporation losses.

In fig. 3 it can be seen that the final temperature distribution can only be determined after about 50% of the life. At first sight this seems to be in disagreement with the model. Moreover, initial temperature differences seem to be greater than the predicted ones. The explanation for these two effects is that the local emissivity of the wire varies. This arises from multiple reflections caused by surface roughness and leads to an apparent increase in temperature. The influence of this effect on the pyrometer signal is enhanced by the small area of measurement. The evaporation rate is however controlled by the real temperature of the wire. To

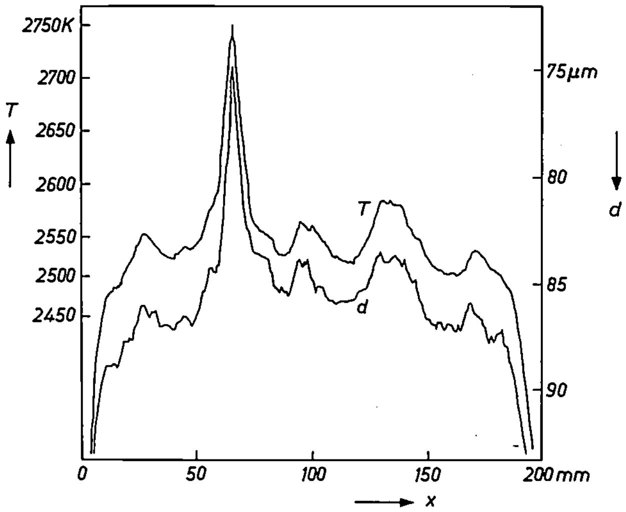


Fig. 4. Comparison of the temperature (T) and the diameter (d) distribution along a tungsten wire at the end of the life.

eliminate the influence of the roughness very smooth wires were produced. In *fig. 5* the cross-section of a conventional and a smoothed wire are compared. The temperature profile of a smoothed wire is given in *fig. 6*. The initial temperature differences are considerably decreased (compare with fig. 3) and moreover it is possible to recognize in curve $t/\tau = 0$ the position at which the temperature is a maximum and at which the spot develops. This indicates that by diminishing the surface roughness the initial true temperature distribution of a smooth wire can be measured. By increasing the area from which the pyrometer signal is derived, from 5 μm to a wire surface of length 12 r_1, the influence of the local surface roughness of a rough wire can be eliminated. The signal then corresponds to the true temperature. In *fig. 7* the temperature distribution of a rough wire measured by this method is compared with that of a smoothed wire at the beginning of the life. A comparison of both curves shows that the smoothing process considerably reduces the initial temperature differences. In the lower part of fig. 7 the corresponding diameter profiles are shown. These clearly again indicate the reduction of the surface roughness. From this it is to be expected that a smoothed wire with its smaller initial temperature differences should have a longer life and its deadly weight loss should increase. The experiment showed that the deadly weight loss $\Delta W/W_0$ of a conventional wire is about 18% and that of the smoothed wires was increased up to 32%, i.e. it is possible to obtain a relative increase of $\Delta W/W_0$ and subsequently in $\Delta\tau/\tau$ of about 70%, which signifies a considerable increase in the quality of incandescent lamps.

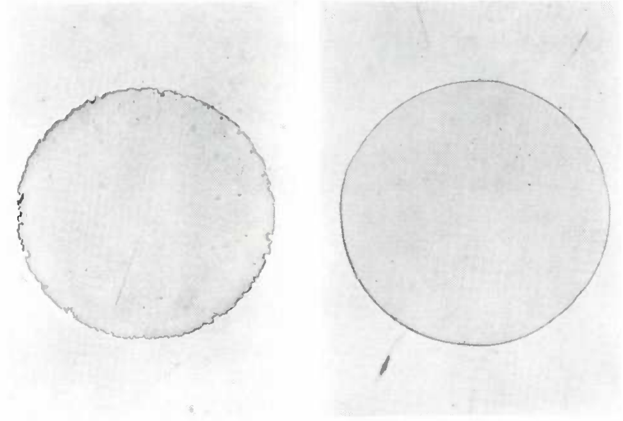


Fig. 5. Micrograph of the cross-section of a conventionally produced tungsten wire (*left*, diameter 118 μm) and of a smoothed tungsten wire (*right*, diameter 105 μm).

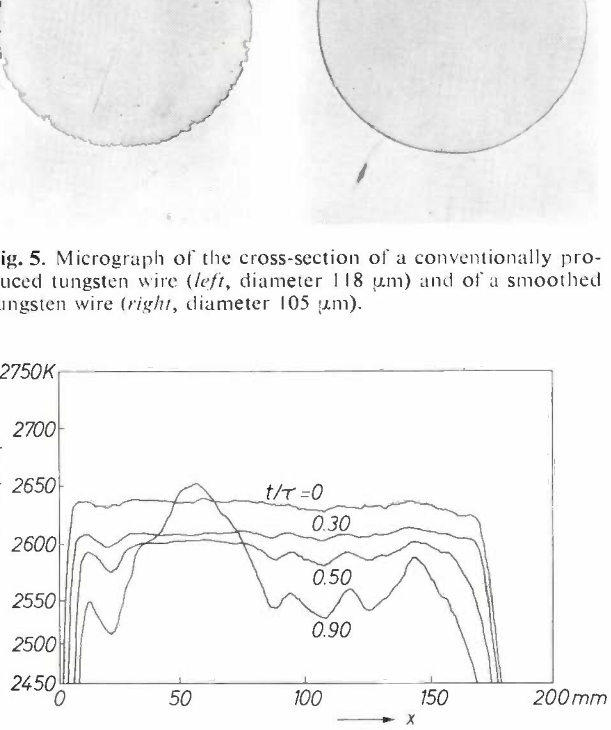


Fig. 6. As fig. 3, but now for a smoothed tungsten wire at different stages of the life. The initial temperature differences are much smaller and the spot can already be recognized in the curve for $t/\tau = 0$, at the beginning of the experiment.

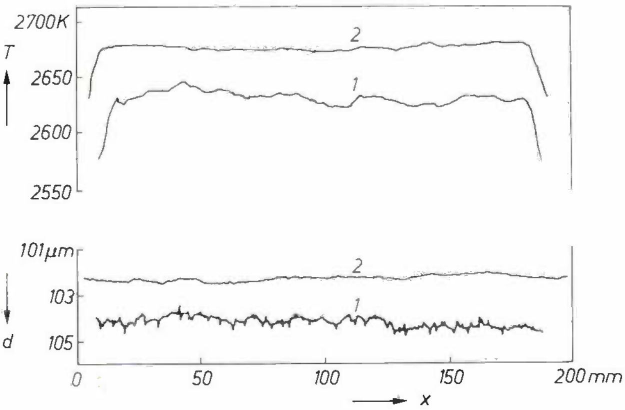


Fig. 7. Comparison of the temperature and of the diameter distribution along a conventionally produced tungsten wire (curves *1*) and a smoothed tungsten wire (curves *2*) at the beginning of the life.

From these experiments it can be concluded that the life-determining process of conventionally produced wires is the surface roughness which controls the wire diameter and the emissivity. This can be deduced from the fact that the smoothing process affects only the surface and not the crystal structure, doping concentration or other internal properties.

Life and weight loss of a coil in vacuo

By coiling a straight wire the free surface area, i.e. the area responsible for mass and energy losses, is reduced. The inner surface of a coil can be disregarded in the energy- and mass-loss equations because the inner part of the coil acts approximately as a Knudsen cell and as a black-body radiator. For vacuum coils the free surface is about 50% of the corresponding wire surface S_{wire}.

It can be shown that the progressive decrease of the wire diameter of a coil compared with that of a straight wire is reduced in the ratio $S_{\text{coil}}/S_{\text{wire}} \approx \frac{1}{2}$. Another interpretation is that the evaporation rate of a coil compared with that of a straight wire is reduced in the same ratio. This is in agreement with evaporation experiments performed by G. R. Fonda [6]. In terms

[6] G. R. Fonda and A. A. Vernon, J. Opt. Soc. Amer. 22, 223, 1932.

of the spot model this extends the temperature-time dependence of the spot T_2 and of the defect-free wire T_1 by the ratio $S_{\text{coil}}/S_{\text{wire}}$, see fig. 2. Consequently this leads to an increase in life by the ratio $S_{\text{wire}}/S_{\text{coil}}$.

The deadly weight loss of a coil compared to a straight wire can be calculated from the defining equation:

$$\left|\frac{\Delta W}{W_0}\right|_{\text{coil}} = \frac{1}{W_0}\int_0^{\tau_{\text{coil}}} S_{\text{coil}}\, v\{T_1(t)\}\mathrm{d}t =$$

$$= \frac{1}{W_0}\int_0^{\frac{S_{\text{wire}}}{S_{\text{coil}}}\cdot\tau_{\text{wire}}} \frac{S_{\text{coil}}}{S_{\text{wire}}}\, S_{\text{wire}}\, v\{T_1(t)\}\mathrm{d}t = \left|\frac{\Delta W}{W_0}\right|_{\text{wire}}. \quad (7)$$

This result is in agreement with the experimental work of Fonda [6]. After consideration is taken of the reduced evaporation loss caused by the supports it can be shown that the results presented in the previous sections are also in agreement with this equation.

The fact that the deadly weight loss should be the same for both coil and wire could be used to investigate any additional defects introduced by coiling, i.e. any defect caused by the coiling process would be revealed by a decrease in deadly weight loss.

Burn-out mechanism of gas-filled lamps

Life and weight loss of a straight wire in an inert gas

We now consider the behaviour of a straight wire in an inert-gas atmosphere in much the same way as the vacuum case. In this case it is necessary to modify the spot model to take into account the change in the energy balance due to heat dissipation into the gas and also its influence on the tungsten evaporation rate. The contribution of the heat losses of the gas is of the order of 27% at a wire temperature of 2800 K and requires an additional current of 19% in this example to maintain the same operating temperature. On the other hand the change from free evaporation to diffusion-controlled mass transport is a much more pronounced effect. This mass transport M can be described by the following equation:

$$M = -D \operatorname{grad} n + nv, \quad (8)$$

where D is the diffusion coefficient of tungsten in the inert gas, n the tungsten concentration, and v the convection velocity of the inert gas. The complete solution of this equation requires a knowledge of the velocity field, which in general is not known. An approximate solution of this problem has been suggested by Fonda [7] and W. Elenbaas [8]. By assuming the existence of a stationary gas film surrounding the filament the problem can be treated as one of concentration diffusion: the boundary conditions being given by the saturation concentration of the tungsten at the wire surface $n(T_1)$, and $n = 0$ at the film boundary. Using this approach the following solution is obtained:

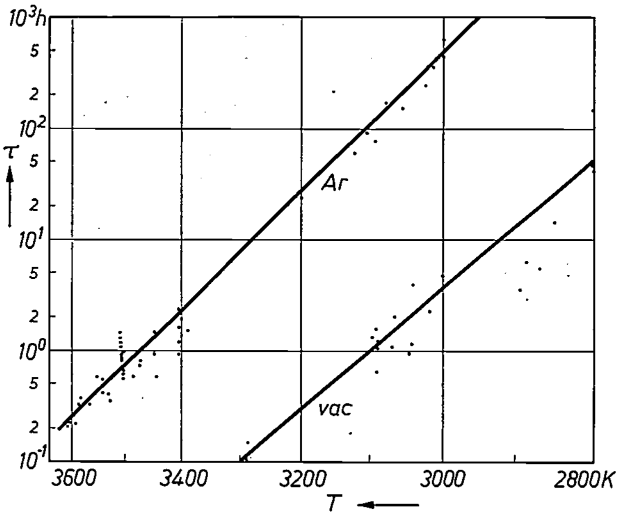


Fig. 8. Life τ of straight tungsten wires *in vacuo* (*vac*) and in inert gas (*Ar*, 600 torr) as a function of the burning temperature T. For the gas-filled lamp τ is about 100 times larger, this being the ratio between the mass losses; the "deadly weight loss" has remained constant.

$$M = 2\pi l\, D\, \frac{n(T_1)}{\ln r_{\text{L}}/r_1}, \quad (9)$$

where l is the length of the wire and r_{L} the radius of the Langmuir film. The ratio g of the mass loss of tungsten in an inert gas to the vacuum case is given by:

$$g = \frac{M}{m} = \frac{cD}{T_1^{\frac{1}{2}} r_1 \ln r_{\text{L}}/r_1}. \quad (10)$$

This ratio g depends only slightly on the wire temperature and for the example of $r_1 = 100$ μm, 1 atm argon and a wire temperature of 3000 K, g is about 1/100.

Introducing this result into the spot model we obtain an expansion of the time scale by the factor $1/g$. For the temperatures $T_1(t)$ and $T_2(t)$, the life τ increases by the same factor, thus we obtain:

$$\tau_{\text{gas}} = \tau_{\text{vac}}/g. \quad (11)$$

Referring to equation (4) and introducing the modified quantities m and τ it is evident that the deadly weight loss remains constant.

This conclusion is in agreement with experimental results. An increase of the life τ by the corresponding factor $1/g$ is found. This is shown in *fig. 8*. Similar results have been obtained by Fonda [7]. For the deadly weight loss of straight wires an average value of 15% was obtained, which comes close to the 18% obtained for the vacuum case.

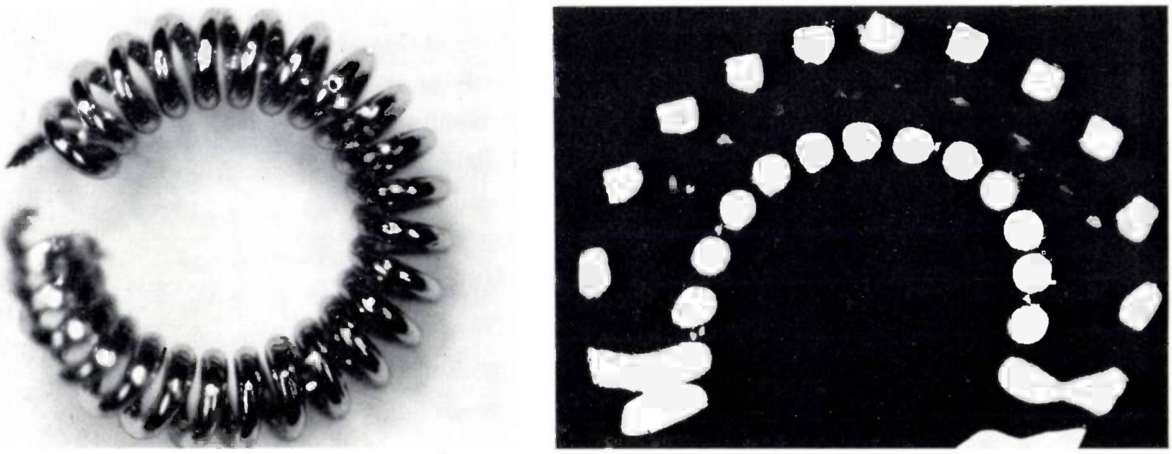


Fig. 9. *a*) Typical coiled-coil filament. *b*) Micrograph of cross-section of a coiled-coil filament after burn-out. The diagonals of the outer parts of the turns which have extended beyond the original diameter of the wire verify mass transport from the inner parts of the turns to the outer parts.

Life and deadly weight loss of a coil

A comparison of measurements on coils in gas and *in vacuo* yields some interesting results. As is to be expected the life increases in an inert-gas atmosphere. The deadly weight loss however changes from 15 to about 2%. From previous considerations on the straight-wire model it is known that the deadly weight loss and the life are controlled by the same physical process, radial tungsten diffusion. If the spot model were still valid one would expect no change in deadly weight loss. The observed discrepancy requires a revision of the spot model.

Well-known experimental facts such as the dependence of the life on coil temperature, pressure and type of gas indicate however that tungsten diffusion is still the life-determining process. It is therefore necessary to assume the existence of another diffusion process which overrides the radial diffusion.

In the search for an alternative diffusion mechanism, one likely possibility which occurs is the tungsten vapour transport within the coil itself [3]. This process can be caused by temperature differences between various parts of the coil. One example of this is demonstrated in *fig. 9*. Fig. 9*a* shows a typical coiled-coil filament and fig. 9*b* its cross-section near to the end of its life. As can be seen there is a mass transport from the inner parts of the turns to the outer parts. A quantitative study on the diffusion process within such a complicated geometrical configuration is however difficult. The situation is simplified in the case of a low-voltage lamp. Here, the strong influence of the axial heat conduction gives rise to a marked temperature decrease in the axial direction and consequently to an axial tungsten concentration gradient [4]. This leads to a high axial mass transport $M_{\rm ax}$ which under certain conditions may be higher than the radial transport.

The tungsten transport within a single coil can be calculated by assuming a temperature difference ΔT between adjacent turns, see *fig. 10*. This temperature difference gives rise to a concentration difference given by $n(T) = n_0 T^\gamma$ and $\Delta n = \gamma\, n(T) \Delta T/T$. The axial diffusion between adjacent turns at temperatures T and $T - \Delta T$ is then:

$$M_{\rm ax} = -\, 2\pi r_2\, D \int \operatorname{grad} n \, {\rm d}f =$$

$$= \frac{2\pi r_2\, D\, \pi \Delta n}{\ln\left[\dfrac{S}{r_1} + \sqrt{\left(\dfrac{S}{r_1}\right)^2 - 1}\right]}, \tag{12}$$

where $2S$ is the separation between centres of adjacent turns and r_2 the radius of the coil.

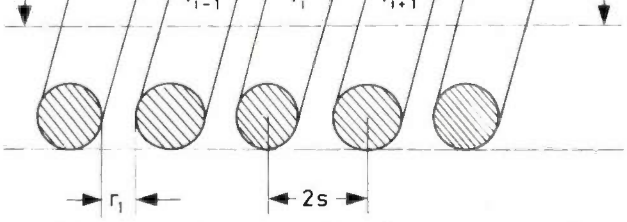


Fig. 10. Diagrammatic representation of tungsten transport in a single coil by radial ($M_{\rm rad}$) and axial diffusion ($M_{\rm ax}$). Indices $i-1, i, i+1, \ldots$ indicate the turn number.

[7] G. R. Fonda, Phys. Rev. **21**, 343, 1923 and **31**, 260, 1928.
[8] W. Elenbaas, Philips Res. Repts. **18**, 147, 1963.

Using the values given in fig. 10 the axial flow is:

$$M_{\mathrm{ax}} = \frac{2\pi r_2\ D\ \pi\Delta n}{\ln 2.6} \approx 2\pi^2 r_2\ D\ \gamma\ n(T)\ \frac{\Delta T}{T}. \quad (13)$$

This flow has to be compared with the radial flow of one turn. This can be calculated by using the approximate assumption of a constant tungsten concentration on a cylindrical envelope of the coil of length $3r_1$ and zero concentration at the Langmuir boundary:

$$M_{\mathrm{rad}} = D\ .\ 2\pi\ .\ 3r_1 \frac{n(T)}{\ln r_{\mathrm{L}}/r_2}. \quad (14)$$

The ratio of the axial to the radial mass flow is approximately:

$$\frac{M_{\mathrm{ax}}}{M_{\mathrm{rad}}} = \frac{\pi r_2\ \gamma \ln r_{\mathrm{L}}/r_2}{3r_1} \cdot \frac{\Delta T}{T}. \quad (15)$$

Of particular interest is the condition $M_{\mathrm{ax}}/M_{\mathrm{rad}} \geqq 1$. This defines the temperature difference at which the axial transport is equal to or greater than the radial transport:

$$\Delta T \geqq \frac{3Tr_1}{\pi\ \gamma\ r_2 \ln r_{\mathrm{L}}/r_2}. \quad (16)$$

For a temperature $T = 3000$ K, $r_1/r_2 = 0.1$ and $r_{\mathrm{L}}/r_2 = 5$, ΔT is about 5 K. The temperature difference is smaller than that between adjacent turns in low-voltage lamps. Hence we conclude that the life in low-voltage lamps is controlled by axial transport. A typical example in which the axial transport is predominant is the coil of a projection lamp (12 V, 100 W). The temperature distribution of such a coil is given in *fig. 11a*. From (13), and simplifying the notation, the axial transport varies along the length x of the coil as:

$$M_{\mathrm{ax}} = c\ n\ (T)\Delta T/T = c\ T^{34}(x)\ \Delta T(x), \quad (17)$$

since $\gamma = 35$ at the relevant temperature (3200 K).

From this formula it can be seen that the temperature distribution given in fig. 11*a* leads to a minimum in the axial transport at the centre and both ends. This results from the small ΔT at the centre and from the considerable decrease in temperature over the last turns of the coil.

Between these two minima a maximum must exist. Thus it is to be expected that if the axial transport is predominant the position of burn-out will be at the position of maximum M_{ax}, which in this case is not the position of highest temperature. The variation of temperature of a turn with time is, however, not a function of M_{ax} but of ΔM_{ax}, the net axial transport, which for the turn i is the difference between the mass flow from the turn $i - 1$ to the turn i and the flow from the turn i to the turn $i + 1$ (see fig. 10):

$$\Delta M_{\mathrm{ax},i} = C \left[T^{\gamma}_{i-1}\ \Delta T_{i-1,i} - T_i^{\gamma}\ \Delta T_{i,i+1} \right]. \quad (18)$$

where $\Delta T_{i-1,i} = T_{i-1} - T_i$ and $\Delta T_{i,i+1} = T_i - T_{i+1}$. For an initial temperature distribution as given in fig. 11*a* the calculated value of the net axial transport takes the form shown in fig. 11*b*. The negative values of ΔM_{ax} represent the mass loss of a turn. A photograph of the same coil after burn-out is shown in fig. 11*c*. A comparison of figs. 11*b* and *c* shows that the burn-out position is at the maximum of the mass loss. The characteristic features of fig. 11*c* are indicative of axial mass transport: the turn with the maximum mass loss has thinned and become smooth by thermal etching, whereas the neighbouring turn shows dendritic growth.

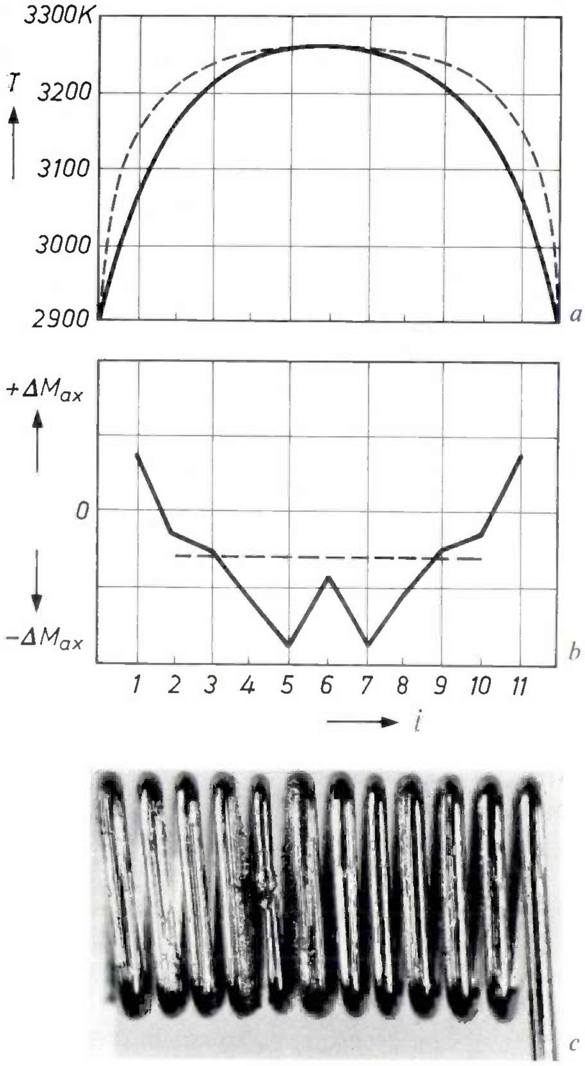


Fig. 11. *a*) Temperature distribution along the length of a coil of a projection lamp (12 V, 100 W). The turn numbers are again indicated by *i*. The solid line shows the measured results; the dashed line is in accordance with eq. (19) to achieve constant axial mass transport. *b*) Net axial mass transport calculated in accordance with eq. (18) from the measured temperature distribution (solid line) and the calculated temperature distribution (dashed line). *c*) Photograph of the coil after burn-out shows the spots at the maxima of axial mass loss.

Another example of this effect is shown in *fig. 12*, which depicts the coil of an automobile lamp after burn-out.

As can be seen the lamp does not burn through at the centre, the point of highest temperature, but at a position of lower temperature where the net axial flow has a maximum. As can be expected from a symmetrical temperature distribution two spots have developed at both sides of the centre. Again, this shows that the life is controlled by axial diffusion. Hence the life and the burn-out position can be predicted by analysing the temperature distribution which is controlled by the geometrical configuration of the coil.

The above considerations can also be applied to low-voltage halogen lamps. These differ from normal gas-filled lamps in that the tungsten is no longer transported to the wall but is transported from hotter to colder parts of the filament. In the high-temperature region of the filament, however, the tungsten transport is controlled by atomic diffusion as in the case of an inert-gas lamp. This is because of the low stability of iodine and bromine tungsten compounds. The life-determining process is therefore also axial tungsten diffusion in the halogen-lamp case. The observed longer life of a halogen lamp compared with a similar gas-filled lamp is due to the higher inert-gas pressure (for normal gas-filled lamps $p \approx 1$ atm, for halogen lamps $p \approx$ 5-10 atm), which reduces the tungsten diffusion coefficient $D \propto 1/p$ and thus the axial diffusion transport. Evidence of this is given in *fig. 13*, which shows the coil of a halogen lamp (type H3) shortly before the end of its life. Once again this demonstrates the characteristic features of turn-to-turn diffusion.

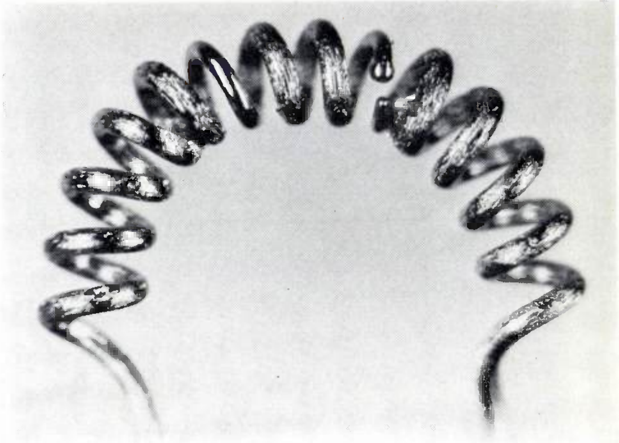


Fig. 12. Automobile lamp (6 V, 40 W) after burn-out showing the typical symmetrical position of the spots beside the centre of the coil.

Application of the turn-to-turn diffusion concept to low-voltage lamps

In the previous section it was established that axial diffusion governs the life of low-voltage lamps. It was also shown that this transport is entirely controlled by the axial temperature distribution. It is therefore reasonable to suppose that by altering the temperature distribution along the filament to reduce the maximum in the axial transport (see fig. 11*b*) a longer life should be obtained. A temperature distribution which fulfils this requirement can be derived from the condition that ΔM is constant along the length of the filament. This can be described mathematically by:

$$\partial M/\partial x = \text{constant},$$

where M is the total flow per turn, which in the general

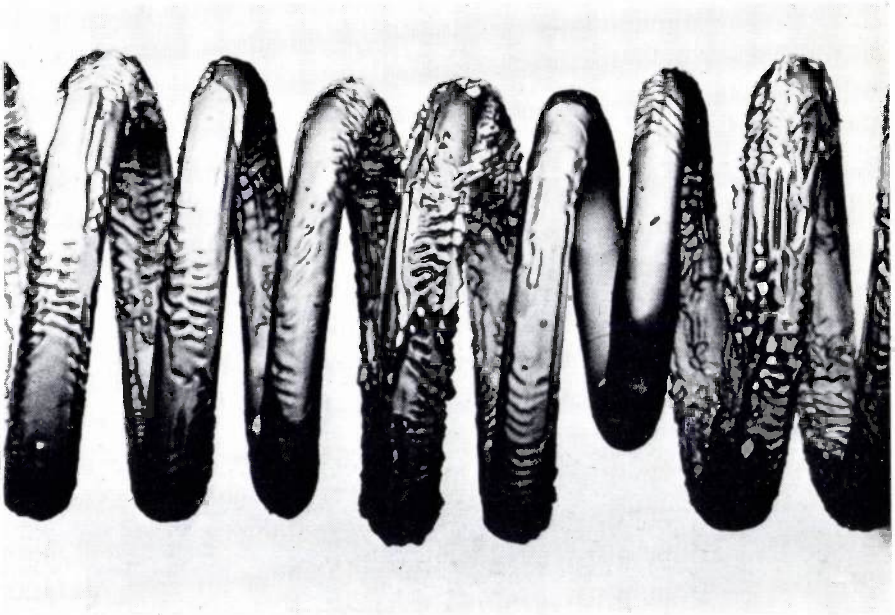


Fig. 13. Halogen automobile lamp (H3, 12 V, 55 W) at about the end of its life showing the characteristic features of turn-to-turn diffusion.

case includes both the axial and radial contributions to the mass transport. The solution of this differential equation yields the following temperature distribution:

$$T(x) = T_{\max}\left[1-\left\{1-\left(\frac{T_{\min}}{T_{\max}}\right)^{\nu}\right\}\times \times\frac{\exp(|x|/r_{\mathrm{L}})-(1+|x|/r_{\mathrm{L}})}{\exp(L/r_{\mathrm{L}})-(1+L/r_{\mathrm{L}})}\right]^{1/\nu}, \qquad (19)$$

where $T_{\max}$ is the maximum temperature at the centre of the filament ($x = 0$), $T_{\min}$ the minimum temperature at both ends ($x = \pm L$), r_{L} the radius of the Langmuir zone, and L half the length of the filament. In *fig. 11a* this temperature distribution is compared with one measured on a practical coil (12 V, 100 W) with the same value of $T_{\max}$ and $T_{\min}$. The calculated values of $\Delta M(x)$ corresponding to these temperature profiles are shown in fig. 11*b*. As can be seen the $\Delta M(x)$ for the ideal temperature distribution is constant and less than the maximum values of the practical coil. Its value can be further decreased by reducing the temperature difference between the centre and the ends of the coil. Experiments performed in this laboratory verify this and show that the life can be increased by a factor of about two. This topic will be dealt with in a separate article.

Until a few years ago the knowledge of incandescent lamps could be characterized by a number of well-known experimental facts. These did not provide a self-consistent physical picture. By introducing a simple model, the spot model, most of these facts can be related to a few physical quantities. One of the most important results of the model is that the life of an incandescent wire depends only on its starting conditions, i.e. the wire temperature and spot temperature. The model has been used to provide, for the first time, an analytical expression for the life. It is found that this depends not only on the evaporation rate but also on the magnitude of the initial defect. Furthermore it provides a physical background to the deadly weight loss, which until now has been used as a quality measure without recognizing its importance.

For the case of a straight wire the model shows that an initial local temperature difference of about 5 K determines the life. This can be due to surface roughness, which causes variations in the emissivity and in the wire diameter.

A critical inspection of the temperature distribution along the wire indicates the presence of more than one spot. Nevertheless the model is in almost perfect agreement with experiment. A more complicated model which takes into account a more realistic periodic temperature distribution only slightly modifies the results.

By applying this straight wire model to a coil in an inert-gas atmosphere agreement with experiments could not be obtained. This required the introduction of a further process based on tungsten transport within the coil. In the case of a low-voltage lamp this process is axial tungsten diffusion from turn to turn, which is caused by the axial temperature gradients. By calculating the axial mass transport from a given temperature profile of the coil the position of burn-out can be predicted. The model can also be used to derive a temperature profile which yields a minimum in axial mass transport and a corresponding increase in life.

It is expected that the results of this model will promote the development of tungsten-wire technology and improve the quality of incandescent lamps.

Scattering characteristics of a cross-junction of oversized waveguides

C. J. Bouwkamp

Introduction

Some years ago, the author analysed the reflection, transmission and deflection properties of a cross-junction of oversized waveguides in which the junction is excited by the dominant transverse-electric mode incident from infinity and propagating in one of the four waveguide arms forming the cross-junction structure. The study was performed at the request of Dr. H. J. Butterweck, then with our Laboratories and now Professor of Theoretical Electrical Engineering at Eindhoven University of Technology. Our results were given in an internal report which was never published. Presenting them in this paper seems warranted, however, for at least two reasons. First, it appeared that our "exact" solution of the scattering problem may serve as a check example in J. B. Keller's formalism of geometric-optics diffraction when applied to waveguide structures, as shown by the work of L. B. Felsen and others at the Brooklyn Polytechnic Institute. Secondly, our analysis has a bearing on problems outside the realm of electromagnetism, as is evident from work carried out recently under the supervision of Prof. J. Boersma of the Department of Mathematics of Eindhoven University of Technology. It is perhaps appropriate to mention here that Boersma was able to eliminate some of the difficulties previously encountered and to answer questions that had been left open.

With reference to *fig. 1*, the primary wave is assumed to be incident from infinity in region *I*. As a consequence, an infinite system of secondary waves is set up in each of the four arms of the structure. Only a finite number of these secondary waves propagate, the remainder being evanescent. In particular, if the waveguide arms are oversized, propagating higher-order modes occur as well as the dominant mode. In region *I* there will be reflected waves, in region *II* transmitted waves, and in regions *III* and *IV* deflected waves.

The aim of this paper is to analyse mathematically the scattering problem indicated above, and to derive relations between the various reflection, transmission, and deflection coefficients that will prove useful for numerical evaluation of these quantities by electronic computer and provide numerical data in support of experimental evidence obtained by Dr. Butterweck.

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

Mode expansion

The time variation of the field is assumed to be given by $\exp(-i\omega t)$, where i is the imaginary unit; this time factor is suppressed throughout. Further, k denotes the wave number, $k = \omega/c$, where c is the velocity of light.

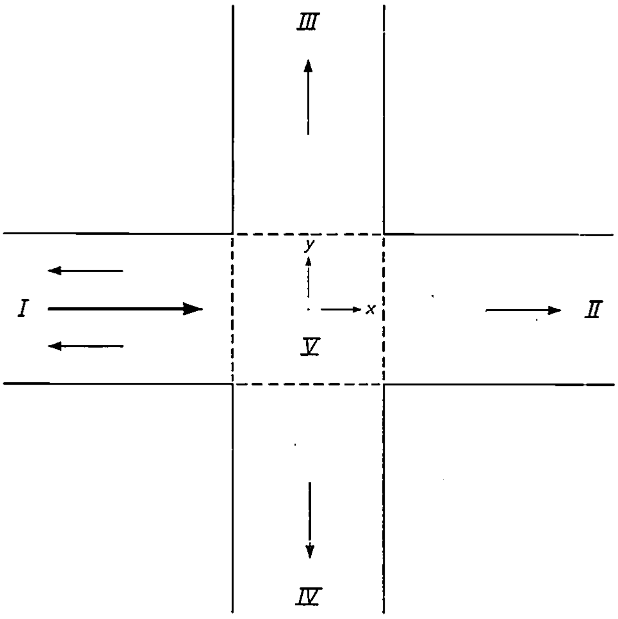


Fig. 1. Cross-junction of waveguides. The incident wave comes from the left in region *I*.

Let x,y denote rectangular cartesian coordinates in the plane of the drawing, with the origin at the centre of the cross as depicted in fig. 1, and with the z axis normal to this plane and pointing toward the reader. Since the incident wave is (dominant) transverse electric, the only nonvanishing component of the electric field is E_z, and the magnetic vector lies in the plane of drawing. Maxwell's equations pertinent to this situation are

$$E_x = E_y = H_z = 0,$$
$$ikH_x = \partial E_z/\partial y,$$
$$-ikH_y = \partial E_z/\partial x,$$
$$\partial H_x/\partial y - \partial H_y/\partial x = ikE_z,$$

where we have used Gaussian units.

It thus follows that our scattering problem is essentially a scalar one: the electromagnetic field can be expressed in terms of the scalar wave function $\Phi = E_z$ satisfying the time-independent two-dimensional wave equation $\Delta\Phi + k^2\Phi = 0$ subject to appropriate boundary and radiation conditions. Since we shall assume that the walls of the waveguide are perfectly conducting, we must have $\Phi = 0$ at the boundary.

Let us now consider the possible scalar-wave modes of an infinite waveguide of width a, and choose coordinates x',y' as indicated in *fig. 2*. With n a positive integer, the function

$$\sin(n\pi y'/a)\exp[i(\pi x'/a)\surd\{(ka/\pi)^2 - n^2\}]$$

is a solution of the wave equation that vanishes at the boundary ($y' = 0,a$) and represents a wave travelling in the positive x' direction. To simplify the notation, it is convenient to choose dimensionless quantities $\pi y'/a$, $\pi x'/a$, ka/π, and to denote them again by y', x', k, in that order. In other words, the width of the waveguide is normalized to $a = \pi$. The above solution then becomes

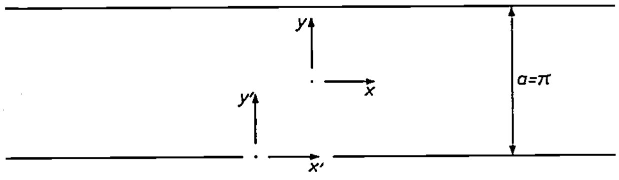


Fig. 2. The two coordinate systems.

$$\sin(ny')\exp[ix'\surd(k^2 - n^2)].$$

If $k > n$, this function represents a propagating wave; if $k < n$, it represents an evanescent wave. In this connection, we should be careful about the definition of the square root. Therefore, let $(p)^{\frac{1}{2}}$ denote the positive value of the square root of the positive number p. Then we define

$$\surd(k^2 - n^2) := \begin{cases} (k^2 - n^2)^{\frac{1}{2}} & \text{if} \quad k > n, \\ i(n^2 - k^2)^{\frac{1}{2}} & \text{if} \quad k < n. \end{cases}$$

In the second case we must have $+i(n^2 - k^2)^{\frac{1}{2}}$ in order to have a damped wave in the positive x' direction.

To avoid difficulties of resonance, we explicitly require k to be different from a positive integer, or else to assume k to be slightly complex, i.e. to have a small but positive imaginary part. Then, generally, $\surd(k^2 - n^2)$ is defined such that $0 \leqslant \arg[\surd(k^2 - n^2)] \leqslant \frac{1}{2}\pi$ with corresponding analytic continuation to the positive axis of real k.

To take full advantage of the symmetry of the cross-junction and its exciting wave, which is an even function of y, the coordinates x,y of fig. 1 are most appropriate. However, mode representation is more simple in the coordinates x',y' of fig. 2. If we displace the origin of x',y' over distances $\frac{1}{2}\pi$, we get the x,y system of fig. 2, the two systems being connected by $x' = x + \frac{1}{2}\pi$, $y' = y + \frac{1}{2}\pi$. In these new coordinates x,y the modal functions are

$$\sin\{n(y + \tfrac{1}{2}\pi)\}\exp[i(x + \tfrac{1}{2}\pi)\surd(k^2 - n^2)],$$

or, depending on the parity of n,

$$\sin(\tfrac{1}{2}n\pi)\cos(ny)\exp[i(x + \tfrac{1}{2}\pi)\surd(k^2 - n^2)]\ (n\ \text{odd}),$$

$$\cos(\tfrac{1}{2}n\pi)\sin(ny)\exp[i(x + \tfrac{1}{2}\pi)\surd(k^2 - n^2)]\ (n\ \text{even}).$$

If we wish, we may drop the phase factor $\exp[\frac{1}{2}\pi i\surd(k^2 - n^2)]$ or replace it by any other so long as it does not depend on x or y.

The dominant mode is obtained for $n = 1$:

$$\cos y\exp[i(x + \tfrac{1}{2}\pi)\surd(k^2 - 1)].$$

For this to be a propagating mode, we require $k > 1$.

As noted above, this dominant mode is an even function of y (see fig. 1; from now on the width of either arm is taken equal to π). This implies that the whole field configuration is an even function of y. Accordingly, in regions *I* and *II* only modes with n odd will be excited. On the other hand, the deflected modes in regions *III* and *IV*, obtained from the modes in regions *I* and *II* by interchanging x and y, will range over all natural numbers n, both odd and even.

With reference to fig. 1, we shall represent the total field in any of the four regions *I-IV* as indicated in *Table I* by means of a set of coefficients A_n, B_n, C_n and D_n. Inspection of these mode expansions shows that the primary wave is taken to be zero in regions *III* and *IV*, whereas both in region *I* and in region *II* it is taken to be $\cos y\exp[i(x + \frac{1}{2}\pi)\surd(k^2 - 1)]$. This function is simply $\cos y$ at the end of region *I*, i.e. at $x = -\frac{1}{2}\pi$. This means that the incident-wave amplitude is normalized to unity at the point where the wave enters the actual cross-junction, region *V*.

The secondary fields in the first four regions are expanded in infinite series of modes. In all terms the phase factors are chosen so that they become unity (phase zero) at the boundary of region *V*. Again, this is a question of normalization but it is highly practical.

What we are trying to find are the coefficients A_n, B_n, D_n (n odd) and C_n (n even). These are complex numbers. From a mathematical analysis of the scattering problem we must be able to determine how they depend on k.

The *reflection* coefficients for region *I* are

$$R_n = 0\ (n\ \text{even}),\quad R_n = B_n - D_n\ (n\ \text{odd}).$$

The *transmission* coefficients for region *II* are

$$T_n = 0\ (n\ \text{even}),\quad T_n = B_n + D_n\ (n\ \text{odd}, \neq 1)$$

$$T_1 = B_1 + D_1 + \exp[\pi i\surd(k^2 - 1)].$$

Table I. Mode-expansion coefficients in the four regions of fig. 1.

Region I $(-\infty < x \leqslant -\frac{1}{2}\pi,\ -\frac{1}{2}\pi \leqslant y \leqslant \frac{1}{2}\pi)$:

$$\phi_I = \cos y \exp[i(x+\tfrac{1}{2}\pi)\sqrt{(k^2-1)}] + \sum_{n\ \mathrm{odd}\ >\ 0} (B_n - D_n)\sin(\tfrac{1}{2}n\pi)\cos(ny)\exp[-i(x+\tfrac{1}{2}\pi)\sqrt{(k^2-n^2)}]$$

Region II $(\frac{1}{2}\pi \leqslant x < \infty,\ -\frac{1}{2}\pi \leqslant y \leqslant \frac{1}{2}\pi)$:

$$\phi_{II} = \cos y \exp[i(x+\tfrac{1}{2}\pi)\sqrt{(k^2-1)}] + \sum_{n\ \mathrm{odd}\ >\ 0} (B_n + D_n)\sin(\tfrac{1}{2}n\pi)\cos(ny)\exp[i(x-\tfrac{1}{2}\pi)\sqrt{(k^2-n^2)}]$$

Region III $(-\frac{1}{2}\pi \leqslant x \leqslant \frac{1}{2}\pi,\ \frac{1}{2}\pi \leqslant y < \infty)$:

$$\phi_{III} = \sum_{n\ \mathrm{odd}\ >\ 0} A_n \sin(\tfrac{1}{2}n\pi)\cos(nx)\exp[i(y-\tfrac{1}{2}\pi)\sqrt{(k^2-n^2)}] - \sum_{n\ \mathrm{even}\ >\ 0} C_n \cos(\tfrac{1}{2}n\pi)\sin(nx)\exp[i(y-\tfrac{1}{2}\pi)\sqrt{(k^2-n^2)}]$$

Region IV $(-\frac{1}{2}\pi \leqslant x \leqslant \frac{1}{2}\pi,\ -\infty < y \leqslant -\frac{1}{2}\pi)$:

$$\phi_{IV} = \sum_{n\ \mathrm{odd}\ >\ 0} A_n \sin(\tfrac{1}{2}n\pi)\cos(nx)\exp[-i(y+\tfrac{1}{2}\pi)\sqrt{(k^2-n^2]} - \sum_{n\ \mathrm{even}\ >\ 0} C_n \cos(\tfrac{1}{2}n\pi)\sin(nx)\exp[-i(y+\tfrac{1}{2}\pi)\sqrt{(k^2-n^2)}]$$

The *deflection* coefficients for both regions III and IV are

$$S_n = A_n\ (n\ \text{odd})\quad\text{and}\quad S_n = -C_n\ (n\ \text{even}).$$

Our aim is to determine and tabulate these coefficients for $k = 1\ (0.1)\ 13$, $k \neq$ integer, in amplitude and phase, up to $n = 30$.

Integral-equation formulation

The scattering problem under discussion can be separated into two independent parts. One part is an even function of x, the other part is odd in x. To see this, we resolve the primary exciting wave into two standing waves, u_1 and u_2, where

$$u_1 = \exp[\tfrac{1}{2}\pi i\sqrt{(k^2-1)}]\cos y \cos\{x\sqrt{(k^2-1)}\},$$

which is an even function of x, and

$$u_2 = i\exp[\tfrac{1}{2}\pi i\sqrt{(k^2-1)}]\cos y \sin\{x\sqrt{(k^2-1)}\},$$

which is an odd function of x.

If the cross-junction is excited by the function u_1 alone, the total field will be even in x; if the junction is excited by u_2, the field will be odd in x. By addition of the two partial fields we get the field due to the incident travelling wave $u_1 + u_2$.

We shall first deal with the even problem. As a preliminary, we solve the following Dirichlet boundary-value problem for the wave equation (see *fig. 3* for notation). Let $f_1(x) = f_1(-x)$ be an even function of x defined in the interval $-\frac{1}{2}\pi < x < \frac{1}{2}\pi$. Let $\Psi(x,y)$ be a solution of the wave equation inside the domain bounded by the two horizontal lines, with $\Psi = 0$ at the boundary except for $y = \pm\frac{1}{2}\pi$, $-\frac{1}{2}\pi < x < \frac{1}{2}\pi$, where $\Psi = f_1(x)$. Further, let $\Psi(x,y)$ be even in x, even in y, and let it finally satisfy the radiation condition at $x = \pm\infty$. This function $\Psi(x,y)$ exists and is unique if certain conditions of smoothness, not to be specified here, are satisfied. It is given by

$$\Psi(x,y) := \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} f_1(t)\,\mathrm{d}t\ \times$$

$$\times\ \frac{1}{\pi}\int_0^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(x-t)\}\,\mathrm{d}s. \qquad (1)$$

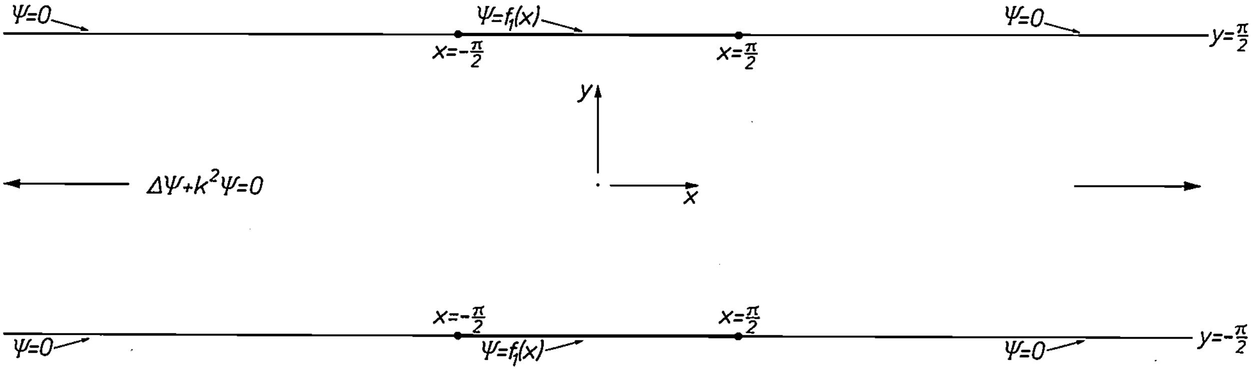


Fig. 3. Illustrating the even boundary-value problem.

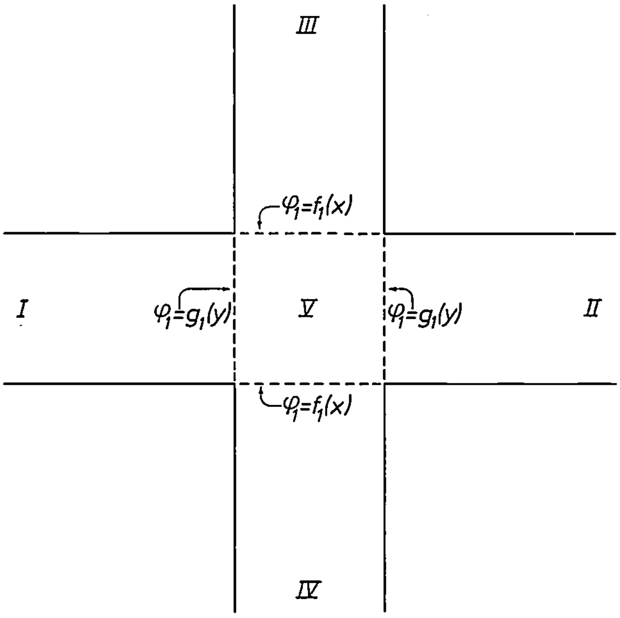


Fig. 4. Boundary values for the even problem.

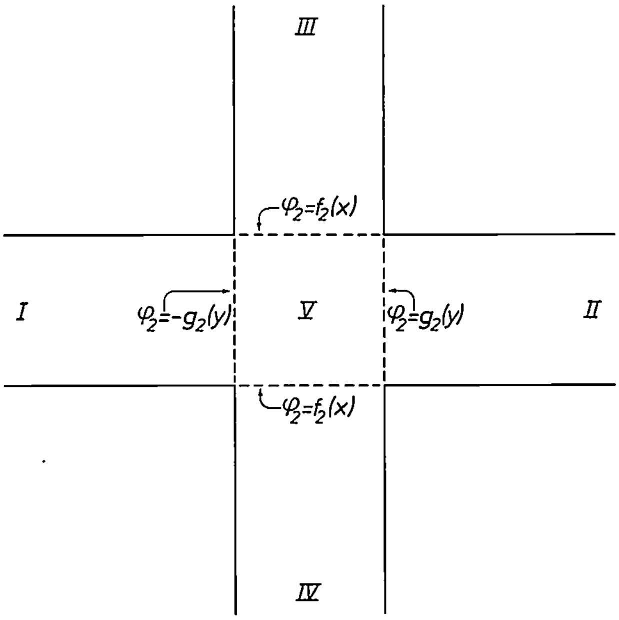


Fig. 5. Boundary values for the odd problem.

To prove this, we observe that the right-hand side (RHS) of (1) is an even function of y. It is also an even function of x, because we may replace $\cos\{s(x-t)\} = \cos(sx)\cos(st)+\sin(sx)\sin(st)$ in (1) by $\cos(sx)\cos(st)$ without altering the integral (note that f_1 is even). Thus $\Psi(x,y)$ has the proper symmetry. Further, RHS reduces to $f_1(x)$ for $y = \pm\frac{1}{2}\pi$ if $-\frac{1}{2}\pi < x < \frac{1}{2}\pi$ and to zero outside this interval, from Fourier's theorem. Hence, $\Psi(x,y)$ satisfies the proper boundary conditions. Next, inside the strip of fig. 3, RHS is indeed a solution of the wave equation, because $\cosh\{y\sqrt{(s^2-k^2)}\}\cos\{s(x-t)\}$ is so, for any value of s and t. This property remains valid upon integration with respect to s and t after multiplication by weighting factors independent of x and y. Some detailed analysis is required to see that RHS satisfies the radiation condition at infinity. For the latter purpose, let us consider, for t real, the function (even in t) defined as follows:

$$\chi(t) := \frac{1}{\pi}\int_0^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos(st)\,\mathrm{d}s.$$

Irrespective of the ambiguous square root, the integrand is analytic and an even function of s. Therefore,

$$\chi(t) = \frac{1}{2\pi}\int_{-\infty}^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\exp[is|t|]\,\mathrm{d}s.$$

The integrand has an infinity of simple poles at $s = \sqrt{(k^2-n^2)}$ (n odd, > 0) in the upper plane of complex s, with residues

$$\frac{2}{\pi}\,\frac{n}{\sqrt{(k^2-n^2)}}\sin(\tfrac{1}{2}n\pi)\cos(ny)\exp[i|t|\sqrt{(k^2-n^2)}].$$

Hence

$$\frac{1}{\pi}\int_0^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(x-t)\}\,\mathrm{d}s =$$

$$= \frac{2i}{\pi}\sum_{n\text{ odd }>\,0}\frac{n}{\sqrt{(k^2-n^2)}}\sin(\tfrac{1}{2}n\pi)\cos(ny)\exp[i|x-t|\sqrt{(k^2-n^2)}].$$

The left-hand side is a solution of the wave equation; the right-hand side is the corresponding mode expansion in which each term satisfies the radiation condition at infinity: the first few represent propagating modes radiating toward infinity, and the remaining modes represent exponentially damped waves as $|x| \to \infty$. On multiplication by $f_1(t)$ and integration over $-\frac{1}{2}\pi \leqslant t \leqslant \frac{1}{2}\pi$, these properties remain valid for $|x| > \frac{1}{2}\pi$. Thus RHS of eq. (1) does indeed satisfy the required radiation condition.

This concludes our proof that the function $\Psi(x,y)$ defined by eq. (1) is the (unique) solution of our Dirichlet boundary-value problem.

Now, let $\phi_1(x,y)$ denote the total field of the even problem, and let $f_1(x)$ and $g_1(y)$ denote its values at the boundary of region V (see *fig. 4*); f_1 and g_1 are both even functions of their respective arguments. Note that the primary wave does not contribute to f_1 because it is identically zero if $y = \pm\frac{1}{2}\pi$.

In the union of regions I, V and II, we have

$$\phi_1(x,y) = \exp\,[\tfrac{1}{2}\pi i \sqrt{(k^2-1)}]\cos y \cos\{x\sqrt{(k^2-1)}\} + \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} f_1(t)\,\mathrm{d}t\,\frac{1}{\pi}\int_0^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(x-t)\}\,\mathrm{d}s,$$

on account of the representation theorem (1).

On the other hand, in the union of regions *III*, *V* and *IV*, we have

$$\phi_1(x,y) = \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} g_1(t)\,\mathrm{d}t\;\frac{1}{\pi}\int_0^\infty \frac{\cosh\{x\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(y-t)\}\,\mathrm{d}s,$$

by applying eq. (1) to the "vertical" arms; note that x and y are interchanged.

The regions of validity of these two equations overlap in region *V*, including its boundary. Applying the first equation to $x = \frac{1}{2}\pi$ gives

$$g_1(y) = \tfrac{1}{2}(\exp\,[\pi i\sqrt{(k^2-1)}]+1)\cos y + \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} f_1(t)\,\mathrm{d}t\,\frac{1}{\pi}\int_0^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(\tfrac{1}{2}\pi-t)\}\,\mathrm{d}s\,,$$

and application of the second equation to $y = \frac{1}{2}\pi$ gives

$$f_1(x) = \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} g_1(t)\,\mathrm{d}t\,\frac{1}{\pi}\int_0^\infty \frac{\cosh\{x\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(\tfrac{1}{2}\pi-t)\}\,\mathrm{d}s\,.$$

These last two equations form a system of coupled linear integral equations for the two unknown functions f_1 and g_1. They can be simplified a little by using the parity properties of f_1 and g_1: the factor $\cos\{s(\frac{1}{2}\pi-t)\}$ may be replaced by $\cos(\frac{1}{2}\pi s)\cos(st)$. To simplify the notation, let the kernel function for the even problem be defined by

$$K(x,t) := \frac{1}{\pi}\int_0^\infty \frac{\cosh\{x\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos(\tfrac{1}{2}\pi s)\cos(st)\,\mathrm{d}s\,. \tag{2}$$

We then get the integral equations of the even problem in the following form:

$$\left.\begin{aligned} f_1(x) &= \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} K(x,t)\,g_1(t)\,\mathrm{d}t \qquad (-\tfrac{1}{2}\pi < x < \tfrac{1}{2}\pi),\\ g_1(y) &= \tfrac{1}{2}(\exp[\pi i\sqrt{(k^2-1)}]+1)\cos y + \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} K(y,t)\,f_1(t)\,\mathrm{d}t\; (-\tfrac{1}{2}\pi < y < \tfrac{1}{2}\pi).\end{aligned}\right\} \tag{3}$$

We now turn to the odd problem. The total field will be denoted by $\phi_2(x,y)$, which is even in y but odd in x. Its values at the boundary of region *V* are denoted by $f_2(x)$ and $\pm\, g_2(y)$, where g_2 is even in y and f_2 is odd in x (see *fig. 5*).

In the union of regions *I*, *V* and *II*, we have

$$\phi_2(x,y) = i\exp\,[\tfrac{1}{2}\pi i\sqrt{(k^2-1)}]\cos y\sin\{x\sqrt{(k^2-1)}\} + \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} f_2(t)\,\mathrm{d}t\,\frac{1}{\pi}\int_0^\infty \frac{\cosh\{y\sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(x-t)\}\,\mathrm{d}s\,,$$

again by application of the representation theorem (1).

On the other hand, in the union of regions *III*, *V* and *IV*, we have

$$\phi_2(x,y) = \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} g_2(t)\,\mathrm{d}t\,\frac{1}{\pi}\int_0^\infty \frac{\sinh\{x\sqrt{(s^2-k^2)}\}}{\sinh\{\frac{1}{2}\pi\sqrt{(s^2-k^2)}\}}\cos\{s(y-t)\}\,\mathrm{d}s,$$

which is obtained on applying a representation theorem quite analogous to that of eq. (1).

These two representations overlap in region V, and upon applying them to the boundary of region V we again get a system of two simultaneous integral equations. The difference from the even problem is that we now need two kernel functions.

These two kernels are defined as follows:

$$L(x,t) := \frac{1}{\pi} \int_0^\infty \frac{\cosh\{x \sqrt{(s^2-k^2)}\}}{\cosh\{\frac{1}{2}\pi \sqrt{(s^2-k^2)}\}} \sin(\tfrac{1}{2}\pi s) \sin(st) \, \mathrm{d}s, \tag{4}$$

$$M(x,t) := \frac{1}{\pi} \int_0^\infty \frac{\sinh\{x \sqrt{(s^2-k^2)}\}}{\sinh\{\frac{1}{2}\pi \sqrt{(s^2-k^2)}\}} \cos(\tfrac{1}{2}\pi s) \cos(st) \, \mathrm{d}s. \tag{5}$$

The resulting system of integral equations for the odd problem is then

$$\left.\begin{aligned} f_2(x) &= \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} M(x,t)\, g_2(t) \, \mathrm{d}t \quad (-\tfrac{1}{2}\pi < x < \tfrac{1}{2}\pi), \\ g_2(y) &= \tfrac{1}{2} (\exp[\pi i \sqrt{(k^2-1)}] - 1) \cos y + \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} L(y,t)\, f_2(t) \, \mathrm{d}t \quad (-\tfrac{1}{2}\pi < y < \tfrac{1}{2}\pi). \end{aligned}\right\} \tag{6}$$

Of course, we have not attempted to solve the systems (3) and (6) of coupled linear equations in closed form, were such a thing at all possible. In addition, the engineer would not be interested in numerical tables for the unknown functions f and g, so that a direct numerical solution of (3) and (6) by computer is of little interest. What the engineer is interested in is a table of values of the reflection, transmission, and deflection coefficients defined in the previous section, because they have a clear physical meaning and are not difficult to measure experimentally. Fortunately, these coefficients are simple combinations of those of the Fourier-series expansions of the functions f and g. In this way we can transform both (3) and (6) into an infinite system of coupled inhomogeneous linear algebraic equations with an infinite number of unknowns A_n, B_n and C_n, D_n, respectively. This transformation is discussed in the next section.

Fourier-series expansion

First of all, the three kernel functions $K(x,t)$, $L(x,t)$ and $M(x,t)$ can be expanded in Fourier series with respect to x for $|x| \leqslant \frac{1}{2}\pi$ and $|t| \leqslant \frac{1}{2}\pi$. This can be accomplished by contour integration and the calculus of residues. We shall not go into details but will only mention the final results:

$$K(x,t) = \frac{2i}{\pi} \sum_{n \text{ odd} > 0} \frac{n}{\sqrt{(k^2-n^2)}} \exp[\tfrac{1}{2}\pi i \sqrt{(k^2-n^2)}] \sin(\tfrac{1}{2}n\pi) \cos(nx) \cos\{t \sqrt{(k^2-n^2)}\}, \tag{7}$$

$$L(x,t) = \frac{2}{\pi} \sum_{n \text{ odd} > 0} \frac{n}{\sqrt{(k^2-n^2)}} \exp[\tfrac{1}{2}\pi i \sqrt{(k^2-n^2)}] \sin(\tfrac{1}{2}n\pi) \cos(nx) \sin\{t \sqrt{(k^2-n^2)}\}, \tag{8}$$

$$M(x,t) = \frac{2}{\pi i} \sum_{n \text{ even} > 0} \frac{n}{\sqrt{(k^2-n^2)}} \exp[\tfrac{1}{2}\pi i \sqrt{(k^2-n^2)}] \cos(\tfrac{1}{2}n\pi) \sin(nx) \cos\{t \sqrt{(k^2-n^2)}\}. \tag{9}$$

Secondly, we introduce the unknown Fourier coefficients of the functions f_1, g_1, f_2, g_2 as follows (cf. the coefficients in Table I):

$$f_1(x) = \sum_{n \text{ odd} > 0} A_n \sin(\tfrac{1}{2}n\pi)\cos(nx)\,, \tag{10}$$

$$g_1(y) = \tfrac{1}{2}(\exp[\pi i \sqrt{(k^2-1)}] + 1)\cos y + \sum_{n \text{ odd} > 0} B_n \sin(\tfrac{1}{2}n\pi)\cos(ny), \tag{11}$$

$$f_2(x) = -\sum_{n \text{ even} > 0} C_n \cos(\tfrac{1}{2}n\pi)\sin(nx), \tag{12}$$

$$g_2(y) = \tfrac{1}{2}(\exp[\pi i \sqrt{(k^2-1)}] - 1)\cos y + \sum_{n \text{ odd} > 0} D_n \sin(\tfrac{1}{2}n\pi)\cos(ny). \tag{13}$$

These equations are valid for $|x|$ or $|y|$ not greater than $\frac{1}{2}\pi$.

By way of example, we shall indicate how (3) can be converted into algebraic equations in terms of the coefficients A_n and B_n of the Fourier series of f_1 and g_1.

By invoking (7) and (10), we have

$$\int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} K(y,t)\, f_1(t)\,\mathrm{d}t =$$

$$= \sum_{m \text{ odd} > 0} A_m \sin(\tfrac{1}{2}m\pi) \times \frac{2i}{\pi} \sum_{n \text{ odd} > 0} \frac{n}{\sqrt{(k^2-n^2)}} \exp[\tfrac{1}{2}\pi i \sqrt{(k^2-1)}] \sin(\tfrac{1}{2}n\pi)\cos(ny) \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \cos(mt)\cos\{t\sqrt{(k^2-n^2)}\,\mathrm{d}t.$$

The remaining integral is elementary:

$$\int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \cos(mt)\cos\{t\sqrt{(k^2-n^2)}\}\,\mathrm{d}t = \frac{2m\sin(\frac{1}{2}m\pi)\cos\{\frac{1}{2}\pi\sqrt{(k^2-n^2)}\}}{m^2+n^2-k^2},$$

where m is odd and positive.

As a consequence, the coefficient of $\sin(\frac{1}{2}n\pi)\cos(ny)$, for fixed n, becomes

$$\frac{4i}{\pi}\frac{n}{\sqrt{(k^2-n^2)}}\exp[\tfrac{1}{2}\pi i\sqrt{(k^2-n^2)}]\cos\{\tfrac{1}{2}\pi\sqrt{(k^2-n^2)}\} \sum_{m \text{ odd} > 0} \frac{m}{m^2+n^2-k^2} A_m .$$

From (11) and the second equation of (3), this coefficient must be B_n. Thus we get equation (14) of *Table II*.

Similarly, by substituting (11) in the first of eq. (3) and using (7) we can express A_n as a linear combination of the B coefficients apart from the extra term due to the primary field. The result is found to be eq. (15) of Table II.

Table II. The two systems of linear equations.

$$B_n = \frac{2i}{\pi}\frac{n}{\sqrt{(k^2-n^2)}}(1+\exp[\pi i\sqrt{(k^2-n^2)}]) \sum_{m \text{ odd} > 0} \frac{m}{m^2+n^2-k^2} A_m \qquad (n \text{ odd} > 0) \qquad (14)$$

$$A_n = \frac{2i}{\pi}\frac{n}{\sqrt{(k^2-n^2)}}(1+\exp[\pi i\sqrt{(k^2-n^2)}]) \left[\frac{\exp[\pi i\sqrt{(k^2-1)}]+1}{2(n^2+1-k^2)} + \sum_{m \text{ odd} > 0} \frac{m}{m^2+n^2-k^2} B_m\right] \qquad (n \text{ odd} > 0) \qquad (15)$$

$$D_n = \frac{2i}{\pi}\frac{n}{\sqrt{(k^2-n^2)}}(1-\exp[\pi i\sqrt{(k^2-n^2)}]) \sum_{m \text{ even} > 0} \frac{m}{m^2+n^2-k^2} C_m \qquad (n \text{ odd} > 0) \qquad (16)$$

$$C_n = \frac{2i}{\pi}\frac{n}{\sqrt{(k^2-n^2)}}(1+\exp[\pi i\sqrt{(k^2-n^2)}]) \left[\frac{\exp[\pi i\sqrt{(k^2-1)}]-1}{2(n^2+1-k^2)} + \sum_{m \text{ odd} > 0} \frac{m}{m^2+n^2-k^2} D_m\right] \qquad (n \text{ even} > 0) \qquad (17)$$

The conversion of the system (6) of integral equations is quite similar. The result is found in equations (16) and (17) of Table II.

It can be seen that the two systems (14, 15) and (16, 17) look very similar. In fact, they are identical for numerical analysis and computer programming. To see this, we introduce the parity index j, with $j = 0$ for the even problem, and $j = 1$ for the odd problem, and we set, for $n = 1, 2, 3, \ldots,$

$$A_{2n-1} = X(1,n), \qquad B_{2n-1} = Y(1,n),$$
$$C_{2n} = X(2,n), \qquad D_{2n-1} = Y(2,n).$$

Then, as is easy to verify, the two systems merge into one, as follows:

$$Y(j+1,n) = -\frac{2(2n-1)}{\pi i\sqrt{\{k^2-(2n-1)^2\}}}\left[1+(-1)^j\exp[\pi i\sqrt{\{k^2-(2n-1)^2\}}]\right]\times$$
$$\times\sum_{m=1}^{\infty}\frac{2m-1+j}{(2m-1+j)^2+(2n-1)^2-k^2}X(j+1,m), \tag{18}$$

$$X(j+1,n) = -\frac{2(2n-1+j)}{\pi i\sqrt{\{k^2-(2n-1+j)^2\}}}\left[1+\exp[\pi i\sqrt{\{k^2-(2n-1+j)^2\}}]\right]\times$$
$$\times\left[\frac{\exp[\pi i\sqrt{(k^2-1)}]+(-1)^j}{2\{(2n-1+j)^2+1-k^2\}}+\sum_{m=1}^{\infty}\frac{2m-1}{(2m-1)^2+(2n-1+j)^2-k^2}Y(j+1,m)\right]. \tag{19}$$

This infinite system of simultaneous linear equations for the infinite number of unknowns X and Y can easily be transformed into one single system for the X alone by elimination of Y. For numerical computations the latter system is then truncated to $n = 30$, say, and the corresponding finite system of order 30 is solved by one of the well-known techniques. Knowing the first 30 components of X, we then compute the first 30 components of Y by means of (18). Of course, the values obtained are approximations to the true values.

Indeed, the approximation will be very poor for the last few components because they are strongly affected by the truncation. Therefore we shall retain only the first 15 components of X and Y in the final output.

Details of numerical analysis and programming

Our aim is to produce a table of the coefficients R_n, T_n and S_n, in modulus and argument, for $n = 1(1)30$ and for $k = 1(0.1)13$ ($k \neq$ integer). This table consists of 108 pages if the data for fixed k is printed out on one page. We shall also compute and print the total energy reflected, transmitted, or deflected in the arms of the cross-junction.

In the FORTRAN CDC-3600 program (see Appendix), we use the identifier NCP to control the number of coefficients to be printed (it is 15, at most), NEQ to denote the order of the finite system of linear equations after truncation of the infinite system. Further, PI denotes π to 10 digits, CI represents the imaginary unit i, KU is the number of units contained in k, KT its number of tenths, KK represents k^2, and k itself is represented by K. The necessary values of k are covered by a double DO loop (DO 1).

By inspection of eqs. (18) and (19) it is seen that the values of $\pi i\sqrt{(k^2-n^2)}$ and $\exp[\pi i\sqrt{(k^2-n^2)}]$ are needed for $n = 1(1)60$. It is thus wise to compute and store them once and for all. We represent the corresponding arrays by the identifiers RT (short for root) and ER (exponential of root), respectively (DO 2). Note that AR represents k^2-n^2. The IF statement involves a test on the value of AR. If $k^2-n^2 > 0$, there is a jump to the statement labelled 21. That is, if $k > n$, we have RT(n) $= \pi i\,(k^2-n^2)^{\frac{1}{2}}$. On the other hand, if $k < n$ (note that $k = n$ is impossible because k is nonintegral), the statement RT(N) = −PI★SQRTF (−AR) is executed, and the program jumps to statement labelled 22, thus skipping that labelled 21. That is, if $k < n$ we will have RT(n) $= -\pi\,(k^2-n^2)^{\frac{1}{2}}$. This evaluation of $\pi i\sqrt{(k^2-n^2)}$ is exactly in accordance with the definition of the square root as given before. In both cases the result is stored as a complex number. The statement labelled 22 also serves to compute and store the corresponding correct value of the exponential expression by use of the standard subroutine CEXP.

Next we should recall that there are two problems to be solved: the even problem and the odd problem. The computer does this under control of the parity index J representing j. This amounts to a most simple loop (DO 3).

At this stage of the programming we again look at eqs. (18) en (19). They can be written as follows:

$$Y(j+1,n) = f_1(n) \sum_{m=1}^{\infty} \frac{2m-1+j}{(2m-1+j)^2+(2n-1)^2-k^2} X(j+1,m),$$

$$X(j+1,n) = -f(n) + f_2(n) \sum_{m=1}^{\infty} \frac{2m-1}{(2n-1+j)^2+(2m-1)^2-k^2} Y(j+1,m),$$

where

$$f_1(n) = -\frac{2(2n-1)}{\mathrm{RT}(2n-1)} \{1+(-1)^j\,\mathrm{ER}(2n-1)\},$$

$$f_2(n) = -\frac{2(2n-1+j)}{\mathrm{RT}(2n-1+j)} \{1+\mathrm{ER}(2n-1+j)\},$$

$$f(n) = -f_2(n) \frac{\mathrm{ER}(1)+(-1)^j}{2\{(2n-1+j)^2+1-k^2\}}.$$

We need the corresponding values for $n = 1(1)30$; we therefore compute and store them (DO 4). The values for $j = 0$ computed and stored first, then used first, are later on over-written by the values for $j = 1$.

We now come to the elimination of Y and to evaluating the matrix elements and the right-hand sides relating to the resulting equations in X. In fact, the right-hand side appears to be $f(n)$. If the equations are written in the form

$$\sum_{m=1}^{\infty} \mathrm{MAT}(n,m)\, X(m) = f(n),$$

it is not difficult to show that

$$\mathrm{MAT}(n,m) = -\delta_{n,m} + (2m-1+j) f_2(n) \sum_{s=1}^{\infty} \frac{(2s-1) f_1(s)}{\{(2s-1)^2+(2n-1+j)^2-k^2\}\{(2s-1)^2+(2m-1+j)^2-k^2\}}.$$

The infinite sum is of course truncated at $s =$ NEQ. Also, note that MAT is not symmetric. The corresponding piece of program involves a triple DO loop. First there is one (DO 5) for $n = 1(1)30$, then one (DO 6) for $m = 1(1)30$, and lastly one (DO 7) for the evaluation of the sum over $s = 1(1)30$. Here 30 is the maximum admissible value of NEQ below. The relevant loop variables are N, M and I.

At this moment we are ready to actually solve the linear equations because the matrix and the right-hand side are available in store. This solving is done by an independent subprogram, which at this stage is called by the main program.

After return from the subprogram CLINEQ (short for complex linear equations), the solution is available in the array XX. The main programme must then store this solution in the array X(NC,N), which is accomplished by a simple loop (DO 8).

We then have to invoke eq. (18) to compute and store the corresponding elements of array Y. This is done by a double loop (DO 9 and DO 10).

We are now ready with the program to close the NC loop. If the computer has executed this loop (DO 3), we have X(1,N), Y(1,N), X(2,N) and Y(2,N) numerically available for further use; here N $= 1(1)30$. We assume that the first 15 components are really significant and drop the other 15. We then compute and store the reflection coefficients R_{2n-1}, the transmission coefficients T_{2n-1}, and the deflection coefficients $S_{2n-1} = A_{2n-1}$ and $S_{2n} = -C_{2n}$, for $n = 1(1)15$. The corresponding arrays are identified by REF, TRM, and DEF, respectively (DO 11). The statement following upon this loop is needed to include the primary wave in region *II* into the transmitted dominant mode.

We are interested in the absolute values and the phases of the various coefficients (DO 12 and DO 13).

Table III. Reflected, transmitted and scattered energies as a percentage of incident energy.

k	r	t	s	k	r	t	s	k	r	t	s
1.1	37.93	46.60	7.73	5.1	1.30	93.74	2.48	9.1	0.59	96.86	1.27
1.2	16.00	61.69	11.16	5.2	0.93	94.44	2.31	9.2	0.41	97.32	1.14
1.3	7.28	67.02	12.85	5.3	0.76	94.97	2.14	9.3	0.33	97.65	1.01
1.4	3.86	68.90	13.62	5.4	0.68	95.41	1.95	9.4	0.30	97.91	0.90
1.5	2.77	69.68	13.78	5.5	0.67	95.79	1.77	9.5	0.30	98.12	0.79
1.6	2.78	70.35	13.44	5.6	0.71	96.08	1.61	9.6	0.32	98.28	0.70
1.7	3.31	71.38	12.66	5.7	0.80	96.28	1.46	9.7	0.38	98.37	0.63
1.8	4.18	72.92	11.45	5.8	0.96	96.39	1.33	9.8	0.47	98.40	0.56
1.9	5.64	74.70	9.83	5.9	1.26	96.32	1.21	9.9	0.65	98.31	0.52
2.1	6.51	54.24	19.62	6.1	1.08	93.71	2.61	10.1	0.49	97.15	1.18
2.2	4.97	55.63	19.70	6.2	0.84	94.16	2.50	10.2	0.37	97.42	1.10
2.3	4.15	58.84	18.50	6.3	0.75	94.70	2.28	10.3	0.33	97.69	0.99
2.4	3.70	62.57	16.87	6.4	0.73	95.20	2.04	10.4	0.32	97.94	0.87
2.5	3.52	66.68	14.90	6.5	0.74	95.62	1.82	10.5	0.34	98.13	0.76
2.6	3.56	71.56	12.44	6.6	0.79	95.96	1.62	10.6	0.38	98.27	0.68
2.7	3.81	77.49	9.35	6.7	0.89	96.24	1.43	10.7	0.44	98.34	0.61
2.8	4.35	83.63	6.01	6.8	1.07	96.44	1.24	10.8	0.53	98.35	0.56
2.9	6.12	85.87	4.01	6.9	1.42	96.38	1.10	10.9	0.68	98.24	0.54
3.1	3.94	76.78	9.64	7.1	0.88	94.95	2.09	11.1	0.44	97.65	0.96
3.2	2.33	82.66	7.51	7.2	0.61	95.45	1.97	11.2	0.32	97.84	0.92
3.3	1.62	86.32	6.03	7.3	0.49	95.92	1.79	11.3	0.27	98.03	0.85
3.4	1.24	89.01	4.87	7.4	0.43	96.38	1.59	11.4	0.24	98.20	0.78
3.5	1.03	91.04	3.96	7.5	0.40	96.83	1.38	11.5	0.24	98.37	0.70
3.6	0.96	92.52	3.26	7.6	0.40	97.25	1.18	11.6	0.24	98.52	0.62
3.7	1.04	93.47	2.74	7.7	0.43	97.60	0.99	11.7	0.26	98.66	0.54
3.8	1.35	93.80	2.42	7.8	0.51	97.82	0.83	11.8	0.31	98.76	0.46
3.9	2.24	92.99	2.38	7.9	0.72	97.80	0.74	11.9	0.41	98.79	0.40
4.1	2.02	84.46	6.76	8.1	0.73	95.13	2.07	12.1	0.40	97.63	0.99
4.2	1.33	87.65	5.51	8.2	0.54	95.62	1.92	12.2	0.30	97.78	0.96
4.3	1.04	90.35	4.31	8.3	0.45	96.23	1.66	12.3	0.26	97.99	0.87
4.4	0.91	92.43	3.33	8.4	0.40	96.85	1.37	12.4	0.24	98.22	0.77
4.5	0.91	93.92	2.58	8.5	0.39	97.40	1.10	12.5	0.24	98.45	0.66
4.6	1.07	94.85	2.04	8.6	0.41	97.83	0.88	12.6	0.25	98.66	0.54
4.7	1.51	95.05	1.72	8.7	0.49	98.08	0.72	12.7	0.28	98.83	0.45
4.8	2.57	93.75	1.84	8.8	0.68	98.04	0.64	12.8	0.34	98.91	0.37
4.9	4.06	90.31	2.81	8.9	1.14	97.39	0.74	12.9	0.50	98.79	0.36

In a numerical problem such as ours it is difficult to estimate the degree of accuracy of the final output of the program. In this case the energy balance may help a little. The incident energy carried by the exciting dominant mode must theoretically equal the sum of the reflected, transmitted and deflected energies. These are determined by the propagating modes only. Now, energy flow in the x direction is proportional to the imaginary part of $\bar{\Phi}\,\partial\Phi/\partial x$ integrated over the cross-section of the waveguide (the bar denotes complex conjugate). Let us set the total energy incident on the cross-junction per second equal to unity. Then we have for example in region *I* for the reflected energy:

$$\frac{1}{\sqrt{(k^2-1)}} \sum_{0 < n \text{ odd} < k} |T_n|^2 \sqrt{(k^2-n^2)}\,,$$

which can be rewritten as

$$\frac{1}{\text{RT}(1)} \sum_{0 < n \text{ odd} < k} |T_n|^2\,\text{RT}(n)\,.$$

Further, since the evanescent waves do not contribute to the sum, we may extend the summation to all values of odd n up to NEQ. Similar expressions hold for the energy transmitted in region *II* and for the energy deflected in regions *III* and *IV*. Since these energies are clearly of physical interest, we compute them for later output. The identifiers are EREF, ETRM and EDEF. We also calculate SUMS, a quantity that should theoretically be equal to 1. Its deviation from 1 is an indication of the degree of accuracy attained by the numerical process. It is not a check on the effect of the truncation, as was shown by Boersma: energy balance holds at any order N of truncation of the infinite system of linear equations.

All quantities of interest having been computed and stored, we then have to print one page of data (DO 17).

We can now close the double loop of the variables KU and KT controlling the value of k.

The program ends with the proper FORMAT specifications.

The subprogram that solves the linear equations was kindly provided by M. J. Geleijns of the Philips Computing Centre.

Discussion of numerical results

In *Table III* we have assembled the numerical data printed in the headline of each of the 108 pages of output. That is to say, $r = 100$ EREF is the total amount of energy reflected back into region *I* as a percentage of the incident energy, $t = 100$ ETRM is the corresponding quantity for the energy transmitted in region *II*, and $s = 100$ EDEF is the energy scattered sideways into each of the two regions *III* and *IV*. We have $r + t + 2s = 100$, apart from rounding. In

Table IV. Reflected, transmitted and scattered energies as a percentage of incident energy.

k	r	t	s	k	r	t	s	k	r	t	s
1.93	6.37	75.08	9.28	5.93	1.41	96.22	1.18	9.93	0.73	98.24	0.51
1.95	7.04	75.17	8.89	5.95	1.55	96.11	1.17	9.95	0.81	98.17	0.51
1.97	8.03	74.98	8.50	5.97	1.75	95.94	1.16	9.97	0.91	98.07	0.51
1.99	9.84	73.98	8.09	5.99	2.10	95.58	1.16	9.99	1.09	97.88	0.52
2.01	10.59	61.94	13.74	6.01	1.95	94.02	2.02	10.01	0.96	97.16	0.94
2.03	8.95	57.71	16.67	6.03	1.56	93.72	2.36	10.03	0.75	97.06	1.09
2.05	7.98	55.81	18.10	6.05	1.36	93.64	2.50	10.05	0.64	97.05	1.15
2.07	7.29	54.83	18.94	6.07	1.22	93.64	2.57	10.07	0.57	97.08	1.18
2.93	7.64	83.43	4.46	6.93	1.60	96.19	1.10	10.93	0.75	98.16	0.54
2.95	9.51	79.26	5.62	6.95	1.77	95.97	1.13	10.95	0.81	98.08	0.55
2.97	13.02	70.12	8.43	6.97	2.00	95.58	1.21	10.97	0.89	97.96	0.57
2.99	20.69	46.98	16.16	6.99	2.36	94.85	1.39	10.99	1.02	97.75	0.61
3.01	12.07	56.14	15.89	7.01	1.86	94.17	1.98	11.01	0.84	97.45	0.86
3.03	7.86	66.28	12.93	7.03	1.41	94.46	2.07	11.03	0.66	97.50	0.92
3.05	6.05	70.85	11.55	7.05	1.19	94.63	2.09	11.05	0.56	97.54	0.95
3.07	4.97	73.76	10.63	7.07	1.04	94.77	2.10	11.07	0.50	97.59	0.96
3.93	2.79	92.26	2.48	7.93	0.85	97.70	0.73	11.93	0.47	98.76	0.39
3.95	3.31	91.50	2.60	7.95	0.96	97.58	0.73	11.95	0.52	98.71	0.38
3.97	4.06	90.35	2.79	7.97	1.14	97.38	0.74	11.97	0.60	98.64	0.38
3.99	5.39	88.27	3.17	7.99	1.44	96.99	0.79	11.99	0.73	98.49	0.39
4.01	4.76	82.66	6.29	8.01	1.39	95.44	1.59	12.01	0.71	97.82	0.74
4.03	3.48	82.54	6.99	8.03	1.10	95.15	1.88	12.03	0.57	97.68	0.87
4.05	2.85	82.94	7.10	8.05	0.94	95.07	1.99	12.05	0.50	97.63	0.93
4.07	2.44	83.51	7.03	8.07	0.84	95.07	2.05	12.07	0.45	97.62	0.97
4.93	4.10	89.71	3.10	8.93	1.37	96.96	0.84				
4.95	3.92	89.70	3.19	8.95	1.54	96.59	0.93				
4.97	3.60	90.00	3.20	8.97	1.73	96.16	1.06				
4.99	3.15	90.61	3.12	8.99	1.89	95.68	1.22				
5.01	2.29	92.29	2.71	9.01	1.28	95.99	1.37				
5.03	1.92	92.86	2.61	9.03	0.97	96.32	1.36				
5.05	1.68	93.20	2.56	9.05	0.80	96.52	1.34				
5.07	1.50	93.45	2.53	9.07	0.70	96.68	1.31				

passing we should note that SUMS = 1.000 000 for each of the 108 values of k. In fact, a second run of the program for different values of k (in the vicinity of integers; see below) has revealed that SUMS is unity to about 10 places of decimals throughout.

Inspection of Table III shows the occurrence of resonance phenomena for k near an integer. As can be seen from eqs. (14) and (15), if k is near the odd number $2N-1$ then A_{2N-1} and B_{2N-1} are possibly large in absolute value because the factor in front of the summation sign becomes infinitely large for $k=2N-1$. In contrast to this, D_{2N-1} is not necessarily large, as follows from eq. (16). Consequently, R_{2N-1}, T_{2N-1} and S_{2N-1} may be unbounded as $k \rightarrow 2N-1$. Similarly, if k is near the even number $2N$, then C_{2N} becomes large in absolute value, and S_{2N} may be unbounded if $k \rightarrow 2N$. However, since the absolute square of these quantities times $\sqrt{(k^2-n^2)}$ must be finite (for $n = 2N-1$ and $n = 2N$, respectively) because it represents energy, the infinities are at most of order $(k-n)^{-1/4}$ as $k \rightarrow n$.

The critical index, $2N-1$ or $2N$, divides the range of k into two parts: for k less than the critical index (i.e. below cut-off) the mode in question is evanescent, for k greater than the critical index (i.e. above cut-off) the mode propagates. We should therefore be on the alert for infinities in the case of the nth mode if n is equal to k at cut-off. If k is an integer infinities other than cut-off may also occur, although the resonance effect, in that the amplitude of the mode shows large peaks or deep minima as a function of k, is perhaps less pronounced.

It should be noted that the denominators $m^2+n^2-k^2$ in eqs. (14) to (17) are quite harmless. They indicate a sort of resonance with region V, the junction proper. If the denominator becomes zero (e.g., $k^2=2$, $n=m=1$) it is compensated by the very factor in front of the summation sign, which also goes to zero.

For accurate interpolation the data so far obtained is insufficient if k is in the vicinity of an integer. Therefore the program was run a second time so as to produce 88 pages of output relating to k = 1.93, 1.95, 1.97, 1.99, 2.01, 2.03, 2.05, 2.07 in the neighbourhood of k = 2, and so on. The corresponding energy cross-sections have been assembled in *Table IV*.

For values of k up to 4 these cross-sections have also been plotted in *fig. 6*. The reflection cross-section shows sharp peaks at k = 2, 3 and 4. On the other hand, the transmission cross-section has what look like inflexion points at even-integer values of k and shows a sharp minimum at k = 3. Much the same holds for the deflection cross-section. The peaks at k = 3 are seen to be very sharp: the data available is not sufficient to estimate the extreme values within five per cent.

In *fig. 7* curves are shown for AREF (1), ADEF (1) and ATRM (1), in the order from top to bottom. The first curve should tend to infinity if k tends to 1 from above. The peaks at k = 2, 3, 4, . . . seem finite. The curves for deflection and transmission again show a difference in behaviour for k an odd or even integer.

There is no indication that ATRM (1) would tend to infinity as k tends to 1.

No attempt has been made to illustrate the full contents of the numerical tables by graphical means. Both the author and Dr. Butterweck are in possession of the

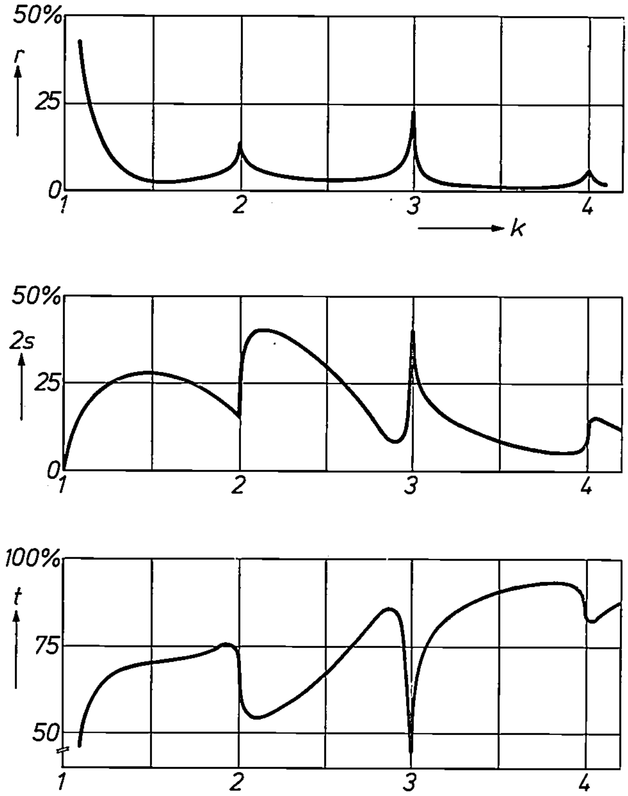


Fig. 6. The reflected (*r*), transmitted (*t*) and deflected (2*s*) fractions of the incident energy as a function of the wave number *k*.

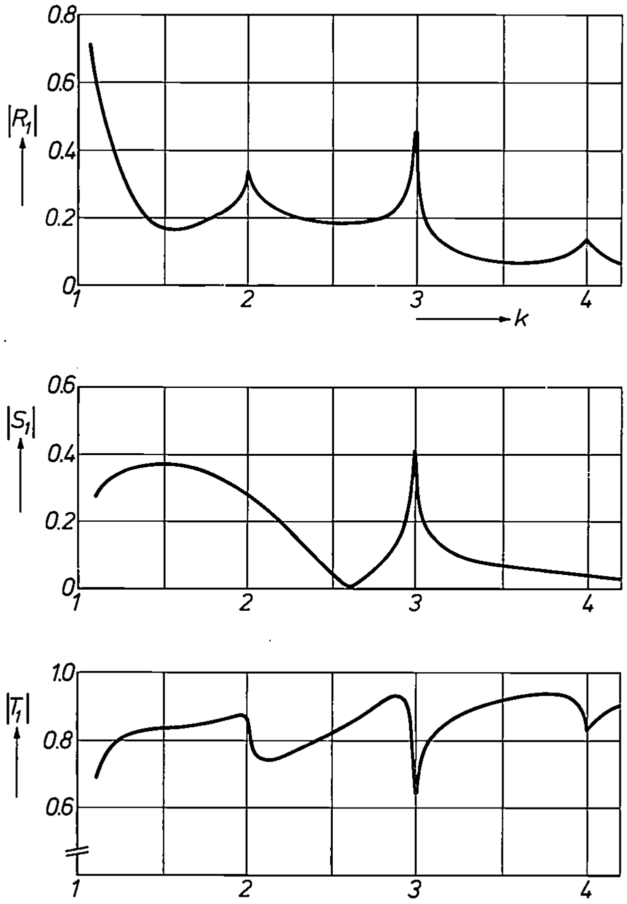


Fig. 7. As fig. 6 but now for the amplitude of the dominant modes.

196-page table, which can be used for further reference if need be.

Run time on the CDC-3600/3200 amounted to less than 2 hours.

Appendix: Fortran CDC-3600 program

```
   PRØGRAM CRØSS
   DIMENSIØN RT(60), ER(60), F1(30), F2(30), F(30),
             MAT(30, 30), X(2,30), Y(2,30), REF(15),
             TRM(15), DEF(30), XX(30), AREF(15),
             ATRM(15), ADEF(30), PREF(15), PTRM
             (15), PDEF(30)
   TYPE CØMPLEX CI, RT, ER, F1, F2, F, SUM, MAT,
             X, Y, REF, TRM, DEF, XX
   TYPE REAL K, KK

   NCP = 15
   NEQ = 2 *NCP
   PI  = 3.141592654
   CI  = (0.,1.)
   DØ 1  KU = 1,12
   DØ 1  KT = 1,9
   K   = KU + 0.1 *KT
   KK  = K*K

   M = 2*NEQ
   DØ 2 N = 1,M
   AR = KK - N *N
   IF(AR.GT.0) GØTØ 21
   RT(N) = -PI *SQRTF(-AR)
   GØTØ 22
21 RT(N) = PI *SQRTF(AR) *CI
22 ER(N) = CEXP(RT(N))
 2 CØNTINUE

   DØ 3 NC = 1,2
   J = NC - 1

   DØ 4 N = 1,NEQ
   NØ = 2 *N - 1
   NE = NØ + J
   F1(N) = -2 *NØ / RT(NØ) *(1. + (-1) * *J *ER(NØ))
   F2(N) = -2 *NE / RT(NE) *(1. + ER(NE))
   F(N)  = -F2(N) *(ER(1) + (-1) * *J) / (2. *(NE *NE
         +1. -KK))
 4 CØNTINUE

   DØ 5 N = 1,NEQ
   NE = 2 *N - 1 + J
   EN = NE *NE - KK
   DØ 6 M = 1,NEQ
   ME = 2 *M - 1 + J
   EM = ME *ME - KK
   SUM = 0
   DØ 7 I = 1,NEQ
   L = 2 *I - 1
   EL = L *L
   SUM = SUM + L *F1(I)/((EL + EN) *(EL + EM))
 7 CØNTINUE
   MAT(N,M) = F2(N) *ME *SUM
 6 CØNTINUE
   MAT(N,N) = MAT(N,N) - 1.
 5 CØNTINUE
```

```
      CALL CLINEQ(MAT, XX, F, NEQ)

      DØ 8 N = 1,NEQ
    8 X(NC,N) = XX(N)

      DØ 9 N = 1,NEQ
      NØ = 2*N - 1
      EN = NØ*NØ - KK
      SUM = 0
      DØ 10 I = 1,NEQ
      L = 2*I - 1 + J
      EL = L*L
      SUM = SUM + L*X(NC,I)/(EL + EN)
   10 CØNTINUE
      Y(NC,N) = F1(N)*SUM
    9 CØNTINUE

    3 CØNTINUE

      DØ 11 N = 1,NCP
      REF(N)  = Y(1,N) - Y(2,N)
      TRM(N) = Y(1,N) + Y(2,N)

      NØ = 2*N - 1
      NE = 2*N
      DEF(NØ) = X(1,N)
      DEF(NE) = - X(2,N)
   11 CØNTINUE
      TRM(1) = TRM(1) + ER(1)

      DØ 12 N = 1,NCP
      AREF(N)  = CABS(REF(N))
      PREF(N)  = CANG(REF(N))
      ATRM(N) = CABS(TRM(N))
      PTRM(N)  = CANG(TRM(N))
   12 CØNTINUE
      DØ 13 N = 1,NEQ
      ADEF(N) = CABS(DEF(N))
      PDEF(N) = CANG(DEF(N))
   13 CØNTINUE

      SUM = 0
      DØ 14 N = 1,NCP
   14 SUM = SUM + AREF(N)*AREF(N)*RT(2*N - 1)
      EREF = SUM/RT(1)
      SUM = 0
      DØ 15 N = 1,NCP
   15 SUM = SUM + ATRM(N)*ATRM(N)*RT(2*N - 1)
      ETRM = SUM/RT(1)
      SUM = 0
      DØ 16 N = 1,NEQ
   16 SUM = SUM + ADEF(N)*ADEF(N)*RT(N)
      EDEF = SUM/RT(1)
      SUMS = EREF + ETRM + 2.*EDEF
```

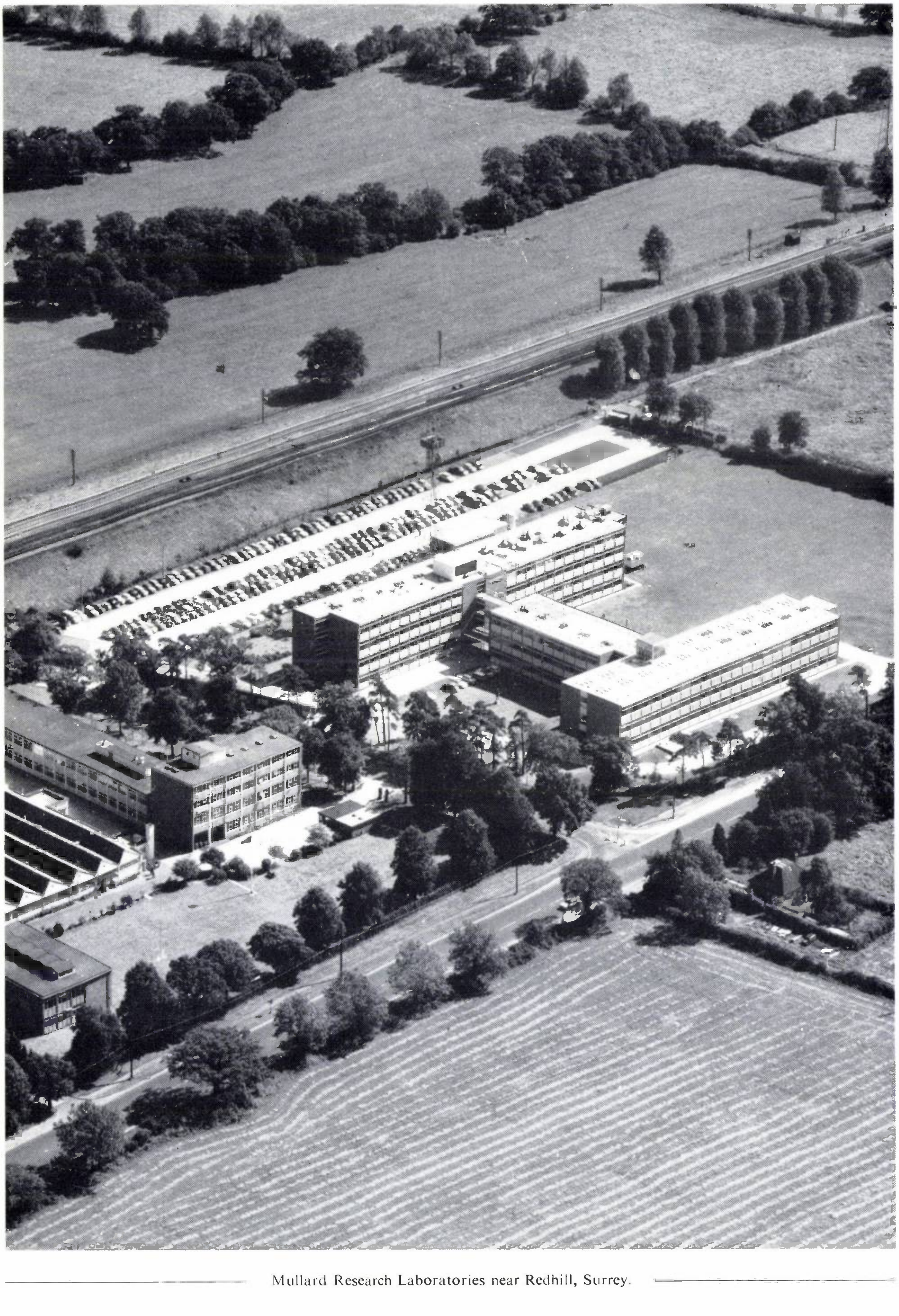
```
      PRINT 100, K, EREF, ETRM, EDEF, SUMS
      PRINT 101
      ZERØ = 0.

      DØ 17 N = 1,NCP
      M = 2*N - 1
      PRINT 102, M, AREF(N), ATRM(N), ADEF(M),
             PREF(N), PTRM(N), PDEF(M)
      MM = M + 1
      PRINT 102, MM, ZERØ, ZERØ, ADEF(MM), ZERØ,
             ZERØ, PDEF(MM)
   17 CØNTINUE

    1 CØNTINUE
      STØP

  100 FØRMAT (1H1,//,10X,*K = *,F5.1,5X,*EREF = *,
              F9.6,5X,*ETRM = *,F9.6,5X,*EDEF = *,
              F9.6,5X,*SUMS = *,F9.6,//)
  101 FØRMAT (10X,*N*,10X,*AREF*,10X,*ATRM*,
              10X,*ADEF*,10X,*PREF*,10X,*PTRM
              *,10X,*PDEF*,/)
  102 FØRMAT (9X,I2,6(5X,F9.6))
      END
```

```
      SUBRØUTINE   CLINEQ(A,X,B,N)
      DIMENSIØN A(N,N),X(N),B(N)
      TYPE   CØMPLEX A,X,B,DIVR
      DØ 5 I = 1,N
      II = I + 1
      DØ 2 J = II,N
    2 A(I,J) = A(I,J)/A(I,I)
      B(I) = B(I)/A(I,I)
      A(I,I) = (1.,0.)
      DØ 4 K = II,N
      DIVR = A(K,I)
      DØ 3 L = I,N
    3 A(K,L) = A(K,L)/DIVR - A(I,L)
    4 B(K) = B(K)/DIVR - B(I)
    5 CØNTINUE
      MM = N + 1
    6 X(MM - 1) = B(MM - 1)
      DØ 7 M = MM,N
    7 X(MM - 1) = X(MM - 1) - X(M)*A(MM - 1,M)
      IF(MM.EQ.2) GØTØ 8
      MM = MM - 1
      GØTØ 6
    8 RETURN
      END
```



Mullard Research Laboratories near Redhill, Surrey.

Acoustic surface-wave filters

R. F. Mitchell

Introduction

High-frequency acoustic waves have been used in signal processing for many years, in such familiar components as crystal oscillators, filters and resonators. In the main these devices rely on the use of sharp resonances produced by the repeated reflection of bulk acoustic waves in regular structures such as plates, bars and discs. The disturbance usually extends over much or all of the resonating structure.

In recent years it has come to be realized that many signal processing functions which do not require sharp resonances can be conveniently achieved by making use of the characteristics of a different kind of acoustic disturbance: the acoustic surface wave. Much research into the characteristics of these waves has been carried out at centres that include Philips Research Laboratories, Philips Forschungslaboratorium Hamburg and Mullard Research Laboratories (MRL). Acoustic surface waves share many of the characteristics of bulk waves, such as dispersion-free propagation, but the disturbance is confined to a small region beneath a surface. The waves can therefore be readily launched or received from a number of points simultaneously and this makes it possible to process signals in a way which is quite different from those used in conventional acoustic devices.

Many useful devices can be made in this way, such as delay lines, pulse compressors, correlators and convolvers, but among the more promising are a class of simple analogue filters. In this article we shall describe some of the work that has been done at this laboratory over the last two or three years on the theory and practice of making such filters. Our attention has been directed primarily at the possibility of making a television intermediate-frequency filter suitable for hybrid integration, and the results obtained to date are discussed later in this article. But the techniques described can be applied to a range of broadband filters for the frequency range 10-1000 MHz with bandwidths of 100 kHz and above. In this article we shall describe only the general principles of surface-wave filters; further details, particularly of the mathematical techniques, have been published elsewhere [1]. We start by briefly outlining the characteristics of acoustic surface waves themselves.

Surface waves in solids

Surface waves on liquid media are a familiar part of everybody's experience of the natural world. But the fact that similar waves can exist on solid materials was not recognized until they were predicted mathematically by Lord Rayleigh in 1885. Experimental confirmation of their existence as a product of earthquakes came later still. These waves, often called Rayleigh waves after their discoverer, travel much faster than their liquid counterparts (typically 1-4 km/s) and involve very much smaller surface motions. Their particle motions are very similar to those found in liquids (*fig. 1*), but with the significant difference that the elliptical motion is retrograde in the Rayleigh wave but in a forward direction in liquids. The differences between the two

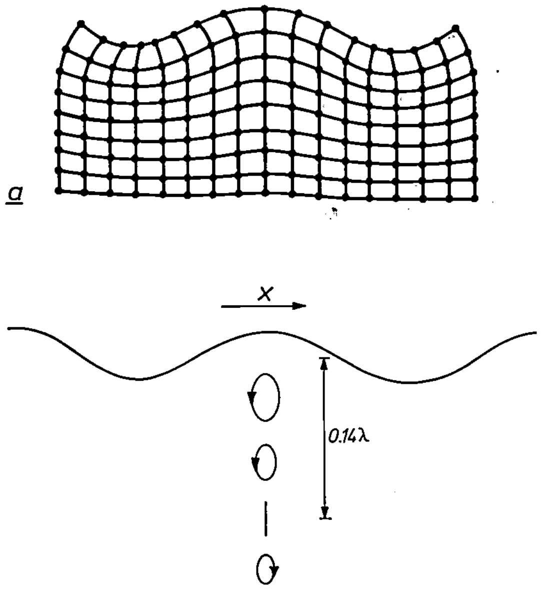


Fig. 1. *a*) To illustrate the displacement caused by an acoustic surface wave (a Rayleigh wave) propagating in an isotropic medium. This section shows the displacements of particles that were originally at the intersections of a rectangular "grid". The wavelength λ at an exciting frequency of 30 MHz would typically be about 0.1 mm. *b*) The elliptical nature of the particle motion in an isotropic medium. At the surface the particle traverses its ellipse in the opposite direction to the direction of propagation x. The elliptical movement reverses direction 0.14λ below the surface, and the disturbance becomes negligible within about 3λ of the surface.

R. F. Mitchell, Ph. D., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] R. F. Mitchell, Generation and detection of sound by distributed piezoelectric sources, Philips Res. Repts. Suppl. 1972, No. 3.

kinds of wave arise because the potential energy in a wave on a liquid is gravitational, whereas in a solid it is elastic.

The disturbance in a Rayleigh wave dies away exponentially from the surface becoming negligible after two or three wavelengths. This means that for device purposes it is sufficient to provide a substrate of two or three wavelengths thick to obtain Rayleigh-wave propagation indistinguishable from that found on a complete half-space.

Rayleigh-wave particle motions in anisotropic materials are more complicated than those shown in fig. 1 although the overall features are usually very similar. However the wave-normal and power-flow directions are not in general in exactly the same direction, as in the isotropic case.

The wavelength of Rayleigh waves is typically 1 mm at 3 MHz, 10 μm at 300 MHz and 1 μm at 3 GHz and so surface-wave devices themselves may also be expected to fall roughly in this range of dimensions. They are thus typically of the order of 10^5 times smaller than their electromagnetic counterparts.

Analogue surface-wave filters

General aspects

Acoustic surface waves can be conveniently generated and detected using the interdigital transducer structure shown in *fig. 2* on a strongly piezoelectric substrate. The metal electrodes will usually be several wavelengths long, in order to launch a "plane" (strictly: straight-crested) surface wave, and will be spaced half a wavelength apart at the fundamental frequency of operation. These dimensions mean that the electrodes can be made by standard photolithographic techniques for frequencies up to several hundred megahertz, providing the substrate surface is adequately prepared.

Fig. 3 is a schematic diagram of a simple delay line using one launching transducer and one receiver. This is the basis of a surface-wave filter. The filtering derives from the fact that the launching and receiving transducers themselves show strongly frequency-dependent behaviour. The frequency response of the total filter is merely the product of the responses of the two transducers.

The schematic diagram of fig. 3 indicates many of the attractive features of surface-wave filters:

1. Their small size; dimensions are typically a few tens of acoustic wavelengths.
2. Their convenience: all that is required is an appropriate piezoelectric substrate with one surface polished smooth to the order of an acoustic wavelength. Provided the substrate is more than a few wavelengths thick and sufficiently large to house the filter its dimensions are otherwise not critical. And it may be mounted — by glueing into a holder, for example — without affecting the performance.
3. Their simplicity: the only parts of the filter which must be accurately defined are the electrode structures, which can be produced by standard means.

In order to understand the behaviour of these devices it is convenient to approach the theory in two steps: first a simple analysis, making a number of approximations, which points clearly to the general principles of their behaviour; second a more detailed treatment, including some of the secondary effects. Surface waves are complex phenomena and even such straightforward problems as the reflection of a surface wave at a corner defy mathematical analysis except under very simple conditions. When elastic and piezoelectric anisotropy are introduced, together with a periodic loading of the surface by electrode structures, the problems become formidable. In fact a full analysis of the behaviour of a finite interdigital transducer making no approximations has never yet been achieved. We therefore make use of a number of approximations and arguments by analogy to produce an adequate model of the situation.

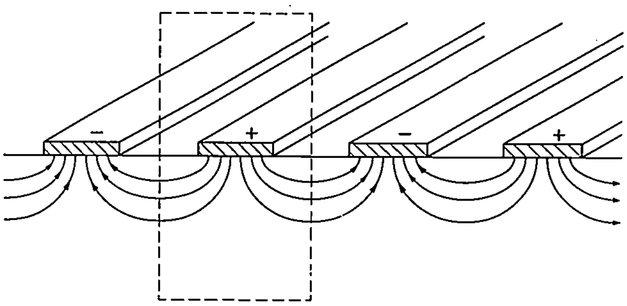


Fig. 2. Interdigital transducer structure for generating or detecting acoustic surface waves. The metal electrodes, deposited on a strongly piezoelectric substrate, are usually several wavelengths long and are spaced at a half-wavelength at the frequency of operation. Alternate electrodes are excited in opposite sense. This structure is conveniently made by standard photolithographic techniques.

The dashed rectangle encloses the cross-section of the elementary unit chosen for detailed analysis later.

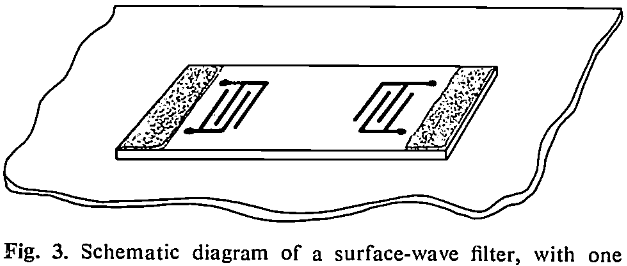


Fig. 3. Schematic diagram of a surface-wave filter, with one transducer for launching and another for receiving. The device acts as a filter because of the marked frequency dependence of the two transducers. Its frequency response is merely the product of the transducer responses. The mechanical design is simple and the device is very small; its dimensions being typically a few tens of acoustic wavelengths. The shaded patches at the ends represent absorbent material to prevent unwanted reflections.

Simple model

The simplest and most straightforward approach is to regard every electrode of the transducer as a simple source of surface waves. The following assumptions are made:

1. The separate sources generate the same surface-wave amplitude at all frequencies — that is, they have no individual frequency response.
2. There is no acoustic attenuation.
3. The sources are independent.
4. There are no diffraction effects; only "plane" waves are launched.

If these assumptions are made the transducer is just a simple end-fire acoustic aerial and the wave which is launched is the simple sum of the contributions from the individual electrodes. The interference of these component waves is constructive at some frequencies and destructive at others and so the device shows a pronounced frequency selectivity. A simple mathematical analysis brings out the salient features of the performance.

Let the amplitude of surface wave generated at x_0 be $B(x_0)$, with the wave propagating along the surface in the x-direction, perpendicular to the electrodes. Then, since the penetration of the wave into the substrate is constant we may write the amplitude of the total wave reaching a point x, outside the transducer as

$$A(x) = c \int_a^b B(x_0) \exp\left[\mathrm{j}\{\omega t - k(x - x_0)\}\right] \mathrm{d}x_0 ,$$

where the transducer stretches from $x_0 = a$ to $x_0 = b$, and c is a constant. Now since $B(x_0)$ is zero outside ab the limits can be extended to $\pm\ \infty$ yielding

$$A(x) = c.\ \mathcal{F}(B) \exp \mathrm{j}\ (\omega t - kx),$$

where

$$\mathcal{F}(B) = \int_{-\infty}^{\infty} B(x_0) \exp\left(\mathrm{j}kx_0\right) \mathrm{d}x_0$$

is the Fourier transform of B. Thus the output is a travelling surface wave whose amplitude is the Fourier transform of the source distribution, $B(x_0)$. The transform is in general complex, giving both the amplitude and phase of the wave. This simple relationship makes it easy to see how the passband of a filter is related to the electrode structure. Consider first the case of a simple, uniform transducer, which can be represented, according to this simple model, by a series of simple sources, one for every electrode as in *fig. 4*. The signs of the sources alternate with the voltage applied to the electrodes. The Fourier transform of the resulting source distribution can be readily found by observing that the distribution is equivalent to a rectangular envelope function sampled at regular intervals. It is well known that the Fourier transform of such a function is the transform of the envelope curve repeated at harmonic intervals whose spacing is determined by the separation of the samples. Thus since the Fourier transform of a rectangular function is of the form (sin x)/x the transform of the function of fig. 4 is as shown in *fig. 5*. This is therefore the expected frequency response of a uniform transducer. The expectation is broadly confirmed by experiment.

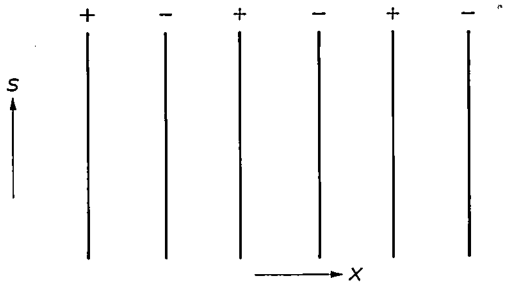


Fig. 4. A simple uniform transducer can be represented by a series of simple sources, with one source for each electrode. The signs of the sources alternate with the voltage applied to the electrodes. The strength S of each source is represented by the length of the line. Here the sources are all of the same intensity and the envelope curve is a rectangular function of the linear coordinate x.

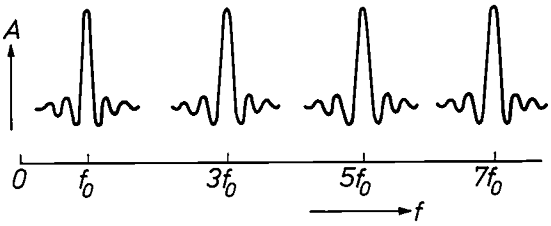


Fig. 5. With the simple transducer model the predicted frequency response is proportional to the modulus of the Fourier transform of the envelope curve, and repeats at harmonic intervals whose spacing is determined by the spacing of the elements. Since the Fourier transform of a rectangular function is of the form (sin$x$)/x the transform of the function of fig. 4 and hence the total frequency response is as shown here. (Note that this diagram shows the real part of the Fourier transform.) A amplitude, f frequency.

The simple relationship between the envelope of the source distribution and the frequency response of the device is convenient in deciding how the source distribution must be changed to produce a particular frequency response. Provided the bandwidth of the device is fairly narrow the harmonic responses do not interfere with each other and the fundamental response can be considered in isolation. *Fig. 6* shows a set of source distributions and the corresponding fundamental frequency responses, calculated from the transform of the envelope.

Possible filter response

The simple analysis in terms of Fourier transforms provides the following general principles, provided, as was assumed, the electrodes are equally spaced.

1. The frequency of operation depends on the spacing of the electrodes: the closer they are, the higher the frequency. We have already mentioned that the fundamental response occurs at the frequency for which the electrode spacing is half a wavelength.

2. The relative bandwidth of the fundamental response is inversely proportional to the number of sources; the more sources there are, the narrower the bandwidth. Since the centre frequency depends on the spacing of the sources and the surface-wave velocity it follows that the absolute bandwidth of the device depends only on its size (in the direction of propagation) and the velocity. For example, if the velocity is 3000 m/s the bandwidth of a one-centimetre device will be about 300 kHz, that of a two-centimetre device 150 kHz and so on. This relationship between bandwidth and size is one of the major drawbacks of surface-wave filters since narrow-bandwidth filters cannot be made without very large substrates. This problem would be removed if a convenient and efficient means of reflecting surface waves were available — which it is not, at the moment.

3. The Fourier transform of a symmetric function is real; that of an antisymmetric function is purely imaginary. Any symmetric or antisymmetric source distribution will therefore generate a signal with a linear phase/frequency relationship. A wide range of amplitude responses is possible with linear phase because surface-wave filters are not minimum-phase networks; amplitude variations do not necessarily imply phase variations. To achieve nonlinear phases, nonsymmetrical arrays or varying electrode periods must be used.

4. The Fourier transform of a real function is Hermitian[*]. This means that, within the restrictions of the model, only symmetrical frequency responses are possible. Departures from this will imply a breakdown of the simple model.

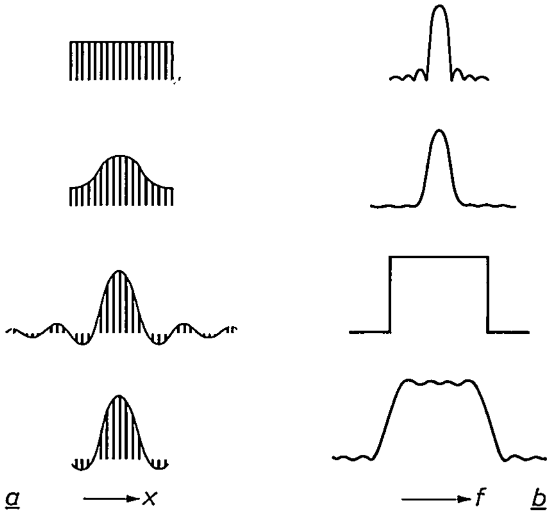


Fig. 6. Some simple source-strength distributions (*a*) with Fourier transforms (shown here as the modulus) of their envelopes (*b*). The Fourier transforms represent the fundamental frequency response. In the source distributions the elements are again assumed to be of alternating sign, and also to change sign where the envelope curves go negative.

Synthesis of frequency response

The Fourier-transform approach gives a useful guide to what frequency responses are physically possible. In theory the same method could be used to determine what envelope of sources is needed to produce a particular frequency response by taking the reverse transform of the response curve. The trouble is that the Fourier transform of most functions extends to plus and minus infinity; this means that the source distribution required to produce exactly some required frequency response will generally contain a large, perhaps infinite, number of sources — an embarrassing structure to manufacture. For a practical device the source distribution must be truncated to leave only a few terms while causing the minimum of departure from the required curve.

This is a difficult problem and forms a study in its own right. It turns out to be best to compute the frequency-response curve of a source distribution numerically without explicitly using transform analysis and then to optimize the result. With care and cunning a source distribution can be found to give the required performance, including asymmetric responses and nonlinear phase/frequency performance if required. At MRL the optimization is done by a "man-machine interaction" process although a computer program has now been written which will automatically compute the source strengths required to achieve a required response curve, for the linear-phase case.

Weighting methods

So far we have discussed the design of surface-wave filters purely in terms of source distributions, with no mention of what methods can be used in practice to obtain the required source strengths. There are two main ways of arranging the strength distribution or "weighting" of the source array: by varying the widths of the electrodes (*fig. 7*) and by varying their lengths (*fig. 8*). The width-weighting technique has the great advantage that it reduces diffraction effects to a mini-

[*] A Hermitian function is one whose real part is symmetrical about the origin and whose imaginary part is antisymmetrical. The modulus of such a function is symmetrical.

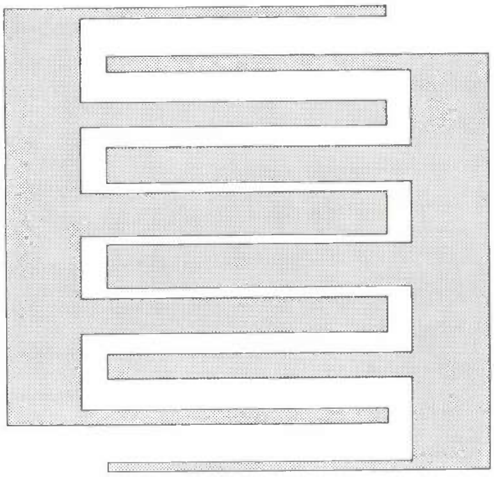


Fig. 7. One way of obtaining an appropriate distribution of source strength ("source weighting") is by varying the width of the "teeth" that form the elements. This method is only suitable at low frequencies because it requires rather narrow teeth. It has the advantage that diffraction is much reduced.

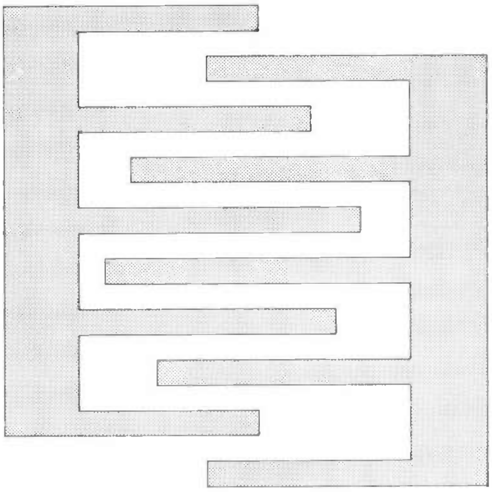


Fig. 8. The distribution of source strength can also be varied by varying the lengths of the elements.

mum. Its disadvantage is that small sources require very narrow electrodes and so demand correspondingly accurate photolithography. This means that width-weighting can be used only in low-frequency devices; at 30 MHz for example, the range of source strengths that can be obtained within the limits of standard photolithography is only 2.5 : 1. The length-weighting approach makes fewer demands on the fabrication accuracy but much increases the effects of diffraction. These effects are, of course, most severe in filters whose "aperture" — the length of the longest electrode — is small; in these diffraction adds a whole new range of design problems. For this reason attention has been concentrated on width-weighting in most of our television i.f. filter work, although it is intended to move to length-weighting in the near future.

Full model

Filter synthesis is best done using the simple model, but the performance of the resulting filter will differ somewhat from that predicted by the model because of the intrusion of the secondary effects which were expressly ignored. *Fig. 9* shows a comparison between the actual performance of a width-weighted filter and a prediction based on this simple model. The agreement is quite good but there are three important discrepancies: the depths of the minima are quite wrongly predicted, there is a tilt of the response in the central region which is not accounted for, and there is a progressive deterioration of the prediction in the side-lobe regions. It is important to refine the simple model so that these departures can be accounted for and "designed out" before the filters are made. A more complicated model is therefore needed which removes some of the assumptions made in the simple one. This full model must take account of the effects of diffraction, acoustic attenuation, interactions between the individual sources and the variation of the output of each source with frequency. We shall now discuss briefly how we have tackled the problem of predicting these effects.

Frequency response of individual sources

To determine the frequency response of an individual electrode we have adapted the published work of

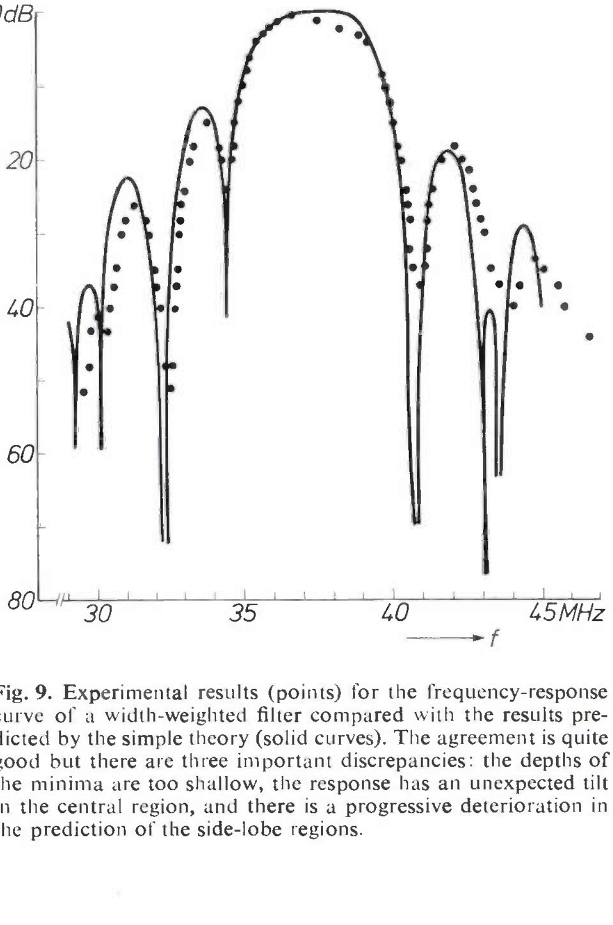


Fig. 9. Experimental results (points) for the frequency-response curve of a width-weighted filter compared with the results predicted by the simple theory (solid curves). The agreement is quite good but there are three important discrepancies: the depths of the minima are too shallow, the response has an unexpected tilt in the central region, and there is a progressive deterioration in the prediction of the side-lobe regions.

H. Engan [2]. The approach used is to make a Fourier expansion of the electric field produced by an interdigital array in components which, like those of the Rayleigh wave, vary sinusoidally along the surface but decay exponentially below it. The generation of surface waves at a particular frequency is associated with the electric-field component of the corresponding spatial frequency. Thus a surface wave of wavelength λ is associated with the electric-field component whose repeat distance is λ. The hypothesis is made that the amplitude of the acoustic wave generated is proportional to the amplitude of the corresponding electric-field component. Engan's original analysis was for an infinite array of interdigital electrodes. We have adapted this for a single electrode by making a similar expansion for the repeat unit shown in the dashed rectangle in fig. 2 — a single electrode forming part of an infinite array. A transducer containing electrodes of varying width can be represented approximately by joining together repeat units of this sort.

Fig. 10 shows the frequency response, determined in this way, of two isolated units taken from arrays with different width/gap ratios. The response does not vary very rapidly with frequency and so if the electrode dimensions are chosen so that the required source strengths are obtained at the fundamental frequency, the choice will remain approximately correct over a considerable bandwidth. But the variation in this region is in opposite directions for wide and narrow electrodes and so the effect on the frequency response of a width-weighted filter, containing both sorts of electrode, may be quite appreciable. At frequencies much higher than the fundamental the differing responses of the various electrodes causes a complete change of the relative source strengths: some electrodes may not generate at all and some may generate in antiphase. The third harmonic response of a width-weighted transducer is therefore found to be completely different to the fundamental. In length-weighted transducers the situation is simpler because the frequency responses of all electrodes are the same; the response calculated from the simple model merely has to be multiplied by the response of a single electrode. This has very little effect on the fundamental passband but does govern the relative strengths of the harmonics. It turns out that if the electrodes have the same width as the gaps between them, the third harmonic response is zero.

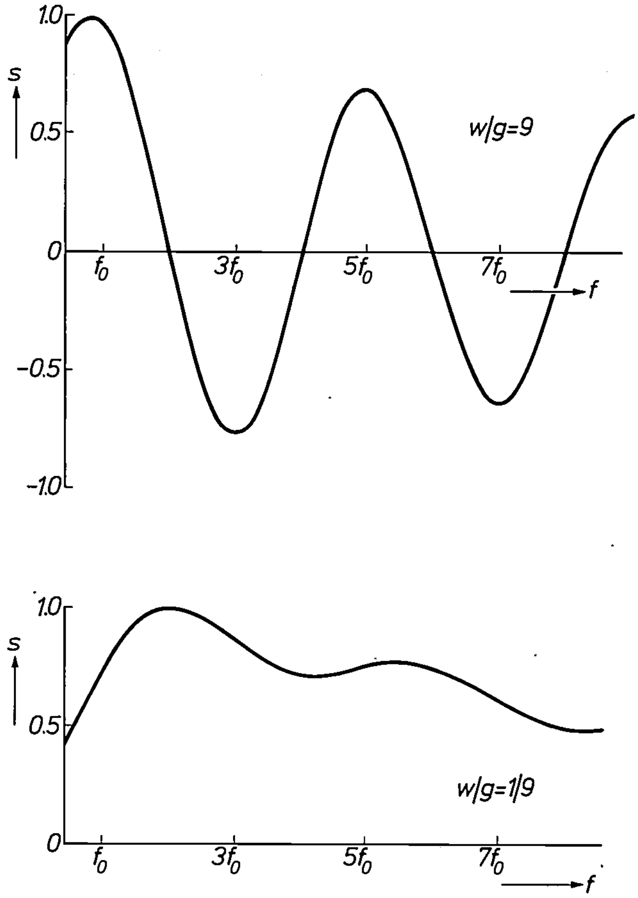


Fig. 10. Frequency response of two isolated units from arrays with different width/gap ratios w/g. S amplitude of surface wave. The response does not vary very rapidly with frequency, so that if the electrode dimensions are chosen to give the required source strengths at the fundamental frequency f_0, this choice will remain approximately correct over a considerable bandwidth. But since the variation in the region of f_0 is opposite for wide and narrow electrodes the effect on the frequency response of a width-weighted transducer, with both kinds of electrodes, is considerable. For length-weighted devices, the simple-model response merely has to be multiplied by the response of a single electrode. With $w/g = 1$ the third harmonic response is zero.

Interactions between sources

There can be two kinds of interactions between the sources. The first is a purely acoustic effect in which a surface wave from one electrode is partially reflected by a neighbouring one because of the change in electrical and mechanical conditions it encounters as it travels from the free surface to the metallized region. The second is a specifically piezoelectric effect in which part of the wave generated by one electrode is converted into an electrical signal by a succeeding one and then re-radiated. This again is a kind of reflection but it is somewhat affected by the external electrical circuit. To include the effects of interactions we have used a development of an idea first suggested by W. R. Smith *et al.* [3]. The idea is that since the propagation of surface waves is basically a one-dimensional process (since apart from diffraction the wave spreads out only in the propagation direction) the generation of surface waves by a periodic transducer can be modelled by analysing the behaviour of an array of *bulk* acoustic-wave transducers connected together in series acoustically but in parallel electrically (*fig. 11*). There is one transducer for every electrode of the surface-wave device and the piezoelectric activity of the transducers is adjusted to a value appropriate for the generation of surface waves (much lower than that for bulk waves).

In order to ensure that the model reproduces interaction effects correctly it is important to choose the right type of bulk-wave transducer. Experiments on simple devices [4] led us to adopt a piezoelectric plate transducer as the basic unit and so to choose the model of fig. 11. It is valid for either width- or length-weighted devices. The dimension x_1 is the same for all transducers and the dimension perpendicular to the page is proportional to the length of the corresponding electrode. The dimension d is chosen so that the static charge developed on each transducer in the model is the same as that developed on the corresponding electrode of the surface-wave device. The transducers are connected together by an acoustic waveguide which represents the action of the surface in ensuring that all the acoustic energy generated by a transducer with a large value of d in the model passes through even the transducer with the smallest value. Finally the frequency response of the individual electrodes is accounted for by making the coupling coefficient of each transducer vary with frequency so that the acoustic output follows the correct form (such as fig. 10).

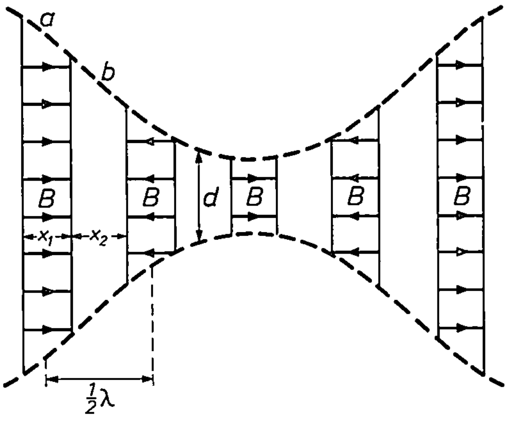


Fig. 11. Schematic cross-section of a bulk-acoustic-wave equivalent model (wavelength λ) for a length- or width-weighted surface-wave transducer. This equivalent is used to study interactions. The dashed lines represent the upper and lower surfaces of an acoustic waveguide. The bulk transducers B (one for each element) are in series acoustically and in parallel electrically. The arrows represent the electric field. The regions a and b are acoustically matched and the ratio x_1/x_2 is immaterial. Dimension d is chosen so that the static charge developed on each transducer is the same as the charge on the corresponding element of the surface-wave device: the voltage applied to each electrode is half that on the corresponding electrode.

The analysis of this model is done by solving a series of simultaneous equations to determine the electric flux in every electrode. Acoustic attenuation can be easily allowed for and so can the effects of diffraction *within* the transducer, although this does not appear to be necessary at the moment.

This model is not very pleasing aesthetically but it gives an excellent prediction of the behaviour of practical devices, as may be seen from *fig. 12,* which is a comparison between theory and practice for a width-weighted filter. The deficiencies of the simple model have now been removed. The passband tilt is found to be due to interaction effects, the depths of the traps are governed by attenuation and the side-lobe levels have been altered by the frequency response of the individual electrodes.

This agreement between theory and practice is perfectly adequate for the design of width-weighted filters, but before describing some applications of this we should mention a further group of secondary effects, which can be important in some circumstances.

Diffraction effects

For filters of small aperture, especially if they are length-weighted, the effects of diffraction must be considered and, if necessary, compensated for. Work is still under way on this problem: we have theoretical predictions but as yet no experimental results. The kind of effects involved can be seen from *fig. 13* which shows the calculated surface-wave diffraction pattern produced by a 20-electrode transducer of 100 wavelengths aperture. This is the "near-field" region of diffraction, the region in which the receiving transducer of the filter is usually placed. The wave has not yet begun to spread out sideways but nevertheless the diffraction field is very complicated and varies rapidly from point to point. Clearly the amplitude of surface wave detected by a receiving electrode in this near-field region will depend in a complex way on its length and precise position — and may vary substantially with frequency.

In *fig. 14* we show plots of the amplitude errors introduced by diffraction in the signal as received by a single-element receiver at various distances from a launching transducer containing 14 electrodes, each 10 wavelengths long. It is interesting that over a range of distances the effect of diffraction is actually to increase the signal if the receiving electrode is small. For an electrode of the same length as the launcher aperture

[2] H. Engan, Excitation of elastic surface waves by spatial harmonics of interdigital transducers, IEEE Trans. **ED-16,** 1014-1017, 1969.

[3] W. R. Smith, H. M. Gerard, J. H. Collins, T. M. Reeder and H. J. Shaw, Analysis of interdigital surface wave transducers by use of an equivalent circuit model, IEEE Trans. **MTT-17,** 856-864, 1969.

[4] R. F. Mitchell, W. Willis and M. Redwood, Electrode interactions in acoustic surface-wave transducers, Electronics Letters **5,** 456-457, 1969.

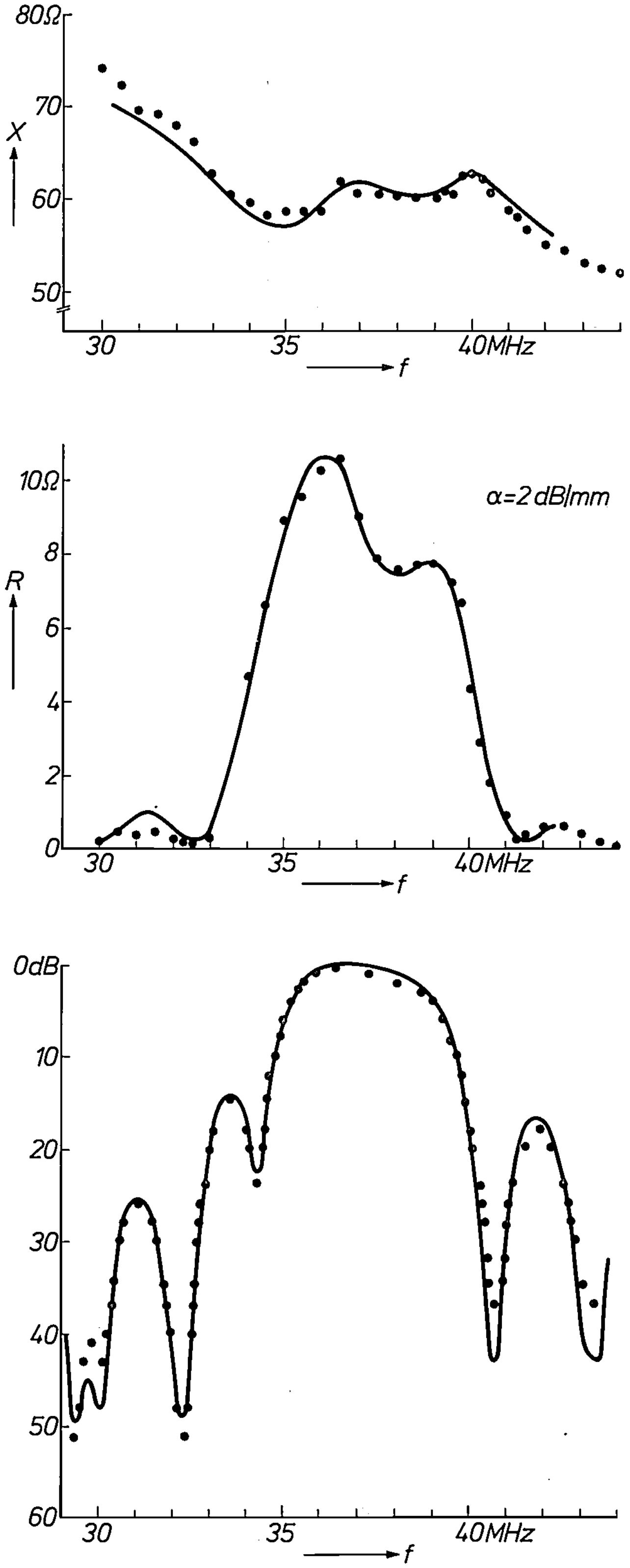


Fig. 12. Comparison of the full theory (from bulk equivalent model; solid lines) with experimental results (points) for a width-weighted acoustic-wave filter. *a*) Variation of reactance X with frequency f. *b*) Variation of resistance R with frequency. α attenuation. *c*) Frequency response of filter. Agreement is better than with the simple model of figs. 4 and 5. The tilt at the centre is found to be due to interaction effects, the depths of the minima are governed by attenuation and the side-lobe levels have been altered by the frequency response of the individual electrodes.

the errors are comparatively small; this is the reason why the width-weighting approach is advantageous. Diffraction errors naturally vary with frequency as well as distance, becoming particularly important at minima where the exact cancellation of all the individual contributions is required. At MRL we have a computer program to calculate the amplitude and phase errors caused by diffraction for each electrode at the centre frequency and then alter the length and position of the electrodes to compensate for the errors. Computation indicates that this compensation carried out for one frequency remains reasonably valid over a considerable bandwidth.

The calculations we have done refer to isotropic surfaces but it has been shown by others [5] that the results can usually be applied to anisotropic materials by a simple scaling of the propagation path length: the intensity profile found a distance x wavelengths from the launcher in the isotropic case is found instead at a distance ax in an anisotropic case where a can be greater or less than one, according to the anisotropy. We have verified this experimentally.

The calculation of diffraction effects at points a few wavelengths away from a radiator is not a problem commonly dealt with in treatises on optical diffraction, for obvious reasons. A formulation of the problem in terms of Green's functions leads to the following form for diffraction of surface waves on an isotropic surface:

$$\psi = -\frac{\mathrm{j}kc}{2\pi\sqrt{r}}\left\{H_1^{(2)}(kr) + \frac{1}{2kr}H_0^{(2)}(kr)\right\}\mathrm{d}S,$$

where ψ is the amplitude reaching a point a distance r away from a small line source $\mathrm{d}S$; $k = 2\pi/\lambda$ and $H_0^{(2)}$, $H_1^{(2)}$ are Hankel functions and c is a constant. Provided r is more than about a wavelength this reduces to

$$\psi = \frac{c'\,k}{\sqrt{kr}}\exp \mathrm{j}\,(\omega t - kr + \pi/4).$$

Thus the signal received from a launcher of length $2l_1$ by a receiver of length $2l_2$, d wavelengths away is

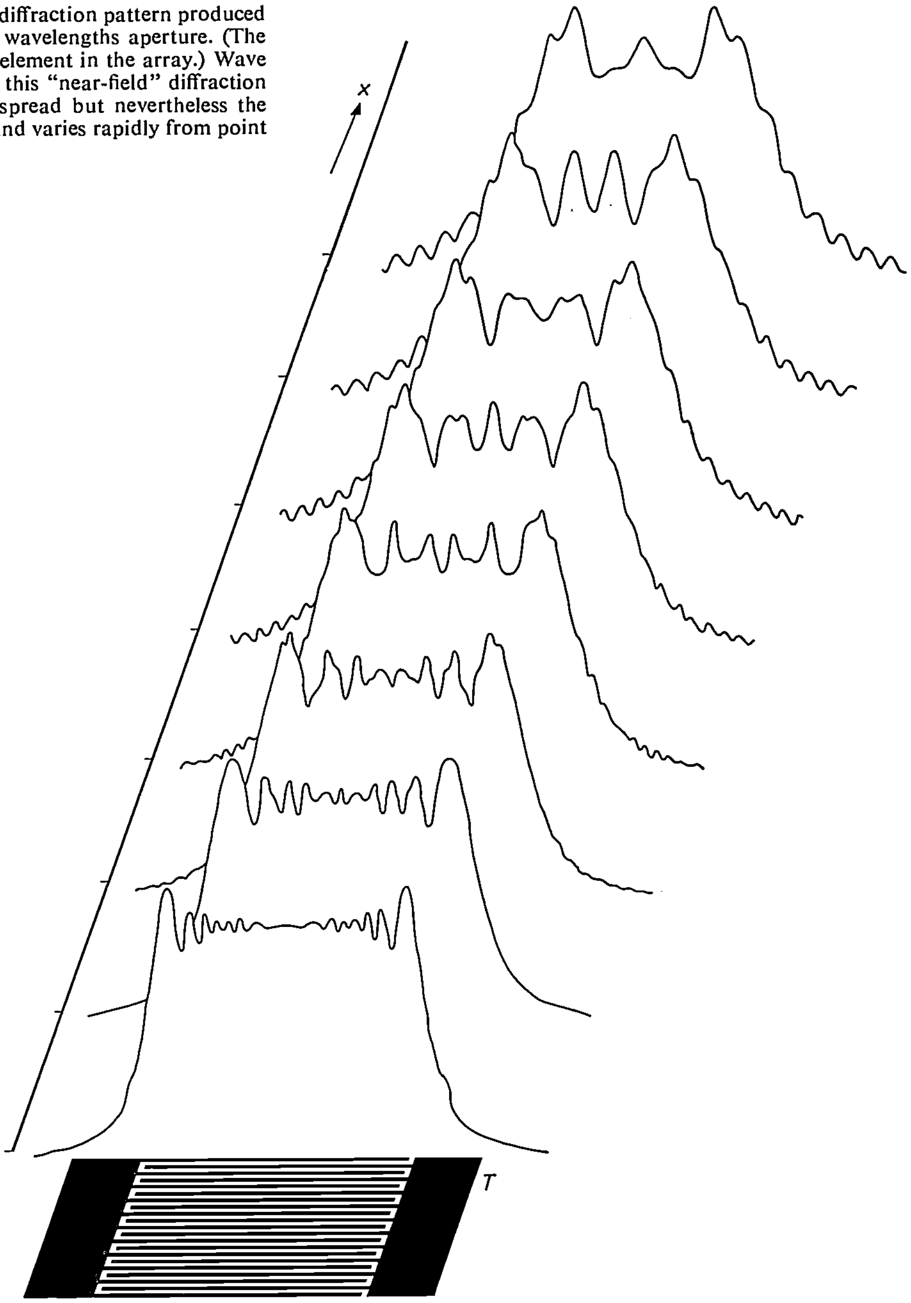
$$c'\,k \int_{-l_1}^{l_1}\int_{-l_2}^{l_2} \frac{d^2 \exp\{\mathrm{j}\,(\omega t - kr + \pi/4)\}}{r^2\sqrt{kr}}\,\mathrm{d}x\,\mathrm{d}x',$$

where $r^2 = d^2 - (x - x')^2$. Expressions of this kind must be evaluated for every electrode in both launcher and receiver to obtain the effect on the full filter.

Choice of substrate material

In designing a surface-wave filter it is very important to choose a substrate material in which the coupling between the acoustic surface wave and the electric field produced by the transducers is as high as possible. Both direct calculation and experimental measurement are far too time-consuming as one wishes to examine a large number of possible propagation directions in each of several materials. Fortunately a convenient approximate method has been suggested [6] [7] which we

Fig. 13. The calculated surface-wave diffraction pattern produced by a 20-electrode transducer at 100 wavelengths aperture. (The aperture is the length of the longest element in the array.) Wave propagation is in the x direction. In this "near-field" diffraction region there is as yet no sideways spread but nevertheless the diffraction field is very complicated and varies rapidly from point to point.



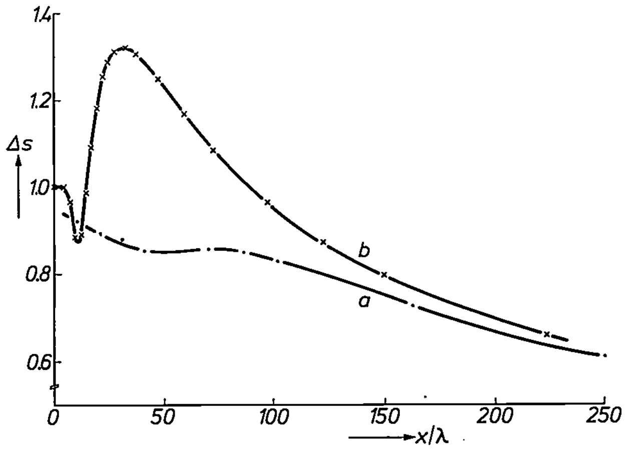


Fig. 14. Curves of the amplitude errors Δs introduced by diffraction in the signal as received by a single-element receiver at various distances from the launching transducer. This has 14 elements and 10λ aperture. Curve a relates to a receiver aperture of 10λ, curve b to a receiver aperture of λ. x direction of propagation, λ wavelength. It can be seen that with a small receiving electrode the signal is actually increased over a certain distance by diffraction. If the receiver and launcher apertures are the same the errors are comparatively small and hence width weighting is advantageous.

[5] I. M. Mason, Two-dimensional surface-wave diffraction from an arbitrary source in an anisotropic medium, Electronics Letters **7**, 344-345, 1971 (No. 12).

[6] J. J. Campbell and W. R. Jones, A method for estimating optimal crystal cuts and propagation directions for excitation of piezoelectric surface waves, IEEE Trans. **SU-15**, 209-217, 1968.

[7] K. A. Ingebrigtsen, Surface waves in piezoelectrics, J. appl. Phys. **40**, 2681-2686, 1969.

have found very useful. The idea is that a surface wave in a piezoelectric material is accompanied by an electric field whose horizontal component is short-circuited if the surface of the material is covered by a thin conducting layer. The result is a change in velocity of the wave, and this change forms a useful measure of the strength of the interaction between electric and acoustic fields in the surface region. The computation is not trivial, but it forms a straightforward extension of a computer program already existing at MRL [8] for calculating the propagation characteristics of surface waves on piezoelectric materials of arbitrary anisotropy. Some typical results of our calculations are shown in *fig. 15.*

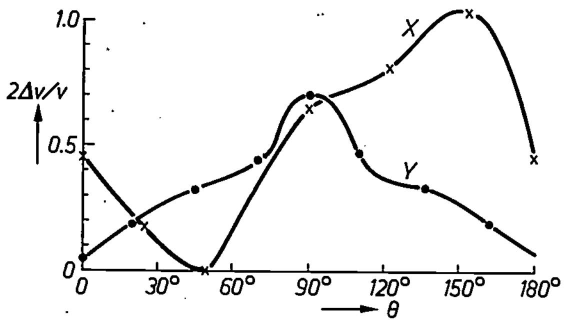


Fig. 15. Directional variation of the coupling (calculated) between electric field and the acoustic surface wave for a lithium-tantalate substrate. A widely used measure of the coupling is $2\Delta v/v$, where v is the velocity of the wave and Δv is the change in velocity on short-circuiting the horizontal electric field by a conducting layer. The curve X relates to an X-cut substrate, and for this curve θ must be taken as the angle between the direction of propagation and the X crystal direction. The curve Y relates to a Y-cut substrate; here θ is the angle between the direction of propagation and the Y crystal direction.

Many other factors must be considered before a final choice of substrate material and propagation direction is made. We list a few below:

1. Dielectric constant. This determines the impedance of the device and so, particularly for large electrode structures, it is important to have a low dielectric constant to maintain a reasonably high impedance.
2. Surface condition must be adequate for photolithography.
3. Velocity must be reproducible and must not vary unduly either with time or with changes in temperature. Variations in velocity will also change the centre frequency of the filter by the same ratio. For a television i.f. filter a temperature coefficient of about 100 ppm/°C is adequate but for more narrow-band filters the requirements are more stringent. Naturally, temperature coefficient usually varies with propagation direction, and this is another property for which optimum directions are most conveniently obtained by computation.
4. In anisotropic materials the direction of propagation is not necessarily perpendicular to the wavefront, and so the surface wave is not launched perpendicular to the electrodes of the launching transducer. This can be an embarrassment in filter design and it is usually convenient to choose a direction in which propagation and wave normal are at least approximately colinear. Once again, this is most conveniently done by computation.

Television i.f. filter

We shall now describe, briefly, the results of applying these techniques to the television i.f. selectivity problem. In many ways this is an ideal application for surface waves. The requirement is for a comparatively wideband filter and the centre frequency is such that the electrode structures can be readily made by standard photolithographic techniques. The frequency response is sufficiently complicated to make full use of the weighting possibilities, and there is much to be gained from producing a filter which can be conveniently encapsulated.

The object of the research programme is to investigate the possibility of producing a filter that is cheap and can be encapsulated in an integrated-circuit package together with its preamplifier and possibly much of the other i.f. circuitry. Such a package has the additional advantage that it needs no adjustment either at the factory or thereafter. It should therefore be an attractive component provided the performance is adequate and the price low. Meeting the commercial requirement of low cost particularly in the encapsulation raises a range of problems which must all be solved before a useful device is made. In many ways these problems are as challenging as the actual synthesis of the frequency response.

The ideal substrate materials are single-crystal piezoelectrics, because of their great uniformity, reproducability and comparative freedom from ageing effects. Price considerations, however, weigh the scales heavily, though not overwhelmingly on the side of artificial piezoelectric ceramics, which are the cheapest materials with the requisite piezoelectric properties. Unfortunately most piezoelectric ceramics are slightly porous, which makes it difficult to define the electrode pattern adequately. They also tend to be non-reproducible. Work at Philips Forschungslaboratorium Aachen, in the Philips Electronic Components and Materials Division and at Philips Research Laboratories has produced a

[8] R. G. Pratt, G. Simpson and W. A. Crossley, Acoustic-surface-wave properties of $Bi_{12}GeO_{20}$, Electronics Letters **8**, 127-128, 1972 (No. 5).

solution to the porosity problem and we now have a ceramic material which appears adequate in most respects. However, ceramic materials have high dielectric constants and so to keep the impedance of the device at a reasonable level we have had to use acoustic apertures as small as 10 wavelengths. Even so the capacitance of the filter is sufficiently low for the special integrated-circuit preamplifier, and to achieve the required level (15 pF) two small devices must be connected in series.

Spurious signals

Many kinds of spurious signals are found in a surface-wave filter: they include reflections of surface waves from the ends of the device, resonant vibrations of the entire substrate in non-surface-wave modes and signals due to bulk acoustic waves which are also launched from the transducer structures. All these spurious signals must be accounted for and reduced to a level where they do not interfere with the performance.

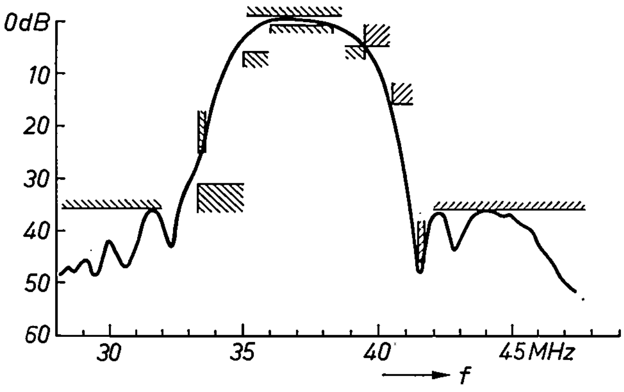


Fig. 16. Frequency response of surface-wave television i.f. filter compared with specification. The filter is mounted in an integrated-circuit package with its own preamplifier. The hatched areas mark the constraints laid down by the specification. These constraints include an allowance for the selectivity of the i.f. coil normally present in a television tuner.

Direct electrical feedthrough is a particular problem. The filter has an insertion loss of about 23 dB and the deepest trap level required is 52 dB. The electrical feedthrough across the device must therefore be more than 75 dB below the input signal. This is not easy to achieve in so small a device.

All these problems are accentuated by the need to keep the substrate as small as possible for low cost and easy encapsulation. Nevertheless we have already produced filters which closely approximate the required performance, although certain of the bulk-wave spurious signals still give ground for concern. These filters are mounted in an integrated-circuit package with a specially-designed preamplifier to make a complete filter unit. The frequency response of such a filter is shown in *fig. 16*, together with the specification points which, it should be noted, include an allowance for the selectivity of the i.f. coil normally present in a television tuner. The insertion loss of the filter is 23 dB, its capacitance 15 pF and size 8×3×0.5 mm. This particular example has been made on potassium sodium niobate, a ceramic which has too high a temperature coefficient of velocity. The new ceramic has the required temperature coefficient and we are currently making filters on this material and also on two single-crystal materials which may prove suitable alternatives.

The author would like to stress that a good deal of the work described in this article was done by others in the team with which he is associated at MRL. In this sense he is also the representative of: M. Chapman, Eileen Hughes, K. W. Moulding, J. S. Palfreeman, D. W. Parker, R. G. Pratt, G. Simpson, J. S. Singleton and R. Stevens.

Optical polarization effects in a gas laser

H. de Lang, D. Polder and W. van Haeringen

Introduction

Any book on classical optics has an important chapter on the characteristics of polarized light and its propagation in physical media. Experiments with polarized light have also greatly contributed to the understanding of the structure of atoms, molecules and matter. The advent of the laser has brought about a revival of optics, mainly because it made available light sources of extremely high intensities and very high temporal and spatial coherence. This article deals with the interaction of laser light with the medium in which it is generated and in particular with the optical polarization effects produced by this interaction.

The first of the following three sections contains a general exposition of the main concepts of laser physics useful for the understanding of the rest of the article. The next section deals with the information that can be obtained from the experimental study of polarization effects in a gas laser. It reveals unorthodox optical behaviour of the generating medium and the section contains a phenomenological description in classical terms. The last section gives the basic elements of a quantitative description. It indicates how a quantum-mechanical theory leads to a detailed understanding of the observations and to some unique predictions which are confirmed by experiment.

General concepts

Lasers and masers are devices in which use is made of the amplification of electromagnetic waves that can be obtained when these waves pass through appropriately prepared gaseous, liquid or solid media. Decrease in the amplitude of an electromagnetic wave propagating through a medium is a well known phenomenon: the absorption coefficient α of the medium is defined as the relative decrease per unit length $-I^{-1}\mathrm{d}I/\mathrm{d}x$ of thc intensity I of the wave. An increase in amplitude and thus amplification can be described by assigning a negative value to the absorption coefficient of the medium. It is possible to create media that have this unusual property and the negative value of α is then caused by the phenomenon of stimulated emission of radiation, which will be discussed later. The mechanism which makes the medium exhibit negative absorption is called the pumping mechanism. The energy required to amplify a wave must be transferred from the medium to the wave and therefore an external energy pump must be present to supply the energy to the medium.

Contrary to what the words LASER (Light Amplification by Stimulated Emission of Radiation) or MASER (Microwave Amplification by S E R) might suggest, any kind of electromagnetic wave can be amplified in principle in appropriately designed devices by stimulated emission. In practice amplification has been realized at many frequencies between those of low-frequency microwaves and short-wave ultra-violet light. The different media used for these purposes must have very specific properties and, generally speaking, a particular medium and its appropriate pumping mechanism will only provide amplification of electromagnetic waves in a small frequency range. This available spectral range differs greatly from case to case.

Once the possibility of amplification exists, the medium can also be used to generate the corresponding electromagnetic waves. Ever present noise can grow to a high intensity through amplification and here the principle of feedback can be used with advantage when the growing wave is reintroduced into the medium several times. The radiation generated in this way has a frequency somewhere in the available spectral range and can be extremely monochromatic, with a spectral purity many orders of magnitude better than that obtained in any other known way.

In the optical region, ranging from the far infra-red to ultra-violet, various gases pumped by a gas discharge offer a large variety of possible laser frequencies. The available spectral range around each frequency is usually very small. For instance, in the helium-neon laser which shows laser action at various frequencies in the red and infra-red the available range is of the order of 500 MHz at centre frequencies of about 3×10^{14} Hz. There are many solids that show laser action in the optical region when pumped by a high-intensity pulsed light source. Examples are the ruby crystal or glass in which ions such as neodymium are present. Lasers are mostly used to generate light of high brilliance in a very small linewidth. Here gas lasers are unique sources

Prof. D. Polder is a Scientific Adviser with Philips Research Laboratories, Eindhoven; Dr. H. de Lang and Dr. W. van Haeringen are also with Philips Research Laboratories.

of light with extreme temporal and spatial coherence, a characteristic that is made use of in holography. Solid-state lasers can generate light of extremely high power, especially if the energy is released in pulses as short as 10^{-8} to 10^{-12} seconds. During a short pulse powers as high as 10^9 watts may be reached. Completely new effects can be observed if matter is irradiated with such a high power.

We shall not try to give a summary of all the different laser sources here. Instead we shall now turn to a discussion of the phenomenon of stimulated emission.

Stimulated emission

In quantum theory, absorption of electromagnetic waves is accounted for by the transition of an atom or molecule from a state (a) of lower energy into one of higher energy (b) under the influence of incident light. If the difference in energy of the two states is ΔE, only light with a frequency $\nu_{\mathrm{ba}} = \Delta E/h$ can be absorbed, where h is Planck's constant. The probability that an atom initially in the lower state (a) jumps to (b) is proportional to the intensity I of the light. The probability that an atom initially in state (b) jumps down to state (a) under influence of the light is also proportional to I and, other things being equal, it is in fact equal to the earlier probability. In the case of a downward transition under influence of incident light the released energy is added to that of the incident light. This is the stimulated emission of radiation. If a unit volume contains N_{a} atoms in state (a) and N_{b} in (b), the decrease of intensity, $-I^{-1}\mathrm{d}I/\mathrm{d}x$, will be proportional to $N_{\mathrm{a}} - N_{\mathrm{b}}$, i.e. amplification can only occur if $N_{\mathrm{b}} > N_{\mathrm{a}}$.

This can be more precisely formulated. In an element of volume the electric field $E = E_0 \cos \omega t$ of the light induces a dielectric polarization P per unit volume, given by

$$P = P_{\mathrm{in}} \cos \omega t + P_{\mathrm{out}} \sin \omega t \qquad (1)$$

with the same angular frequency $\omega = 2\pi\nu$ as that of the light. Both the component P_{in} in phase with E and the out-of-phase component P_{out} are proportional to E_0 as well as to $N_{\mathrm{a}} - N_{\mathrm{b}}$. The part P_{out} that lags by $\pi/2$ is positive if $N_{\mathrm{a}} > N_{\mathrm{b}}$, implying that the medium absorbs energy, while $P_{\mathrm{out}} < 0$ if $N_{\mathrm{a}} < N_{\mathrm{b}}$ in which case the light wave is amplified. Both P_{in} and P_{out} depend on ω. In particular $|P_{\mathrm{out}}|$ as a function of ω is a bell-shaped curve with a maximum at $\omega = \omega_{\mathrm{ba}}$ ($\equiv 2\pi\nu_{\mathrm{ba}}$) corresponding to the fact that for various reasons an atomic absorption line has a finite spectral width. It is precisely over this width that, for a negative absorption ($N_{\mathrm{b}} > N_{\mathrm{a}}$), amplification and laser action is possible. This width was called the available spectral range above. The value and the frequency dependence of P_{in} determine respectively the refractive index and the dispersion of the medium, which, as is well known, can be considerable in the neighbourhood of an atomic transition.

Until now it has been assumed that N_{a} and N_{b} are given quantities. In a medium in thermodynamic equilibrium atomic levels of higher energy are always less populated than the lower ones, i.e. $N_{\mathrm{a}} > N_{\mathrm{b}}$. In a laser or maser a very specific pumping mechanism must be operative to achieve an inverted population $N_{\mathrm{a}} < N_{\mathrm{b}}$ for at least one pair of levels (a) and (b). We shall not discuss pumping mechanisms further here since each different type of laser or maser is pumped in its own particular way. We shall simply assume the presence of a pump that produces Λ atoms in state (b) per unit volume and per unit time. Even in the absence of incident radiation, there are many processes such as atomic collisions which cause the atoms to leave state (b), e.g. for completely different lower-lying states. Such processes obviously hinder the building up of an inverted population. We shall assume the number of atoms disappearing from (b) in this way per unit time to be proportional to the number present in (b), i.e. to be given by $\gamma_{\mathrm{b}}N_{\mathrm{b}}$. Similarly we assume the rate of disappearance from (a) to be equal to $\gamma_{\mathrm{a}}N_{\mathrm{a}}$ caused by transitions to still lower energy levels. The latter transitions favour an inverted population between (a) and (b), but obviously can only be effective if (a) is not the very lowest state of the atom. Indeed in many practical laser systems the lowest laser level (a) is not the lowest state of the atom. For simplicity we assume γ_{a} to be so large that under all circumstances N_{a} remains negligible ($N_{\mathrm{a}} \approx 0$). Finally there are transitions from (b) to (a) which provide the energy required for the amplification of the incident light wave. This stimulated transition rate is given by $c_1(N_{\mathrm{b}}-N_{\mathrm{a}})I$, i.e. is proportional to $N_{\mathrm{b}}-N_{\mathrm{a}}$ and the light intensity. With $N_{\mathrm{a}} \approx 0$ we now have the following rate equation:

$$\mathrm{d}N_{\mathrm{b}}/\mathrm{d}t = \Lambda - \gamma_{\mathrm{b}}N_{\mathrm{b}} - c_1N_{\mathrm{b}}I. \qquad (2)$$

For given I the stationary excess population N_{b}-N_{a} follows:

$$N_{\mathrm{b}}-N_{\mathrm{a}} \approx N_{\mathrm{b}} = \Lambda/(\gamma_{\mathrm{b}} + c_1I). \qquad (3)$$

In all these equations c_1 still depends on the light frequency ν.

From these elementary considerations two conclusions can be drawn. Firstly, the negative absorption coefficient which was proportional to $N_{\mathrm{b}}-N_{\mathrm{a}}$ is by (3) a function of I. Its absolute value decreases with increasing intensity I, an effect that is referred to as saturation of the medium. This means that an electromagnetic wave which is reintroduced again and again into the medium, for example with the aid of mirrors, does not continue to grow indefinitely. The decrease of amplification leads to stabilization of the intensity and the corresponding negative absorption coefficient at a value for which the amplification of the wave just compensates the ever present attenuation due to effects such as absorption in or transmission through the mirrors. We then have a laser operating as an oscillator stabilized by its intensity. Secondly, the dielectric polarization P induced in the medium, which was proportional to E_0 and N_b-N_a, becomes a nonlinear function of E_0, since I in (3) is itself proportional to E_0^2. For small E_0, i.e. $E_0^2 \propto I \ll \gamma_b/c_1$, a series expansion of P in ascending odd powers of E_0 will approximate the nonlinearity. Later in this article the nonlinearity will play an important role in situations where the vector character of the quantities E and P is relevant.

In the foregoing we have not discussed spontaneous emission. The possibility of absorption and stimulated emission by transitions between the states (a) and (b) implies the existence of spontaneous emission of light of frequency ν_{ba} by a spontaneous transition from (b) to (a) whenever an atom is in state (b), irrespective of the presence of incident light. However, the spontaneous transition rate is proportional to N_b and the effect on equation (2) can be accounted for by an adjusted choice of γ_b. Furthermore, spontaneously emitted light does not contribute to amplification of the incident wave. By definition, it is light that is not able to do so, either because of the direction in which it is emitted or because of its incoherence with the incident light.

Spontaneous emission is a source of noise in a stationary oscillating laser. We shall not discuss such effects in this article.

Characteristics of the gas laser

Since the effects that occur in a continuously operating gas laser will be discussed later in this article, the principal features of the device will now be reviewed. Consider a helium-neon laser oscillating at a wavelength $\lambda_l = c/\nu_l$ (≈ 1.15 µm), where c is the velocity of light. It consists of two parallel plane circular mirrors with a diameter d of a few mm at a distance L of about 10 cm with in between them a cylindrical tube containing a helium-neon gas mixture (90% He; 10% Ne; pressure 5 torr, i.e. $\approx 10^{16}$ Ne atoms/cm^3). The mirrors have a large reflection coefficient $R \approx 0.99$. The gas mixture is "pumped" by an electrical gas discharge and as a consequence atomic collisions between He and Ne atoms build up an inverted population between a pair of energy levels (b) and (a) of the neon atoms, their energy difference being $\Delta E = h\nu_{ba} \approx hc/\lambda_l$. Characteristic values for $N_b - N_a$ are 10^9/cm^3, i.e. only a very small fraction of the total number of atoms. The frequency dependence of the resulting unsaturated negative absorption coefficient is given by

$$\alpha(\omega) = \alpha(\omega_{ba}) \exp\left[-(\omega-\omega_{ba})^2/\gamma_D^2\right], \tag{4}$$

where $\alpha(\omega_{ba}) < 0$ and $|\alpha(\omega_{ba})| \approx 10^{-3}$ cm^{-1}. The spectral width of α, for which $\gamma_D \approx 10^9$ s^{-1} is a measure, is caused by the Doppler effect, which will be discussed later.

The fractional gain in intensity of a light wave with angular frequency ω_l travelling back and forth between the mirrors equals $|\alpha(\omega_l)| \times L$ for one transit over a distance L and amounts to about 1%. On the other hand there is a fractional loss $\mathscr{L}$ per round trip caused by transmission and absorption at the mirrors and by a very small part of the light beam spilling over the edges of the mirror. As long as $2L \times |\alpha(\omega_l)| > \mathscr{L}$ the beam undergoes net amplification. In the stationary oscillating situation saturation reduces the negative absorption coefficient at $\omega = \omega_l$ to such a value α_s that its value satisfies

$$2\,|\alpha_s(\omega_l)| = \mathscr{L}/L. \tag{5}$$

The ratio of the unsaturated to the saturated value is

$$F = \alpha(\omega_l)/\alpha_s(\omega_l), \tag{6}$$

and $F-1$ is called the excess fraction above threshold. The greater $F-1$, the greater the intensity I reached by the standing light wave in the stationary oscillator.

As soon as $2|\alpha(\omega_{ba})| > \mathscr{L}/L$, there exists a range of frequencies around ω_{ba} for which

$$2|\alpha(\omega)| > \mathscr{L}/L. \tag{7}$$

The exact frequency $\omega_l = 2\pi\nu_l$ at which the laser in fact will oscillate is now determined by the requirement that the light wave must constructively interfere with itself after one round trip of length $2L$. Ignoring the (frequency-independent) phase jumps at the reflecting mirrors one must then have:

$$2L = p\lambda_l = pc/\nu_l, \tag{8}$$

where the mode number p is an integer of the order of 2×10^5. Generally speaking there now exists a finite number of frequencies ν_l satisfying both (7) and (8), separated in frequency by intervals of $\Delta\nu_l = c/2L$. For $L \approx 10$ cm, $\Delta\nu_l$ is approximately equal to 1.5×10^9 Hz, i.e. of the same order of magnitude as the spectral (Doppler) width of α. It is therefore possible to operate the laser at one single frequency ν_l and also to tune the value within the Doppler profile by precise adjustment of L.

In the foregoing it was more or less implied that the light wave was a plane wave travelling between the mirrors. This cannot be correct since the beam and the mirrors have a finite diameter d of a few millimetres. It is true that since $\lambda/d \ll 1$ and also $d^2/\lambda L > 1$ the deviations from a plane wave are small, but a much more precise analysis is required to show that the simplification to plane waves is adequate for our purpose. In the analysis the system of two small mirrors, which may be plane or slightly curved, at a distance L is considered as an open resonator for electromagnetic waves. As with the more familiar cavity resonator for microwaves, open resonators have a large number of (now closely spaced) eigenfrequencies each with a certain attenuation. The damping is considerable for empty open resonators. Under pumping conditions the medium de-attenuates one or more of the corresponding modes with frequencies ν_l satisfying (7) and a more refined version of (8).

Alternatively the optical system can be considered as a special type of Fabry-Perot interferometer. Because of the small but finite transmission of the mirrors the empty optical system shows the characteristic transmission maxima as a function of frequency for an external incident light beam parallel to the optical axis. For our purpose it is sufficient to assume that in a laser, oscillating at a frequency ν_l, there exists a well defined standing-wave pattern and that this pattern can be thought of as being composed of two plane waves travelling between the mirrors. The finite transmission of the mirrors allows a narrow and nearly plane-parallel beam to leave the system on which the experimental observations can be made.

Doppler broadening

For a proper understanding of a gas laser a discussion of the Doppler width γ_D in equation (4) is necessary. The spectral width γ_D of the positive or negative $\alpha(\omega)$ seen by a propagating wave does not correspond to the width of the absorption or stimulated-emission line of an individual atom, but finds its origin in the velocity distribution of the atoms. If all atoms in states (a) and (b) were at rest a much narrower line $\alpha_0(\omega)$ centred around ω_{ba} would be obtained:

$$\alpha_0(\omega) = A(N_a - N_b)/[(\omega - \omega_{ba})^2 + \gamma_{ba}^2]\,. \qquad (9)$$

Here the characteristic width γ_{ab} is much smaller than γ_D and also the shape of the line differs from that in equation (4). In (9) $A > 0$ and the notation makes explicit the proportionality of α with $N_a - N_b$.

If all atoms had a velocity v_z, a wave propagating in the $+\,z$-direction would experience absorption centred around a displaced frequency ω_{v_z} in accordance with:

$$\alpha_{v_z}{}^+(\omega) = A(N_a - N_b)/[(\omega - \omega_{v_z})^2 + \gamma_{ab}^2]\,, \qquad (10)$$

with

$$\omega_{v_z} = \omega_{ba}(1 + v_z/c). \qquad (11)$$

This is an immediate consequence of the Doppler effect as seen by an observer moving with the atom. In a gas there is a distribution of atomic velocities v_z, so that if $n_a(v_z)\mathrm{d}v_z$ is the number of atoms per unit volume in state (a) in the velocity interval between v_z and $v_z + \mathrm{d}v_z$ and similarly for atoms in (b), then

$$\alpha^+(\omega) = A\int \mathrm{d}v_z\, [n_a(v_z) - n_b(v_z)]/[(\omega - \omega_{v_z})^2 + \gamma_{ab}^2]. \qquad (12)$$

The particular frequency dependence of the unsaturated α expressed by (4) now follows from (12) if, in the absence of saturation, the velocity distribution functions $n_a(v_z)$ and $n_b(v_z)$ are Maxwellian distributions for atoms of mass M and temperature T. In this case γ_D^2 can be identified by $2kT\omega_{ba}^2/Mc^2$ where k is Boltzmann's constant.

The Doppler profile $\alpha(\omega)$ now appears as a weighted superposition of many narrow lines of width γ_{ab}. This width is called the homogeneous linewidth, since in contradistinction to the inhomogeneous Doppler width it cannot be further resolved into contributions from different classes of atoms. One cause of the homogeneous width is the finite lifetime of the levels, because of collision and spontaneous emission. In the helium-neon laser $\gamma_{ab} \approx 2.10^7$ s^{-1} and is indeed smaller than γ_D.

The dielectric polarization P induced in a gaseous medium by the electric field E of a wave will also consist of the sum of the contributions from atoms in the various velocity intervals. This now greatly complicates the evaluation of saturation effects such as saturation of negative absorption or, more generally, the evaluation of the nonlinear relation between P and E. For this a detailed knowledge is required of how an intense beam at frequency ν_l affects the values of $n_a(v_z)$ and $n_b(v_z)$ for each velocity v_z. The point here is that the beam travelling in the positive z-direction reduces the inverted population $n_b - n_a$ of only those atoms for which ω_{v_z} in equation (11) approximately equals ω_l, i.e.

$$|\,\omega_{v_z} - \omega_l\,| \leqslant \gamma_{ab}. \qquad (13)$$

From (11) this means that only atoms for which

$$v_z \approx c(\omega_l - \omega_{ba})/\omega_{ba} \qquad (14)$$

are affected by saturation. Similarly the beam returning in the negative z-direction only affects atoms with

$$v_z \approx -\, c(\omega_l - \omega_{ba})/\omega_{ba}. \qquad (15)$$

Therefore, in a stationary oscillating laser two holes are burned in the inverted distribution $n_b(v_z) - n_a(v_z)$.

A quantitative description can be obtained by setting up rate equations of the type of equations (2) and (3) for atoms in each velocity interval separately. As a consequence of hole burning $|\alpha(\omega_l)|$ is reduced to its saturated value $|\alpha_s(\omega_l)|$; in addition both holes influence the refractive index experienced by each of the two waves. Very special effects occur when $\omega_l \approx \omega_{ba}$ as in this case atoms with $v_z = 0$ are saturated by the two waves simultaneously, so that the laser intensity reached will be lower than it would otherwise have been. The phenomenon is known as the Lamb dip in the laser output at $\omega_l \approx \omega_{ba}$.

In what follows the general concepts exposed so far will be made use of, but the attention will be focused on questions related to the vector character of the electric field in a light wave. A stationary oscillating laser normally produces light of a definite state of optical polarization. If the laser system has ideal cylindrical symmetry about its z-axis one does not expect the laser output to be optically polarized, in, say, the x-direction. Polarization in the y- or any other direction would be equally likely or for that matter any superposition of linear polarizations, e.g. elliptical polarization. It appears, both experimentally and theoretically, that these considerations are not at all complete and that the experimentally observed polarization phenomena are not exclusively due to deviations from ideal cylindrical symmetry but are also related to nonlinear properties of the medium under the influence of polarized light.

Optical polarization states in laser interferometers

As argued above there would be no constraint on the state of polarization of a mode in an isotropic, i.e. cylindrically symmetric, laser. In practice such an ideal situation does not exist. Even if the interferometer only consists of two parallel plane mirrors there will be a small anisotropy, caused for instance by imperfections in the reflecting films on the mirrors. Indeed, in an article [1] on the first helium-neon laser, which was of the Fabry-Perot type, the radiation was reported to be plane polarized in a direction apparently controlled by a linear pattern of fine cracks in the films. In the literature at that time the direction of polarization was generally assumed to be the direction for which the interferometer shows minimum loss. Such an interpretation in terms of loss anisotropy would imply independence of the direction of polarization from tuning conditions.

Experiments at Philips Research Laboratories with a small Fabry-Perot helium-neon laser ($\lambda = 1.15$ μm) [2] have shown however that the polarization direction more or less abruptly changed by $\pi/2$ when the laser was tuned through the centre of the Doppler profile [3] [4]. This "polarization flip", also reported by other workers, has forced us to discard the interpretation in terms of loss anisotropy. We have replaced it by one involving phase anisotropy of the interferometer, i.e. a different optical path length for different polarization states, and, in addition, anisotropy of the gaseous medium induced by saturation effects [3] [4] [5]. In the later treatment it will be shown how experimental evidence has necessitated the introduction of the concept of saturation-induced anisotropy and how it can be understood in terms of the atomic properties of the medium.

In order to see how interferometer anisotropy affects the polarization state of a mode, the basic repetitive properties of the nearly plane waves travelling back and forth between the mirrors will be recalled. In a stationary mode of oscillation the wave must, after one round trip, reproduce a) its phase b) its amplitude and c) its state of polarization. The mode number p being given, condition (a) fixes the frequency of the wave. Condition (b) can only be met in the presence of an amplifying medium. Condition (c) will be reformulated with the aid of the complex amplitudes of the linearly polarized components E_x and E_y, into which a general state of polarization can always be resolved. For instance, a purely imaginary ratio of E_x and E_y characterizes circular polarization, a purely real ratio plane polarization at an angle arc tan (E_y/E_x) to the x-axis.

Now let E_x and E_y be the complex amplitudes before a round trip and E_x', E_y' those after one round trip. Condition (c) then requires:

$$E_y'/E_x' = E_y/E_x. \tag{16}$$

It is clear that relation (16) is independent of conditions (a) and (b). Indeed a phase factor or an amplitude factor common to E_y' and E_x' does not affect (16). Therefore, if the mirrors are the only anisotropic elements in the interferometer, the consequences of (16) can be studied with the empty interferometer. In this case there will be a linear relation between E' and E, so that

$$\begin{aligned} E_x' - E_x &= a_{11}E_x + a_{12}E_y \\ E_y' - E_y &= a_{21}E_x + a_{22}E_y \end{aligned} \tag{17}$$

where the matrix elements a_{ij} are complex quantities. The ratios E_y/E_x that satisfy both (16) and (17) are given by

$$E_y/E_x = \frac{a_{22} - a_{11}}{2a_{21}} \pm \left[\left(\frac{a_{22} - a_{11}}{2a_{21}}\right)^2 + \frac{a_{21}}{a_{12}}\right]^{\frac{1}{2}} \tag{18}$$

and define two states of (generally) elliptical polarization, the "eigenstates of polarization" of the inter-

[1] D. R. Herriott, J. Opt. Soc. Amer. 52, 31, 1962.

[2] H. G. van Bueren, J. Haisma and H. de Lang, Physics Letters 2, 340, 1962. See also J. Haisma, S. J. van Hoppe, H. de Lang and J. van der Wal, Philips tech. Rev. 24, 95, 1962/63.

[3] H. de Lang, Physica 33, 163, 1967 (Proc. Zeeman Centennial Conference, Amsterdam 1965).

[4] H. de Lang and G. Bouwhuis, Physics Letters 19, 481, 1965.

[5] H. de Lang, Thesis, Utrecht 1966 (also published as Philips Res. Repts. Suppl. 1967, No. 8).

[6] H. de Lang, Philips Res. Repts. 19, 429, 1964.

ferometer [6] [5]. They are fixed by the characteristics of the mirrors and are independent of frequency at least to the extent that these characteristics are frequency insensitive. The mirrors are isotropic if $a_{22} - a_{11} = a_{12} = a_{21} = 0$. In this case E_y/E_x in (18) is undefined, corresponding to a previous statement that there is then no constraint on the state of polarization. In practice $a_{22} - a_{11}$, a_{12} and a_{21} are small quantities of the order 10^{-4}. However, it is the ratios of these small quantities which determine the eigenstates in accordance with (18).

Assuming again that in a stationary oscillating laser the role of the medium is to provide enough amplification to make up for the losses, it may be concluded that a laser should oscillate in one or both of the eigenstates of polarization determined by the empty interferometer. The frequencies corresponding to these states may be different as a result of the phase anisotropy of the mirrors. Since polarization-dependent phase jumps at the mirrors are of the order of the coefficients a_{ik}, i.e. 10^{-4}, and the mode number is of the order 2×10^5 a relative frequency shift of the order of $10^{-9}/2\pi$, i.e. about 50 kHz is expected. The empty cavity loss corresponding to the two states of polarization may also be different as a result of loss anisotropy of the mirrors. In view of the fact that the 50 kHz frequency difference is much smaller than the inhomogeneous linewidth γ_{ab}, so that the two eigenstates suffer saturation effects from a common "burned hole", then only the eigenstate of polarization with lowest empty-cavity loss will actually be expected to appear.

Experimental evidence shows however that a laser does not necessarily oscillate in a polarization eigenstate of the empty interferometer and that the medium does not play the neutral rôle so far assumed. This will be discussed later. First of all the Poincaré representation of polarization states will be introduced.

The Poincaré sphere

A very useful way of visualizing polarization states was introduced by Poincaré in 1892. According to Poincaré a state is represented by a point on a sphere, its location being defined by a longitude 2ψ and a latitude 2χ (see *fig. 1*). The correspondence between a state of polarization and a point $(2\psi, 2\chi)$ on the sphere is that ψ corresponds to the angle between the long axis of the polarization ellipse and the x-axis of the laser while $\tan \chi$ is the axis ratio of the ellipse $(-1 \leqslant \tan \chi \leqslant 1; -\pi/4 \leqslant \chi \leqslant \pi/4)$. In particular, the two poles represent circular polarization with opposite sense of rotation, whereas the equator is the locus of plane polarizations. The eigenstates of polarization discussed before can therefore be represented by two points somewhere on the sphere. They are diametrically opposite in the case of mirrors with pure phase anisotropy and pure loss anisotropy.

An equivalent way of expressing the correspondence is with the aid of the complex amplitudes E_x and E_y introduced before. It is not difficult to check the expressions

$$\begin{aligned} \tan 2\psi &= (E_x{}^*E_y + E_y{}^*E_x)/(E_x{}^*E_x - E_y{}^*E_y) \\ \sin 2\chi &= -\,\mathrm{i}(E_x{}^*E_y - E_y{}^*E_x)/(E_x{}^*E_x + E_y{}^*E_y), \end{aligned} \qquad (19)$$

where the starred quantities are the complex conjugates. Note that these expressions are functions of (E_y/E_x) or $(E_y/E_x)^*$ only.

The Poincaré sphere is also convenient for visualizing the time evolution of a polarization state [5]. To illustrate this, consider what happens if at some initial

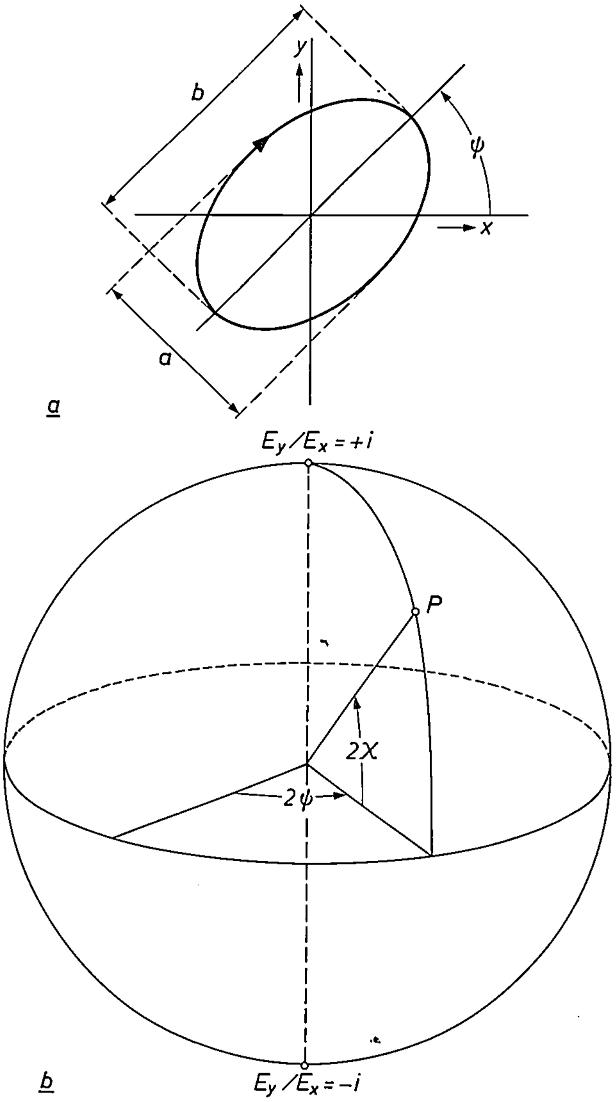


Fig. 1. Polarization ellipse (*a*) and its representation *P* on the Poincaré sphere (*b*); the longitude on the sphere is twice the azimuth ψ of the long axis and the latitude is 2χ where $\tan \chi = a/b$ $(-1 \leqslant \tan \chi \leqslant +1)$. In particular the "equator" $(2\chi = 0)$ is the locus of plane polarizations whereas the "poles" $(2\chi = \pm \pi/2)$ represent opposite circular polarizations.

time the wave inside the laser has a frequency ν, an intensity I and a state of polarization such that the repetitive conditions (a), (b) and (c) of page 194 are not exactly satisfied. It stands to reason that then the change of E_y/E_x per round trip time is $E_y'/E_x' - E_y/E_x$ and follows from equation (17), again under the assumption that only the mirrors affect the polarization state. The temporal change in the state of polarization at a given time then only depends on the state of polarization at that time.

This makes possible the construction of trajectories on the Poincaré sphere. One such trajectory is shown in *fig. 2.* In general it starts from the eigenstate of polarization with highest empty-cavity loss and ends at the state of lowest loss, in agreement with the idea that the latter state will be found in the stationary laser. The length of the arrows in the flow field in fig. 2 indicates the magnitude of the time derivative. It is zero at the unstable point of departure and at the stable final position. In the case of an interferometer with pure loss anisotropy the trajectories are semicircles directly connecting the (opposite) eigenstates; in the case of pure phase anisotropy they are closed circles perpendicular to the axis connecting the two eigenstates. Such a case will be encountered in the next section (fig. 8).

A crucial check on the validity of the considerations given so far is provided by an experiment [3] [7] [5] in which a mirror is used that on reflection converts part of one circular component of the incident light into circularly polarized light of the opposite sense. Such a mirror does not exist, but its function can be realized by making use of the light transmitted through a normal isotropic mirror M_1 in such a way that a fraction is reflected back into the interferometer by means of an additional external mirror M_3. The arrangement also involves a polarizer and a $\lambda/4$ plate between M_1 and M_3 as shown in *fig. 3.*

For such an artificially anisotropic mirror $a_{12} = \mathrm{i}a_{11} = -\mathrm{i}a_{22} = a_{21}$, so that from (18) the two eigenstates of polarization degenerate into a single one at the pole for which $E_y/E_x = \mathrm{i}$. The flow pattern corresponding to this case is given in *fig. 4.* It can be shown that the meridian trajectory crosses the equator at a longitude $2\psi_c = 4\pi z_M/\lambda +$ constant, where z_M is the axial position of mirror M_3. This position determines the arguments of the matrix elements a_{ik}.

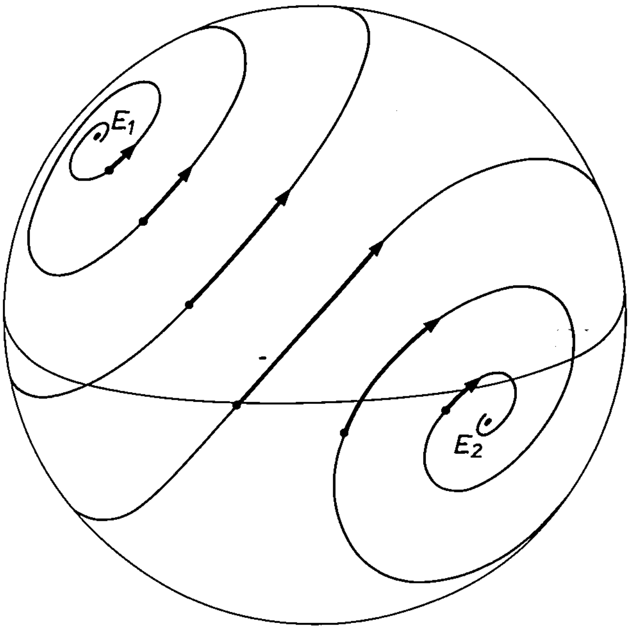


Fig. 2. Flow line on the Poincaré sphere of mode polarization under the influence of a general interferometer anisotropy. The flow line spirals from the high-loss eigenstate E_1 towards the low-loss eigenstate E_2. The sense of rotation of the spiral is determined by the sign of the frequency difference $\nu_2 - \nu_1$ of the eigenstates. In this example $\nu_2 - \nu_1 > 0$.

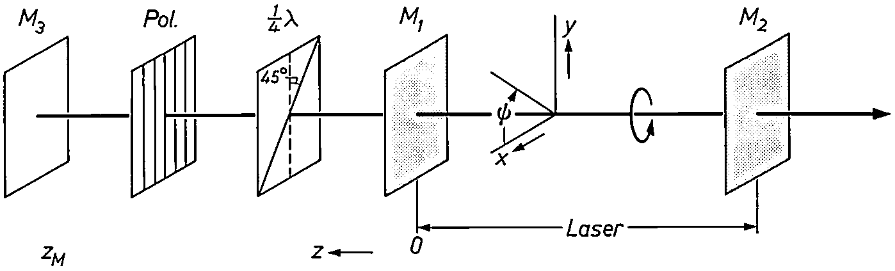


Fig. 3. Fabry-Perot laser with an external anisotropic reflecting system causing the interferometer to have a degenerate anisotropy with a circular eigenstate; mode polarization is nearly linear due to saturation-induced anisotropy of the medium; the azimuth ψ of mode polarization is a measure of the axial position z_M of the external mirror M_3 through the relation $\psi = 2\pi z_M/\lambda +$ constant.

The important point now is that experimentally a laser provided with this reflection arrangement oscillates in a nearly plane-polarized state with azimuth ψ depending on z_M through the relation $\psi = 2\pi z_M/\lambda +$ constant. This experimental fact is interesting from various points of view. First of all it is an example of an arrangement in which the plane of polarization of an intense light source changes periodically as a distance

(z_M) varies. Such an arrangement is of technical importance in that it allows one to construct an apparatus that counts and measures displacements electronically in complete wavelengths and fractions. It is also a surprising fact since, as can be seen in fig. 4, there is no point near the equator where the time derivative of the polarization state might be expected to be zero. It is therefore reasonable to conclude that not only the interferometer but also the gaseous medium itself plays a role in determining the polarization state.

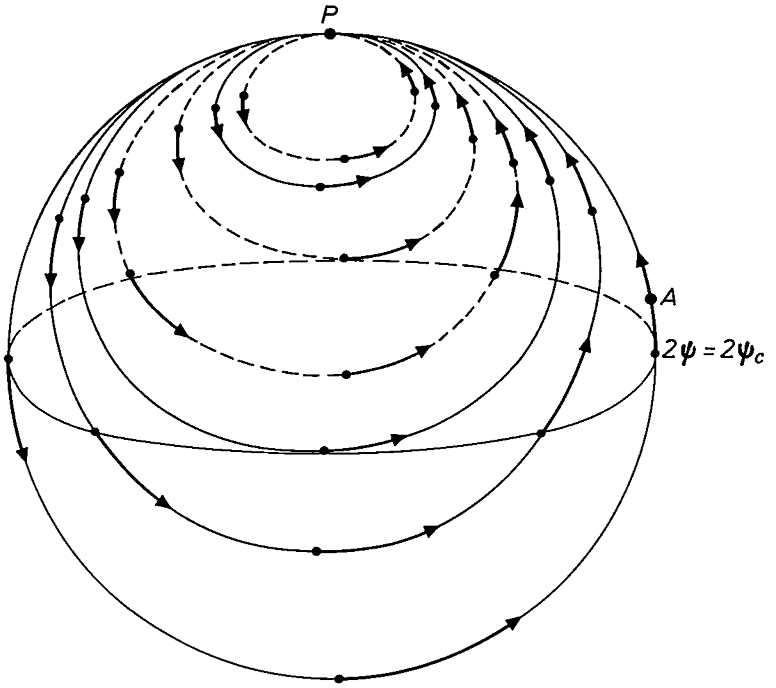


Fig. 4. Flow pattern due to interferometer anisotropy of the arrangement shown in fig. 3. All the flow lines are circles through the circular eigenstate and are tangential to the meridian with azimuth $2\psi_c$.

To show this more clearly we ought to mention an additional observation. The polarization state of the laser is slightly elliptical, i.e. represented by a point slightly off the equator in the direction of the expected circularly polarized eigenstate. The deviation decreases as the excess fraction F— 1 above threshold increases. All this strongly points to an effect due to nonlinear properties of the medium. It is as if a strong additional "force" is present driving the polarization state towards the equator, this force increasing as nonlinear or saturation effects increase. The force would be proportional to the light intensity in the medium. On the Poincaré sphere it would be represented by arrows pointing from the poles to the equator, the arrows being zero at the poles and at the equator (see *fig. 5*). The last two statements follow from the symmetry: the medium might show preference for plane polarization but it should not discriminate either between different values of ψ or between the two possible senses of rotation. Within this hypothesis it can be seen that the laser will find a stationary state of oscillation A in fig. 4 on the meridian trajectory where the arrows due to the mirror anisotropy and the saturation-induced medium anisotropy just compensate each other. A second stationary state diametrically opposite to A is easily seen to be unstable. The z_M-dependence of the longitude 2ψ is then also explained.

A more precise analysis given in the next section confirms that this is essentially a correct interpretation but only part of the story. It will be shown there that saturation-induced anisotropy also gives rise to driving forces with a component parallel to the equator on Poincaré's sphere as depicted by the flow pattern in *fig. 6*. They are caused by nonlinear dispersive effects in the medium. The superposition of the flow patterns of figs. 5 and 6 yields the net flow pattern resulting from saturation effects and is shown in *fig. 7*. In the crucial experiment discussed above the superposition causes a displacement of the stationary point A.

It is of interest to note that the existence of the net saturation-induced flow pattern was postulated to account logically for the polarization-flip effect men-

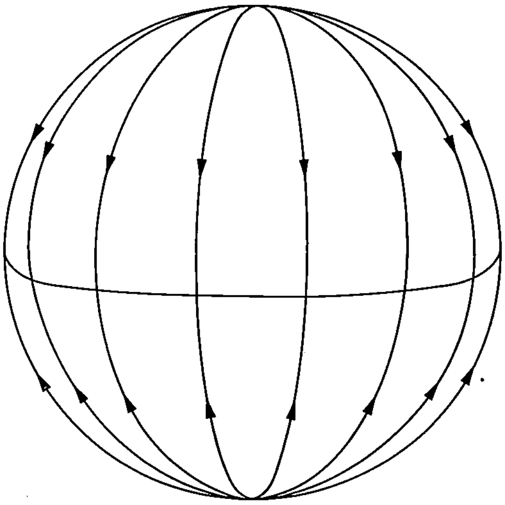


Fig. 5. Pattern of meridional flow lines towards the equator due to saturation-induced (absorptive) anisotropy in a 1.152 μm He-Ne laser.

[7] H. de Lang, G. Bouwhuis and E. T. Ferguson, Physics Letters **19**, 482, 1965.

tioned earlier in this article, before a detailed quantum-mechanical theory of nonlinear polarization effects was available. We shall return to a discussion of the polarization flip in the next section.

Theoretical treatment

In a more quantitative description of polarization effects the electromagnetic field in the laser is thought of as a standing wave sin kz, with angular frequency ω and an amplitude and phase slowly varying in time [8] [9] [10] [11]. In particular for the electric field

$$\boldsymbol{E}(z,t) = \tfrac{1}{2}\,[\boldsymbol{E}(t)\mathrm{e}^{-\mathrm{i}\omega t} + \boldsymbol{E}^*(t)\mathrm{e}^{-\mathrm{i}\omega t}]\sin kz;\quad (L = p\pi/k). \tag{20}$$

Here $\boldsymbol{E}(t)$ is a slowly varying complex variable. The aim is to find the equations of motion for the two-dimensional vector $E_x(t)$, $E_y(t)$. Since the mirrors are discrete elements in the laser, such a description is not immediately applicable. As long as the effect of the mirrors on the cavity waves is only small for each round trip, it can however be spread out over the cavity length so that the purely sinusoidal z-dependence in (20) is guaranteed. The spreading-out procedure consists in formally assigning anisotropic dielectric and lossy properties to vacuum in such a way that the polarization-dependent optical path length and the anisotropic losses caused by the mirrors are accounted for in the mean. The total dielectric polarization $\boldsymbol{P}(z,t)$ at any point is then the sum of a fictitious polarization in vacuum and the actual dielectric polarization of the medium.

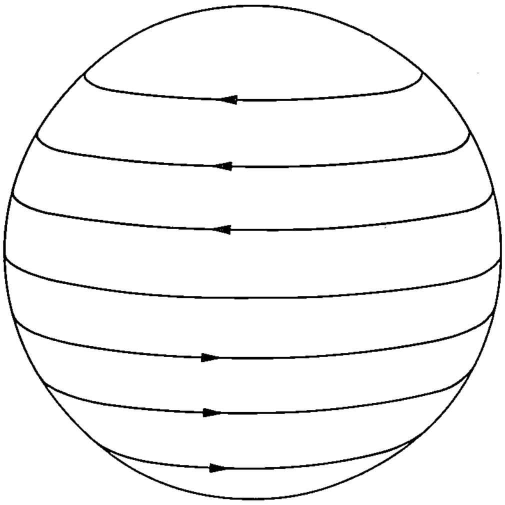


Fig. 6. Pattern of flow lines parallel to the equator due to saturation-induced (dispersive) anisotropy in a 1.152 μm He-Ne laser. The direction of flow is determined by the sign of ($\nu_{\mathrm{mode}} - \nu_{\mathrm{line}}$) which in this example is positive.

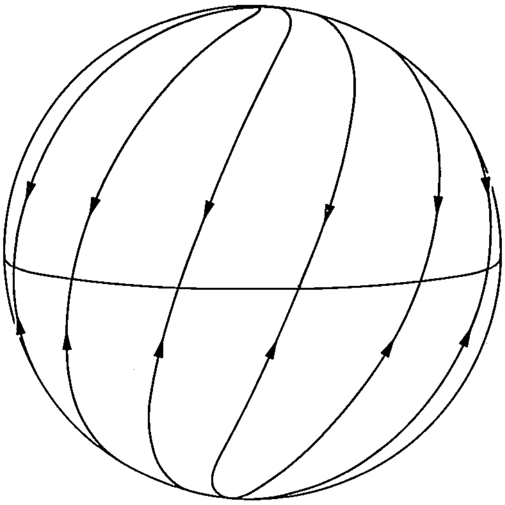


Fig. 7. Flow pattern resulting from the combined effects of the absorptive and the dispersive part of saturation-induced anisotropy (superposition of the flow patterns of figs. 5 and 6).

From Maxwell's equations for nearly plane waves where ε_0 is the permittivity of free space,

$$k^2\boldsymbol{E}(z,t) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\boldsymbol{E}(z,t) + \frac{\partial^2\boldsymbol{P}(z,t)}{\varepsilon_0 c^2 \partial t^2} = 0\,. \tag{21}$$

On applying equation (20) the second derivative of $\boldsymbol{E}(t)\exp\mathrm{i}\omega t$ gives three terms. By choosing $\omega = ck$, $k^2\boldsymbol{E}(t)\exp\mathrm{i}\omega t$ is cancelled, the mixed derivative is retained and the term in $\partial^2\boldsymbol{E}/\partial t^2$ is neglected because of the assumed slow variation of $\boldsymbol{E}(t)$. If for $\boldsymbol{P}(z,t)$ an expression analogous to (20) is used, the vector equation of motion is obtained in what is known as the "adiabatic" approximation:

$$\frac{2\mathrm{i}\omega}{c^2}\frac{\partial\boldsymbol{E}(t)}{\partial t} = \frac{\omega^2\boldsymbol{P}(t)}{\varepsilon_0 c^2}, \tag{22}$$

together with its complex conjugate. In deriving (22) the largest term arising from the second derivative of $\boldsymbol{P}(z,t)$ is retained.

Equation (22) contains only slowly varying quantities. In order to solve it the dependence of $\boldsymbol{P}(t)$ on $\boldsymbol{E}(t)$ must be known. $\boldsymbol{P}(t)$ consists of the sum of three terms.

The first of these terms is the fictitious vacuum polarization $\boldsymbol{P}_v$. It is linearly related with $\boldsymbol{E}$ by means of a two-dimensional tensor $\tilde{A}$:

$$\boldsymbol{P}_v = \varepsilon_0\tilde{A}\boldsymbol{E}. \tag{23}$$

The dielectric permittivity ε_0 of free space has been added to make $\tilde{A}$ dimensionless. The real part of $\tilde{A}$ must be chosen so as to account for the phase anisotropy of the mirrors, the imaginary part for the anisotropic losses [12]. Remembering that equation (17) was the change in E per round-trip time $2L/c$ we find using (22) the following relation between A_{ij} and a_{ij}:

$$a_{ij} = -\ \mathrm{i}\omega L A_{ij}/c. \tag{24}$$

Since a_{ij} was of the order 10^{-4}, A_{ij} is of the order 10^{-10}.

The second term is the dielectric polarization $\boldsymbol{P}_1$ of the medium as far as it is linearly dependent on $\boldsymbol{E}$, i.e. characteristic for the unsaturated medium. In view of the isotropy of the unsaturated medium we have:

$$\boldsymbol{P}_1 = S_1\varepsilon_0\boldsymbol{E}, \tag{25}$$

where S_1 is a complex, strongly frequency-dependent scalar. Its (positive) imaginary part is proportional to the (negative) absorption coefficient α present around the centre ω_{ba} of the Doppler profile:

$$\alpha = -\ (\omega/c)\,\mathrm{Im}(S_1). \tag{26}$$

The third term is the nonlinear dielectric polarization $\boldsymbol{P}_3$ of the medium, occurring in the first term of a series expansion of $\boldsymbol{P}(z,t)$ in odd powers of $\boldsymbol{E}(z,t)$ describing nonlinear effects. Remembering that $\boldsymbol{P}(t)$ is the part of $\boldsymbol{P}(z,t)$ approximately proportional to exp $\mathrm{i}\omega t$ and that an expression involving the third power of $\boldsymbol{E}(z,t)$ produces terms proportional to exp $\mathrm{i}\omega t$ only if two factors $\boldsymbol{E}(t)$ and one factor $\boldsymbol{E}^*(t)$ occur, then

$$\boldsymbol{P}_3 = \varepsilon_0{}^2S_2(\boldsymbol{E}\cdot\boldsymbol{E}^*)\boldsymbol{E} + \varepsilon_0{}^2S_3\,(\boldsymbol{E}\cdot\boldsymbol{E})\boldsymbol{E}^*. \tag{27}$$

The special form of (27) involving two complex scalars S_2 and S_3 is dictated by the requirement that the expression should be invariant for the choice of the coordinate axes x and y, i.e. by the cylindrical symmetry of the laser medium. Note the absence of any quadratic terms in $\boldsymbol{E}$ and $\boldsymbol{E}^*$. This is a consequence of the inversion symmetry of the medium.

The equation of motion now reads:

$$\frac{2\mathrm{i}}{\omega}\frac{\partial\boldsymbol{E}}{\partial t} = \bar{A}\boldsymbol{E} + S_1\boldsymbol{E} + S_2\varepsilon_0(\boldsymbol{E}\cdot\boldsymbol{E}^*)\boldsymbol{E} + S_3\varepsilon_0(\boldsymbol{E}\cdot\boldsymbol{E})\boldsymbol{E}^*. \tag{28}$$

It is instructive to observe that at any time the right-hand-side vector can be resolved into a vector $\boldsymbol{P}_{||}/\varepsilon_0$ parallel to $\boldsymbol{E}$ and a vector $\boldsymbol{P}_{\perp}/\varepsilon_0$ perpendicular to $\boldsymbol{E}$. Only the latter can cause a change in state of polarization. In other words the terms with S_1 and S_2 do not contribute to such a change. The component $\boldsymbol{P}_{||}$ to which all four terms in (28) contribute causes a change in intensity by means of its out-of-phase part $\boldsymbol{P}_{||,\mathrm{out}}$ and a correction to the frequency ω by its in-phase part $\boldsymbol{P}_{||,\mathrm{in}}$.

With the aid of equations (19) and (28) the rate of change of the polarization-state parameters ψ and χ can now be calculated. It is found that:

$$\frac{1}{\omega}\frac{\partial\chi}{\partial t} = \Gamma_1(\psi,\chi) - \mathrm{Im}(S_3)(I/4)\sin 4\chi\,, \tag{29a}$$

$$(\cos 2\chi)\,\frac{1}{\omega}\frac{\partial\psi}{\partial t} = \Gamma_2(\psi,\chi) - \mathrm{Re}(S_3)(I/4)\sin 4\chi\,. \tag{29b}$$

Here Γ_1 and Γ_2 depend on the elements of the matrix $\bar{A}$, and describe cavity anisotropies. $\Gamma_1 = \Gamma_2 = 0$ for isotropic cavities. Moreover I is proportional to the intensity and is defined by

$$I = \varepsilon_0(\boldsymbol{E}\cdot\boldsymbol{E}^*). \tag{30}$$

The time derivative of I itself also follows from (28):

$$\frac{1}{\omega}\frac{1}{I}\frac{\partial I}{\partial t} = \Gamma_3(\psi,\chi) + \mathrm{Im}(S_1) + I\,[\mathrm{Im}(S_2 + S_3) - \\ -\ \mathrm{Im}(S_3)\sin^2 2\chi]\,. \tag{31}$$

Here Γ_3 depends on the lossy part of the matrix $\bar{A}$. It reduces to a negative scalar for isotropic cavities. It equals $c/2\omega L$ times the cavity loss $\mathscr{L}$ per round trip and its order of magnitude is therefore 10^{-8}. In an anisotropic cavity the average empty-cavity attenuation $\Gamma_{\mathrm{av}} = \mathrm{Im}(A_{11} + A_{22})/2 < 0$ always predominates over the (ψ,χ)-dependent part of Γ_3. In a laser Γ_{av} is overcompensated by the positive $\mathrm{Im}(S_1)$ representing the unsaturated deattenuation of the medium. The excess fraction $F - 1$, which in a typical case amounts to about 10% is given by

$$F - 1 = \frac{[\mathrm{Im}(S_1) - |\Gamma_{\mathrm{av}}|]}{|\Gamma_{\mathrm{av}}|}\,, \tag{32}$$

the value of $\mathrm{Im}(S_1)$ to be taken at the appropriate frequency ω_l. Finally, the term in (31) proportional to I is always negative and represents gain reduction by saturation, allowing stationary laser oscillation as discussed in an earlier section. Its order of magnitude is $(F-1)|\,\Gamma_{\mathrm{av}}|$, i.e. 10^{-9}. From this it follows that saturation effects are an order of magnitude larger, in the experiments discussed in this article, than interferometer anisotropies.

The terms proportional to I in (29) represent the "driving forces" on the state of polarization due to saturation-induced anisotropy. In particular the term containing $\mathrm{Im}(S_3)$ defines the flow pattern of fig. 5 and the term containing $\mathrm{Re}(S_3)$ defines the pattern of fig. 6: their superposition is shown in fig. 7.

The stationary state in which a laser will eventually oscillate is determined by those values of I, χ and ψ for

[8] D. Polder and W. van Haeringen, Physics Letters **19**, 380, 1965.

[9] W. van Haeringen, Physics Letters **24A**, 65, 1967.

[10] W. van Haeringen, Phys. Rev. **158**, 256, 1967.

[11] W. van Haeringen and H. de Lang, Phys. Rev. **180**, 624, 1969.

[12] Strictly speaking this is only true for reflection arrangements in which no magnetic materials or magnetic effects are involved. In this case the matrix $\bar{A}$ is symmetric. In the general case, the Hermitean part of $\bar{A}$ describes phase effects and the anti-hermitean part loss effects.

which all time derivatives in (29) and (31) are zero. These equations are quite complex because of their interrelation. In many practical cases, i.e. for not too small values of $F-1$, saturation-induced effects are predominant. In the case of the helium-neon laser at $\lambda = 1.152$ µm this means that χ is small, so that I can be taken to be practically constant. Its value follows from (31) with $\partial I/\partial t = 0$ and $\chi = 0$:

$$I \approx \frac{\mathrm{Im}(S_1) - |\Gamma_{\mathrm{av}}|}{|\mathrm{Im}(S_2 + S_3)|} \approx \frac{(F-1)}{F} \frac{\mathrm{Im}(S_1)}{|\mathrm{Im}(S_2 + S_3)|}. \qquad (33)$$

A constant value of I greatly simplifies the discussion of (29).

All the constants in (28) can in principle be measured. The mirror anisotropies can be calibrated and the values of S_1, S_2 and S_3 can be determined by a careful study of the dependence of the stationary polarization states on the intensity. The quantities S_1, S_2 and S_3 can also be calculated with the aid of a quantum-mechanical treatment of the response of the medium to an electromagnetic field [10]. Such a calculation also gives the frequency dependence: the imaginary (absorptive) parts of S_1, S_2 and S_3 are even functions of $(\omega - \omega_{\mathrm{ba}})$, the real (dispersive) parts are odd functions of $(\omega - \omega_{\mathrm{ba}})$, where ω_{ba} is the centre of the Doppler profile.

Interpretation of some experiments

The theory will now be applied to explain the polarization-flip effect mentioned earlier, on the basis of a small phase anisotropy of the mirrors, due for instance to stress birefringence in the mirror coating, and saturation-induced anisotropy. Without loss of generality we may take the x-axis as the anisotropy axis of the cavity, so that the real matrix $\tilde{A}$ reduces to $A_{11} = - A_{22} = a$, $A_{12} = A_{21} = 0$. In that case one deduces:

$$\Gamma_1 = \tfrac{1}{2}\, a \sin 2\psi,\ \Gamma_2 = - \tfrac{1}{2}\, a \cos 2\psi \sin 2\chi,\ \Gamma_3 = 0. \qquad (34)$$

The corresponding flow pattern is given by the circles in *fig. 8*. The presence in equation (29a) of the term with $\mathrm{Im}(S_3)$, assumed to be positive, drives the polarization state towards the equator and the eventual stationary state will be found to be plane polarized at $\chi = 0$, $2\psi = 0$ or π. In fact for these values $\mathrm{d}\chi/\mathrm{d}t$ and $\partial\psi/\partial t$ are zero. Apparently there are two solutions, the solution $\psi = 0$ being plane polarized in the x-direction, the one with $\psi = \pi/2$ in the y-direction. The corresponding stationary values I_0 of the intensity are identical, in agreement with the absence of loss anisotropy in the cavity. The question is which solution is stable.

In order to investigate this, consider the superposition of the flow pattern from interferometer anisotropy (fig. 8) and that from saturation-induced anisotropy (fig. 7) with I given by (33). The resulting superposition is shown in *fig. 9* for a small region around the stationary point ($\chi = 0$, $\psi = 0$). Looking on to the sphere around the other stationary point ($\chi = 0$, $2\psi = \pi$) *fig. 10* is obtained. The picture with the converging flow lines corresponds to the stable state, the diverging flow lines indicate instability.

Inspection of the figures 9 and 10 shows that the inclination of the saturation-induced flow lines with respect to the meridian (see fig. 7) plays an essential role in the considerations given above. The inclination exists because of the term with $\mathrm{Re}(S_3)$ in equation (29b). A change in sign of this term would cause an inclination in the opposite direction, thereby interchanging the stable and unstable points. This is none other than the polarization flip; since $\mathrm{Re}(S_3)$ is an odd function of $(\omega - \omega_{\mathrm{ba}})$ tuning of the laser from $\omega > \omega_{\mathrm{ba}}$ to $\omega < \omega_{\mathrm{ba}}$ brings about this change of sign and makes the laser flip from one state of plane polarization to the other with an abrupt change in azimuth by $\pi/2$.

There exist interesting effects in the case $\omega = \omega_{\mathrm{ba}}$ where $\mathrm{Re}(S_3) = 0$. A discussion of these effects, bistability and hysteresis, can also be given on the basis of the general theory.

The analysis leading to equation (28) can be extended to the case in which a constant external magnetic field B is applied. Limiting the situation to for instance an axial field, i.e in the z-direction, additional terms are obtained. For small B the main effect is on the relation between $\boldsymbol{P}_1$ and E_1 notwithstanding the simultaneous appearance of B-dependent terms in the nonlinear part of the response of the medium. The main effect is the

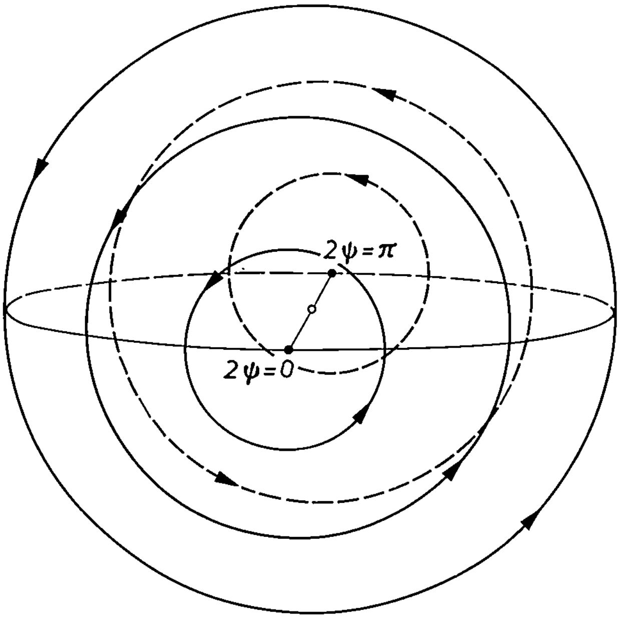


Fig. 8. Flow lines due to phase anisotropy of the interferometer. The eigenstates are plane polarized, the low-frequency one at $\psi = 0$, the high-frequency one at $\psi = \pi/2$.

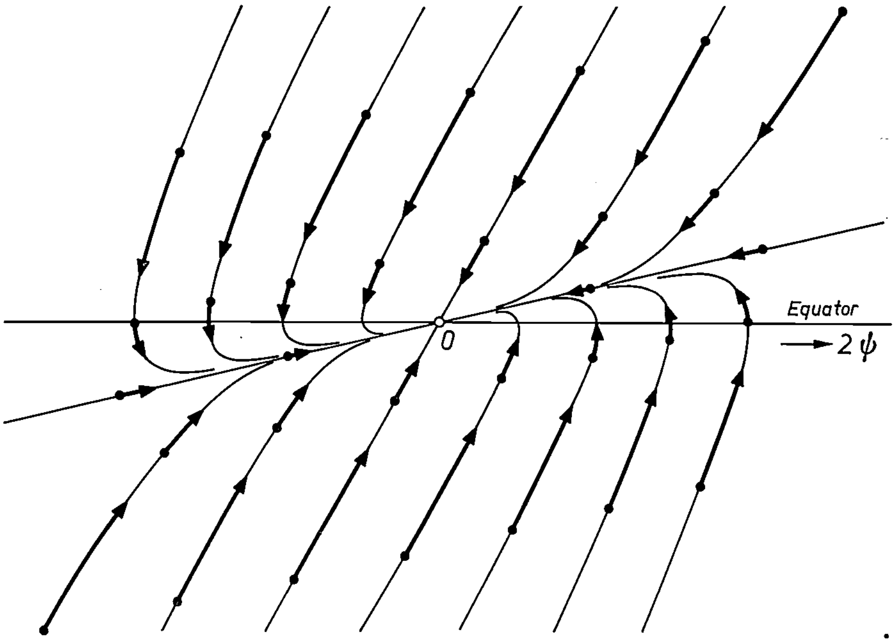


Fig. 9. Superposition of the flow patterns from saturation-induced anisotropy (fig. 7) with that of interferometer anisotropy (fig. 8) in a small region around the point $2\psi = 0, 2\chi = 0$. Converging flow lines confirm the experimentally observed stability of this point. The lengths of the arrows correspond to about 1 μs for a typical case.

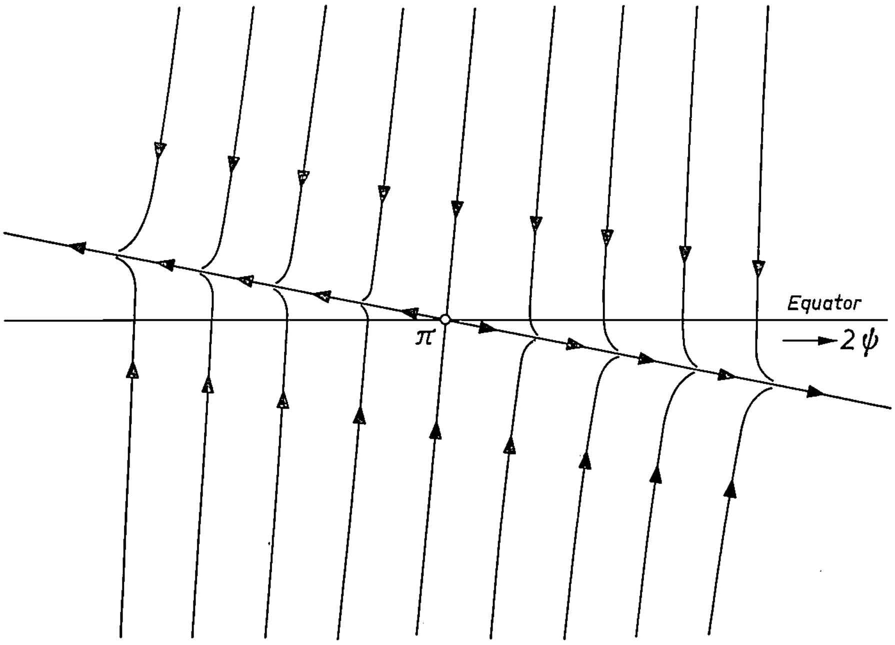


Fig. 10. Superposition of the flow pattern of saturation-induced anisotropy (fig. 7) with that of interferometer anisotropy fig. 8) for a small region around the point $2\psi = \pi, 2\chi = 0$. The presence of diverging flow lines indicates the instability of this point.

Faraday rotation in the medium: the magnetic field causes a difference between the propagation velocities of circularly polarized light of opposite sense of rotation. In addition there may occur a difference in (negative) absorption coefficient, which will be ignored.

In this situation S_1 is no longer a scalar and (25) becomes

$$P_1 = \bar{b}\, E, \tag{35}$$

where $\bar{b}$ is an antisymmetric tensor ($b_{12} = - b_{21}$) with $b_{11} = b_{22}$. In particular b_{12}, for small B, is proportional to B. Its imaginary part β describes the Faraday effect. The flow pattern corresponding to this type of linear-medium anisotropy consists of circles parallel to the equator with a unique sense of rotation (see *fig. 11*). In (29b) an additional term appears:

$$\frac{1}{\omega}\frac{\partial\psi}{\partial t} = -\frac{\beta}{2}. \tag{36}$$

The stationary states of an isotropic laser in a magnetic field are clearly the two poles on the Poincaré sphere, each one corresponding to circular polarization. The presence of the saturation-induced anisotropy given in fig. 7 makes these points unstable. In fact there is no stable state of polarization. The laser polarization state will eventually perform a uniform rotation along the equator. It will show plane polarization, the azimuth being a linear function of time. Seen through a

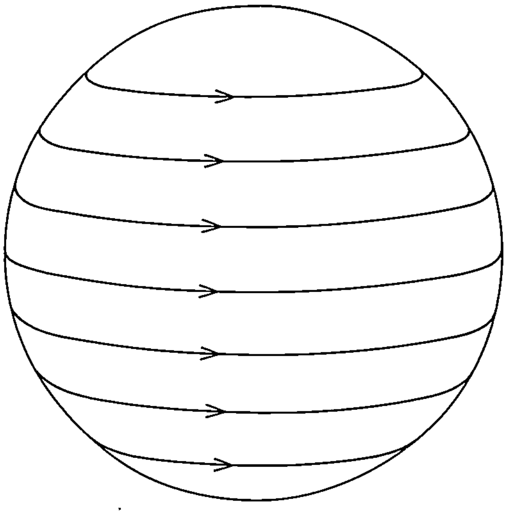


Fig. 11. Flow pattern due to Faraday rotation in the laser medium as caused by a longitudinal magnetic field.

polarizer a periodic output will be observed with a modulation frequency proportional to B which amounts to 50 kHz for $B = 1$ gauss, corresponding to β of the order 10^{-10} for $B = 1$ gauss. In this situation the equatorial flow line in fig. 11 acquires a very real physical significance.

Interesting effects occur when, in addition to a magnetic field, there is a small interferometer-phase anisotropy, such as shown in fig. 8. The equations (29) with the additional term (36), combined with (34) then have an approximate solution determined by

$$\chi \approx [2I\,\mathrm{Im}(S_3)]^{-1}\,a \sin 2\psi \ll 1, \qquad (37a)$$

$$\frac{1}{\omega}\frac{\partial\psi}{\partial t} \approx -\,[2\,\mathrm{Im}(S_3)]^{-1}\,a\,\mathrm{Re}(S_3)\sin 2\psi - \tfrac{1}{2}\beta, \qquad (37b)$$

if saturation anisotropy is dominant, i.e. if both $\mathrm{Im}(S_3)$ and $\mathrm{Re}(S_3)$ are much larger than a/I and β/I. The solution (37a) says that the state of polarization proceeds along an ecliptic passing through the eigenstates of the cavity anisotropy and slightly inclined towards the equator (*fig. 12*). Equation (37b) says that on this ecliptic the dispersive saturation-induced anisotropy $\mathrm{Re}(S_3)$ periodically makes itself felt, alternately increasing and decreasing the "Faraday" velocity $\tfrac{1}{2}\omega\beta$. As a result the motion along the ecliptic is no longer uniform. While the intensity I remains practically constant the output observed through a polarizer will be a non-sinusoidal function of time (*fig. 13*).

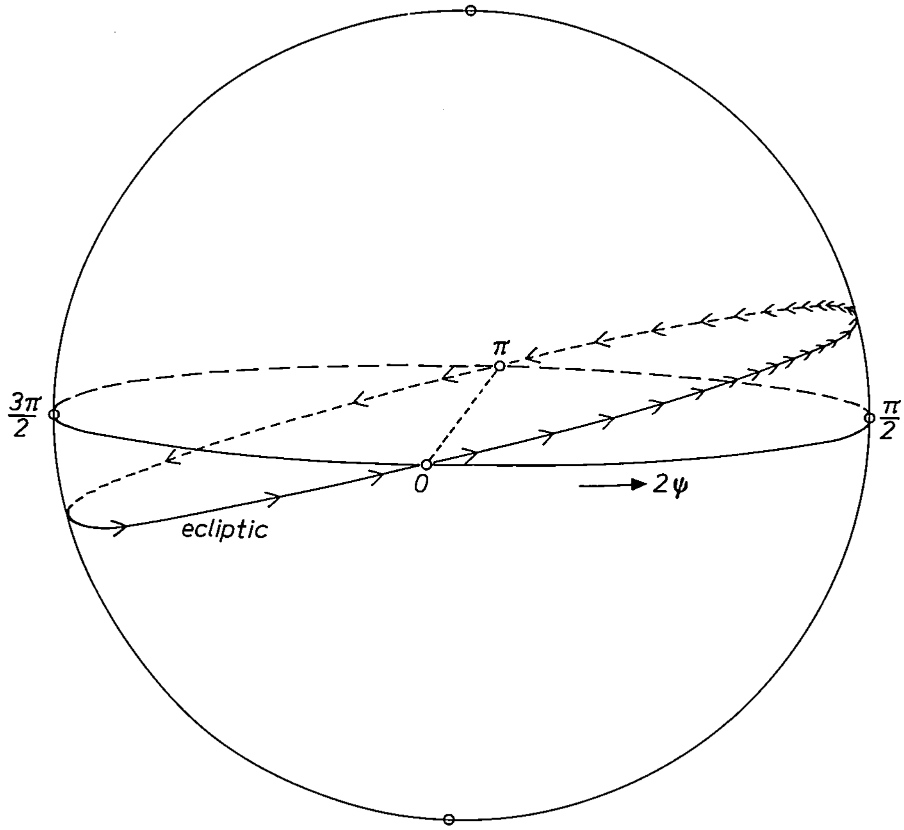


Fig. 12. Periodic motion of mode polarization along an ecliptic. The component flow patterns are those of figs. 7, 8 and 11.

For sufficiently large phase anisotropy a, or sufficiently small β, the continuous motion around the ecliptic will be broken up and a stationary state $\partial\psi/\partial t = 0$ becomes possible. The laser then exhibits nearly plane polarization with an azimuth ψ sensitively depending on the the ratio β/α. All these effects have been observed and interpreted [3] [4] [5] [9] [10] [11]. In actual fact these experiments with a magnetic field were the ones that first made us introduce the concept of saturation-induced anisotropy.

Quantum-mechanical considerations

The physics behind saturation-induced anisotropy lies in the behaviour of the individual atoms, which can only be properly described in quantum-mechanical terms. Nevertheless it is possible to make clear what is actually going on. To do this let us consider an atom of which both the upper state (b) and the lower (a) have angular-momentum quantum number $j = \frac{1}{2}$. This is not the situation in the helium-neon laser discussed previously, but we shall return to this later. Both states are twofold degenerate and the sublevels can be distinguished by the projection of j on the z-axis, i.e. one has $a_{\frac{1}{2}}$ and $a_{-\frac{1}{2}}$ and the other $b_{\frac{1}{2}}$, $b_{-\frac{1}{2}}$. It is a matter of simple symmetry considerations that optical transitions by light propagating in the z-direction are only possible between the pair $a_{\frac{1}{2}} \leftrightarrow b_{-\frac{1}{2}}$ and $a_{-\frac{1}{2}} \leftrightarrow b_{\frac{1}{2}}$, the former only interacting with, say, the clockwise-rotating circular component, the latter pair only with the anticlockwise circular component of the light. The (negative) absorption coefficients of the clockwise and anticlockwise components are therefore determined by the relative overpopulation of $b_{-\frac{1}{2}}$ over $a_{\frac{1}{2}}$ and of $b_{\frac{1}{2}}$ over $a_{-\frac{1}{2}}$ respectively.

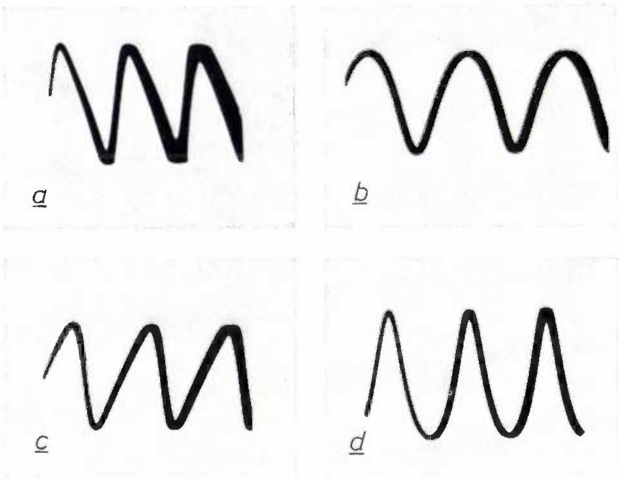


Fig. 13. Oscillograms demonstrating non-uniform rotation of polarization azimuth in fig. 12. The laser beam was sent through a polarizer and the transmitted intensity as a function of time was recorded oscillographically (linear time base of about 3.10^{-4}s). The oscillograms (*a*), (*b*), (*c*) and (*d*) were taken with the transmission azimuth of the analyser at $\psi = 0, \pi/4, \pi/2$ and $3\pi/4$ respectively. The waveform for each of these cases confirms the velocity modulation along the ecliptic as depicted in fig. 12.

Suppose that at some initial time an elliptical state of polarization is present, in which the clockwise circular component has a larger amplitude than the anticlockwise one. Saturation then causes a greater reduction in overpopulation of the $a_{\frac{1}{2}} \leftrightarrow b_{-\frac{1}{2}}$ pair than in the other pair. As a result the anticlockwise component will grow relative to the clockwise component, causing the elliptical polarization to become more plane polarized. This, in essence, is the explanation of the equator-directed saturation-induced driving force on the Poincaré sphere for a $j = \frac{1}{2} \leftrightarrow j = \frac{1}{2}$ transition.

In the laser with $\lambda = 1.152$ μm a transition between a threefold-degenerate level with $j = 1$ and a fivefold-degenerate level with $j = 2$ is operative. We have therefore worked out a detailed theory [8] for calculating the values of S_1, S_2 and S_3 for any possible set of integer or half integer j values (j,j') satisfying the selection rule $|j' - j| = 0$ or 1.

It is not unexpected that the values of S_1, S_2 and S_3 turn out to be proportional to $N_b - N_a$. Unfortunately this is a quantity that depends strongly on experimental conditions and it cannot even be directly measured. The ratios, such as S_2/S_1, S_3/S_1, or $\mathrm{Im}(S_3)/\mathrm{Im}(S_1)$, will not contain this quantity. The imaginary part of S_1, however, can be expressed in terms of measurable quantities with the aid of equation (32). It is equal to $F|\Gamma_{av}|$. Therefore $\mathrm{Im}(S_3)$ can be expressed in terms of F and $|\Gamma_{av}|$. The detailed calculation gives

$$\mathrm{Im}(S_3) = \frac{F|\Gamma_{av}|\, p_{ab}{}^2}{8\, h^2 \varepsilon_0\, \gamma_a \gamma_b} \left[\frac{(\omega - \omega_{ba})^2 + 2\,\gamma_{ab}{}^2}{(\omega - \omega_{ba})^2 + \gamma_{ab}{}^2}\right] g(j), \quad (38)$$

where p_{ab} is the electric-dipole matrix element between the upper and lower states and g(j) is defined by

$$\begin{aligned} g(j) &= -\frac{1}{5}(j-1)(j+2) \quad \text{for a } j \leftrightarrow j \text{ transition} \\ g(j) &= +\frac{2}{5}j(j+2) \quad \text{for a } j \leftrightarrow j+1 \text{ transition} \end{aligned} \qquad (39)$$

Similarly, expressions for S_3 itself and S_2 can be written.

Equations (38) and (39) are particularly interesting for the following reasons. Until now it has been implicitely assumed that $\mathrm{Im}(S_3)$ was a positive quantity, thus providing saturation-induced preference for plane polarization. Indeed, (39) shows that for the neon $\lambda = 1.152$ μm $j = 1 \leftrightarrow j = 2$ transition, as well as the $j = \frac{1}{2} \leftrightarrow j = \frac{1}{2}$ transition considered above, $\mathrm{Im}(S_3)$ is positive. However, this is not always the case. From (39) the rules are [8]:

$$\mathrm{Im}(S_3) > 0 \text{ for } \tfrac{1}{2} \leftrightarrow \tfrac{1}{2} \text{ and all } j \leftrightarrow j+1\ (j > 0), \quad (40a)$$

$$\mathrm{Im}(S_3) < 0 \text{ for } j \leftrightarrow j\ (j > 1), \quad (40b)$$

$$\mathrm{Im}(S_3) = 0 \text{ for } 0 \leftrightarrow 1 \text{ and } 1 \leftrightarrow 1. \quad (40c)$$

This striking result of the theory predicts three possible types of saturation-induced anisotropy. It may either lead to preference for plane polarization (case 40a), or to preference for circular polarization (case 40b), i.e. equal preference for the poles on the Poincaré sphere. Finally saturation-induced anisotropy may be absent (case 40c), in which case the dispersive part Re(S_3) also vanishes.

The theoretical predictions induced us to investigate a $j = 2 \leftrightarrow j = 2$ transition in the helium-neon laser, which is operative at a wavelength of $\lambda = 1.207$ μm. The experiments did indeed show that in this case a strong preference for circular polarization exists and that bistability corresponding to the two poles on Poincaré's sphere occurs [13] [5].

Experiments [13] [5] on the $\lambda = 1.523$ μm $j = 1 \leftrightarrow j = 0$ transition however did not confirm the neutral behaviour predicted by theory. Instead a clear preference for circular polarization was found, though of a much smaller magnitude than for the cases of type (40b). This residual effect could be explained [14] [15] by a mechanism not considered so far, the relaxation mechanism between the levels of a degenerate ($j = 1$) state. We could show that, if the nature of the operative collision processes is such that angular momentum relaxation of the degenerate level is faster than quadrupole relaxation, the observed preference is obtained It is of interest to note that, although in a quite different context and in a different atom, a difference between the two relaxation rates has already been found by other workers.

In this article it has been the aim of the authors to build up what is in their view a coherent picture of the subject. Full reference to important work of other workers in the field can be found in the articles quoted.

[13] H. de Lang and G. Bouwhuis, Physics Letters **20**, 383, 1966.
[14] D. Polder and W. van Haeringen, Physics Letters **25A**, 337, 1967.
[15] G. Bouwhuis, Physics Letters **27A**, 693, 1968.

Prof. Casimir presenting an award to a group of girl participants in the Final of the European Contest for Young Scientists and Inventors, an annual event arranged by Philips and held at Eindhoven in the Evoluon.

The temperature distribution in gas-filled incandescent lamps

J. Fitzgerald and H. Hörster

Important physical quantities in gas-filled incandescent lamps such as life and efficiency are controlled by the mass and heat flux from the filament. The description of these processes requires an accurate knowledge of the temperature and velocity fields of the inert gas in the bulb. In principle these fields can be calculated from the equations of fluid motion (Navier-Stokes), energy and continuity. Because of the mathematical difficulties with this differential-equation system only limited solutions are known. In the case of the gas-filled incandescent lamp the situation is further complicated by low symmetry and extreme temperature differences. An approximate solution has been suggested by I. Langmuir [1]. By assuming the existence of a stationary gas film surrounding the filament he reduced the problem to one of heat conduction. This concept was later used by W. Elenbaas [2] to calculate the mass transport. Although this approach adequately describes the total mass and heat losses in a normal incandescent light bulb it is insufficient to analyse the chemical transport reactions inside a halogen lamp. This requires the knowledge of the spatial concentration of the different tungsten-halogen compounds, which is controlled by the spatial temperature and velocity distributions of the inert gas. Because of the analytical difficulties it is necessary to use an experimental technique to obtain these spatial distributions. In view of the relationship between the temperature and velocity fields only one of these quantities requires to be measured.

One elegant method is the estimation of the temperature field by means of optical interferometry. During steady-state operation the pressure in an incandescent lamp is constant. Consequently the temperature T is inversely proportional to the density ϱ which in turn is proportional to the refractive index n:

$$T^{-1} \propto \varrho \propto n\,.$$

Hence, by experimentally determining the refractive-index field $n(x,y,z)$ the temperature distribution $T(x,y,z)$ can be calculated.

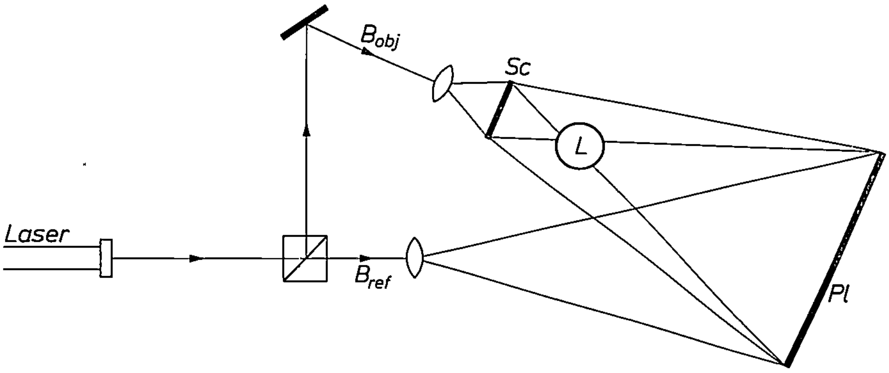


Fig. 1. The experimental arrangement. Coherent light from a laser is split into two beams, an object beam B_{obj} and a reference beam B_{ref}. These two beams recombine and interfere at the photographic plate *Pl*. When the developed photographic plate (hologram) is illuminated by the reference beam, a wave similar to the original object wave is reconstructed. *L* incandescent lamp. *Sc* scatter plate which diffuses the light to provide a spatially extended source.

The most frequently used classical interferometer for determining refractive-index changes in gases is the Mach-Zehnder interferometer. In this instrument two coherent beams of light are made to traverse different optical paths and are brought together to interfere. Any change in the interference pattern caused by the introduction of a test object can then be related to its refractive index. With the incandescent light bulb there are two refractive-index fields, namely the glass enclosure and the inert gas, and any changes in the interference pattern are due to their combined effect. In view of the poor quality of the glass it is not possible to eliminate the effect of the bulb by introducing an identical bulb in the other beam. One possibility would be to introduce optical flats into the bulb wall, but this

Dr. H. Hörster and J. Fitzgerald are with Philips Forschungslaboratorium Aachen GmbH, Aachen, Germany.

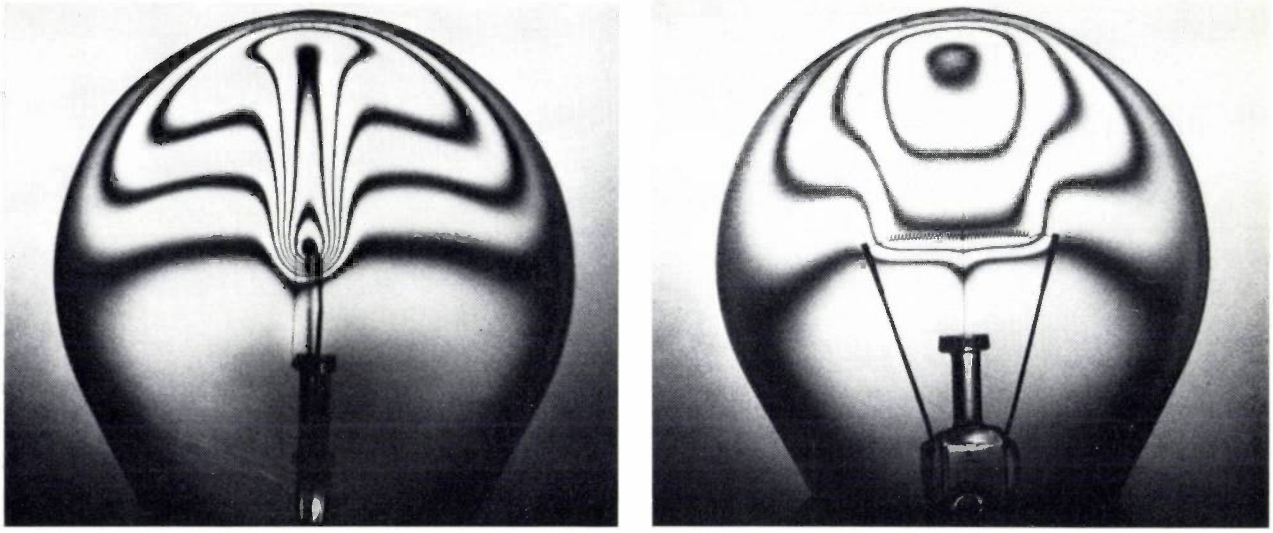


Fig. 2. Fringe pattern in a domestic light bulb (110 V/100 W) obtained by means of interference holography. In double-exposure holographic interferometry two exposures are taken corresponding to two different object conditions. In the reconstruction process the two corresponding object waves are simultaneously reconstructed and, due to their coherence, they interfere. In the case of a domestic light bulb the two exposures are taken with the bulb cold and with the bulb lighted respectively. *Left*: End-on view. *Right*: Side-on view.

would change the bulb symmetry and influence the temperature distribution. The interferometer has two further disadvantages inasmuch as it is expensive and requires precise alignment.

These obstacles can be overcome by the use of holographic interferometry. In general a hologram is formed by recording the interference pattern of two coherent waves (an object wave and a reference wave) on a photographic emulsion. The developed emulsion or hologram has the remarkable property that when illuminated by the reference wave it reconstructs a wave similar to the original object wave. In double-exposure holographic interferometry this procedure is repeated twice and two object waves corresponding to different conditions of the object are recorded on the same photographic plate. These two waves are simultaneously reconstructed in the reconstruction process and interfere to form a fringe pattern.

In the case of a light bulb the first exposure is taken with the lamp cold and the second when the lamp is burning (see *fig. 1*). During the first exposure the density and hence the refractive index n_0 of the inert gas are constant. When the lamp is burning, however, a three-dimensional temperature distribution is established. Associated with this three-dimensional temperature distribution is a three-dimensional refractive-index distribution. In the reconstruction process these two situations are superimposed and the two reconstructed object waves interfere and generate a fringe system. Because there is no change in the glass envelope between the two exposures it does not contribute to the fringe-order pattern.

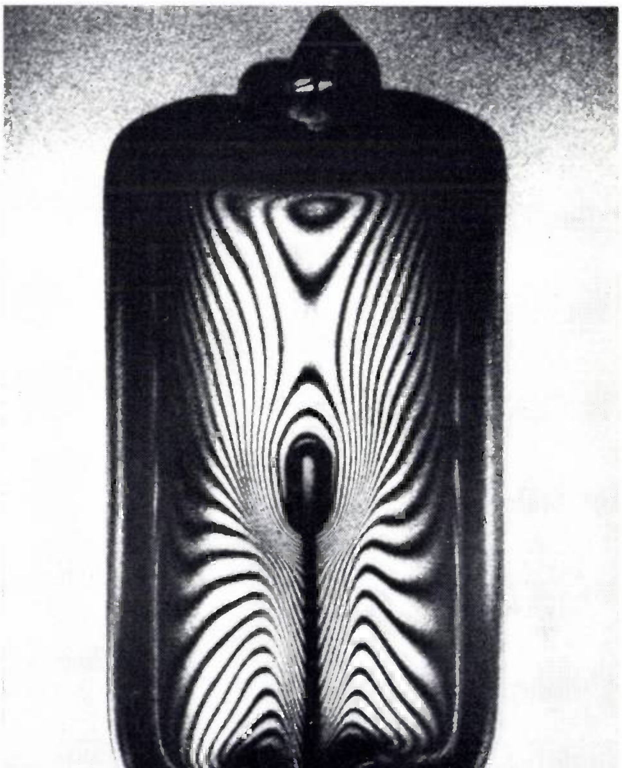


Fig. 3. The same as fig. 2*a* for a halogen projection lamp (12 V/100 W) with a filling of 4 atm xenon. Length 17 mm, diameter 10 mm.

[1] I. Langmuir, Phys. Rev. **34**, 401, 1912.

[2] W. Elenbaas, Philips Res. Repts. **18**, 147, 1963.

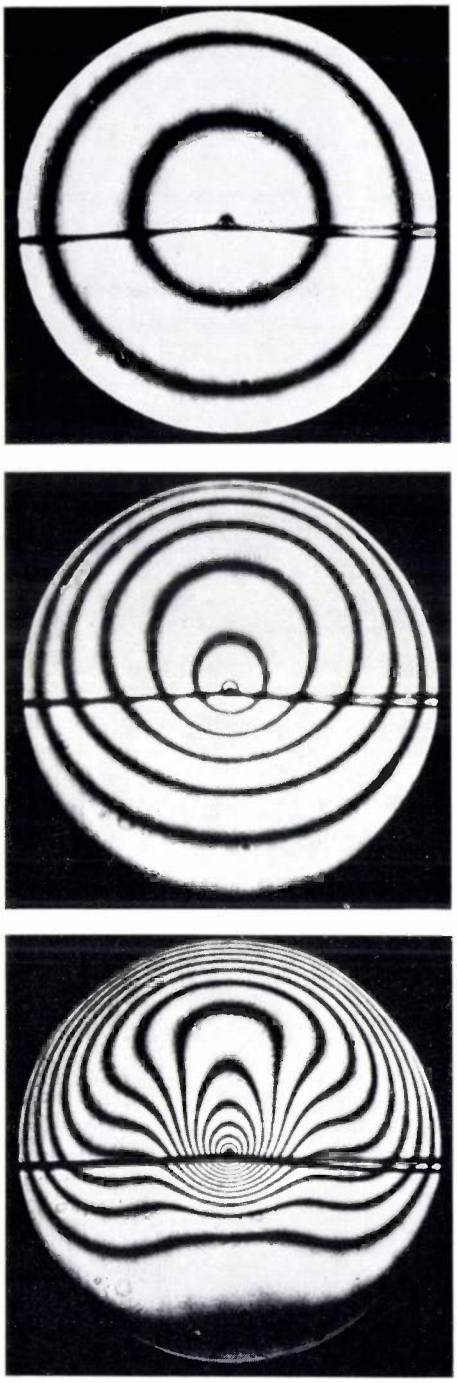


Fig. 4. Interference patterns in an axially symmetric argon-filled lamp (diameter of tube 4 cm; diameter of wire 200 μm) viewed along the filament. In this lamp each fringe order represents an isotherm. The effect of pressure variation at a constant filament temperature (1400 K) is demonstrated. In the tables *m* is the dark fringe order, counted from the centre.

pressure 30 torr, diameter of Langmuir film 2.8 cm

m	*T* (K)
1	1338
2	563
3	356

pressure 75 torr, diameter of Langmuir film 1.7 cm

m	*T* (K)
1	946
2	680
3	532
4	436
5	370
6	313

pressure 200 torr, diameter of Langmuir film 1.0 cm

m	*T* (K)
1	1295
2	1080
3	923
4	809
5	718
6	647
7	588
8	538
9	497
10	462
11	431
12	405
13	381
14	360
15	341

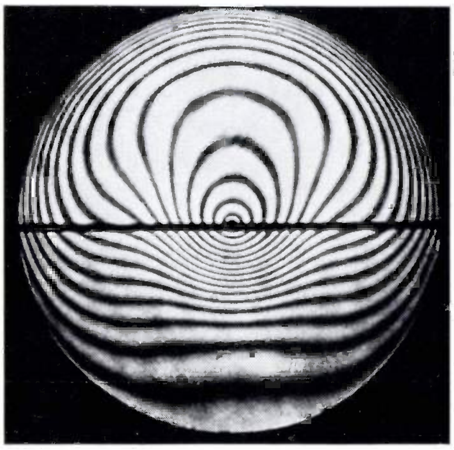


Fig. 5. Interference pattern in an axially symmetric lamp at a pressure of 200 torr argon and a filament temperature of 2350 K. Comparison with fig. 4*c* demonstrates that at a higher filament temperature the temperature distribution in the vicinity of the wire is more axially symmetric.

pressure 200 torr, diameter of Langmuir film 1.1 cm

m	*T* (K)
1	2150
2	1615
3	1295
4	1080
5	923
6	809
7	718
8	647
9	588
10	538
11	497
12	462
13	431
14	405
15	381
16	360
17	341

The fringe orders observed m are related to the value of the path integral of the change in refractive index by:

$$m = \frac{1}{\lambda}\int [n(x,y,z) - n_0]\,\mathrm{d}s,$$

where λ is the wavelength of the light and s is the optical path through the bulb. For a truly three-dimensional distribution the evaluation of $n(x,y,z)$ from this integral equation is a complicated task. Recently, however, H. G. Junginger and W. van Haeringen [3] have shown that the integral can be inverted to give an analytical expression for $n(x,y,z)$ in terms of the measured fringe-order distribution m.

Photographs of the fringe patterns obtained from a domestic light bulb are shown in *fig. 2.* Fig. 2*a* shows the fringe pattern viewed along the length of the filament and fig. 2*b* is a side-on view. In fig. 2*b* it can be seen that close to the wire one fringe order remains practically constant along the length of the filament. If end effects are neglected then to a first approximation the fringe order system in fig. 2*a*, the end-on view, can be regarded as a system of isotherms. Now, if a stagnant gas film exists, as the Langmuir theory predicts, then the temperature distribution in the immediate vicinity of the filament must take the geometrical form of the filament itself. As can be seen from fig. 2*a* this is approximately true just below the filament. The deviation from axial symmetry in the upper part is caused by strong convection. This is also the case in the small halogen lamp shown in *fig. 3.*

To overcome the difficulties in evaluating a three-dimensional refractive-index distribution the validity of the Langmuir film theory was investigated in an axially symmetric lamp. The lamp was designed in such a way that the fringe system represented the temperature distribution. In preliminary investigations the effect of gas pressure on the temperature distribution of the inert gas at a constant filament temperature of 1400 K was investigated. Results are shown in *fig. 4.* From these figures it can be seen that by increasing the pressure the influence of the convection becomes more pronounced and changes the temperature distribution. In the tables beside each of these figures the temperature of each dark fringe and the thickness of the corresponding Langmuir film is given.

By comparing the temperature distributions with the calculated film thickness it can be seen that the Langmuir film theory is only valid at extremely low pressures. By increasing the pressure the influence of convection distorts the axially symmetric temperature distribution. *Fig. 5* shows the temperature distribution again at a pressure of 200 torr but at a higher filament temperature (2350 K). This demonstrates that at higher temperatures the distribution becomes more axially symmetric in the immediate vicinity of the wire, i.e. the Langmuir theory becomes more exact. The remarkable change in the two pictures is due to the increase in gas viscosity with temperature.

This preliminary investigation shows that the application of the Langmuir theory to incandescent lamps is problematic. A detailed analysis of this topic will be given in Philips Research Reports.

[3] H. G. Junginger and W. van Haeringen, to be published in Optics Communications.

Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes, France.

Tunnelling in metal-semiconductor contacts under pressure

P. Guétin and G. Schréder

Introduction

Consider the contact between a metal and a highly doped N-type semiconductor containing free electrons (*fig. 1*). Because of the difference in the work functions of the two materials, the band structure of the semiconductor presents a noticeable curvature in the contact region where a space charge appears [1]. A "potential barrier" is formed which prevents electrons from flowing freely from one side to the other when a voltage is applied between the two materials. The presence of this barrier is responsible for the rectifying behaviour of the well known Schottky diodes. Such devices are made with materials of relatively low concentrations n of the dopant impurities ($n \approx 10^{16}$ cm^{-3}) and the rectifying process is related to the fact that electrons have to overcome the barrier in order to go from one electrode to the other. Thermal excitation is necessary in order to boost them *over* the barrier. The "thermionic emission" prevails at low doping levels and room temperature and is illustrated schematically in fig. 1 by path *1*. For a given total band bending, however, the space-charge layer, i.e. the barrier, is thinner the higher the bulk doping. At very high impurity concentrations, the barrier may become [1] so thin ($\approx$ 100 Å) that electrons possess a non-negligible probability to go *through* the part of the "forbidden" band which constitutes the barrier. At very low temperature, thermal excitation is inefficient and this conduction mechanism is called the "tunnel effect" (path *3*). At higher temperature both the thermal excitation and the tunnel transmission compete with one another (path *2*) and the corresponding regime is commonly called "thermionic field emission" [2]. We are essentially concerned here with the low-temperature (pure) tunnel effect.

Tunnelling is in fact a very general property of quantum systems. In particular, the mechanism is responsible for the nuclear emission of α particles. The first convincing observation of the tunnel effect in semiconductors was made in 1957 by L. Esaki [3]: the appearance of a *negative* incremental resistance in a highly doped P-N diode provided unambiguous evidence for tunnelling through the forbidden energy band. Basic research on "Esaki diodes" rapidly decreased after 1960 and gave way to the development of applications in electronic devices. After some time, physicists turned their attention to metal-semiconductor contacts (M.S.), in which tunnelling was discovered in 1966.

Although this type of tunnel junctions did not appear to be well suited to device applications, it was rapidly recognized as opening the way to highly interesting physical studies. This is largely due to the "spectroscopic" nature of this effect. Indeed, tunnelling constitutes a way of injecting free carriers into a solid at a controlled energy above the Fermi level. Since electrons obey Fermi statistics, it is clear that those which participate in the current move from an occupied state on one side of the barrier to an empty state on the other. Without entering into greater detail, it can therefore be understood that a careful analysis of the

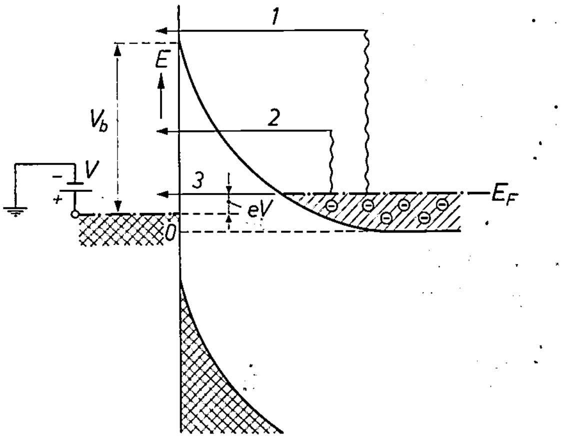


Fig. 1. Band structure in the region of a contact between a metal and an N-type semiconductor. The total energy E of the electrons (vertical axis) is measured from the bottom of the conduction band. The horizontal axis represents the distance from the interface. The semiconductor is so heavily doped that the electronic states of the conduction band are filled with electrons up to the "Fermi energy" E_F. The bending of the bands results in the creation of a potential barrier. e represents the absolute value of the electronic charge and eV is the energy difference between the two Fermi levels which corresponds to the applied voltage V. (V is defined as positive for the case of the figure; the + electrode of the battery is connected to the metal.) Three conduction mechanisms are possible according to whether the electrons are thermally excited over the barrier or go through it: *1* thermionic emission, *2* thermionic-field emission, *3* tunnel effect. In all cases the contact resistance is such that the potential drop in the bulk materials may be neglected and the full potential is applied across the contact.

Dr. P. Guétin and Dr. G. Schréder are with the Laboratoires d'Electronique et de Physique Appliquée at Limeil-Brévannes, France.

[1] See for example L. Heijne, Physical principles of photoconductivity, I. Basic concepts; contacts on semiconductors, Philips tech. Rev. 25, 120-131, 1963/64.

[2] F. A. Padovani, in: R. K. Willardson and A. C. Beer (editors), Semiconductors and semimetals Vol. 7A, Academic Press, New York 1971, chapter 2.

[3] L. Esaki, Phys. Rev. 109, 603, 1958.

conductance yields valuable information about the distribution of available states, i.e. the density of states, and their occupancy. In other words, the tunnel effect constitutes a fine band-structure analyser. In any solid, however, the carriers are not really free but are constantly interacting with their surroundings: other electrons, lattice atoms, impurities and so on. It turns out that the conductance to some extent reflects these interactions, which create small conductance "anomalies" at definite values of the applied voltage. Tunnelling may therefore be used as a tool in studying electron interactions.

Our own experiments have mainly been aimed at bringing new light on two of these aspects: the relationship between tunnel conductance and band structure, and the electron interactions with the lattice vibrations (phonons).

Although different kinds of tunnel junctions may be used for this purpose, we have chosen metal-semiconductor contacts for the following reasons:

They can be prepared much more cleanly and reproducibly than any other type of junctions.

The essential parameters of the contact may be determined by independent experiments so that quantitative agreement between theory and experiment may be hoped for.

They are very resistant to severe treatments like large temperature and pressure cycles.

Most of these aspects are related to the fact that the barrier is not an artificial layer but is an "intrinsic" property of the semiconductor-metal system.

After acquiring a quantitative understanding of the tunnel conductance in simple isotropic materials, we have devoted our attention to the influence of the band structure on the overall conductance and on electron-phonon interactions as well. Generally speaking, the conduction band of a semiconductor has several "valleys." located about different points of high symmetry in the momentum space [4]. For example, the Gunn effect in GaAs is due to the transfer under high electric fields of electrons to higher-energy valleys where their transport properties are drastically modified.

Applying pressure on a semiconductor represents an elegant method of modifying, in an externally controlled way, the band structure and especially the energy separation of the different sub-bands. We have used this technique together with tunnelling in the case of direct semiconductors [5] (GaAs) and also indirect ones (Ge). On one particular material (GaSb) we have been able [6], through tunnelling measurements, to follow the progressive "inversion" of the gap, i.e. the change from a direct gap to an indirect one.

Such experiments can be of great interest for the semiconductors in which the Gunn effect may exist. Indeed the inter-valley energy gap constitutes a fundamental parameter which can be determined directly by the tunnel effect under pressure.

We first present the typical behaviour of metal-to-N-GaAs contacts and show how the relatively weak conductance variations make them well suited to a detailed study of the superconducting properties of the metal under pressure. We then show the drastic changes in metal/N-GaSb junctions which appear near the gap "inversion" and also what kind of information can be extracted from the tunnel characteristics.

Before discussing our tunnel measurements, let us specify the configurations of the bands that we have studied.

Influence of hydrostatic pressure on the band structure

We now describe the modification under the influence of hydrostatic pressure of the conduction band with respect to the valence band of two compounds that we have studied with particular attention: GaAs and GaSb. *Fig. 2* illustrates the behaviour of both materials. In this schematic view we represent the "dispersion relation" $E(k)$, i.e. the electron energy plotted against momentum, in a particular direction. In fact we only show the parts which correspond to the nearest extrema of both bands. In other words, we consider the lowest

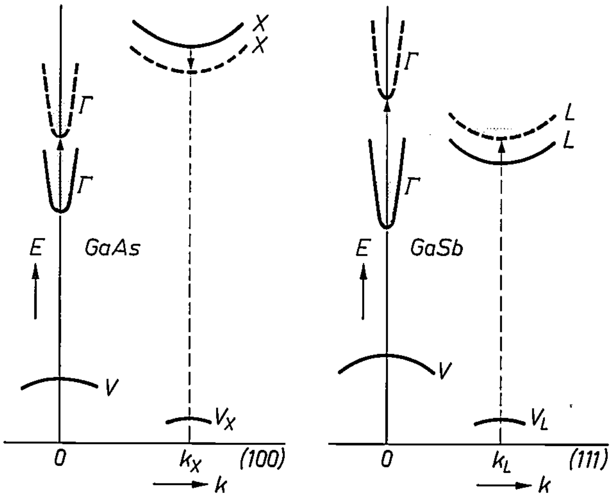


Fig. 2. Schematic view of the energy bands in a particular direction of the momentum space for GaAs and GaSb. The curves labelled V, V_X and V_L correspond to the extrema of the valence band and Γ, X and L correspond to the low-energy valleys of the conduction band. Solid curves and dashed curves refer to atmospheric pressure and high pressure respectively. In both materials the energy difference between the two types of valleys of this conduction band decreases when pressure is raised from the atmospheric value. The highest pressure available is not large enough to give a detectable transfer in GaAs of the electrons from the Γ to the X minima [7]. In GaSb, on the other hand, it is possible to bring the L valleys to a lower energy than that of the Γ valley. At high pressure the energy gap becomes "indirect". At the "inversion" pressure $P = P_0$ the lower edge of the Γ and L valleys lie at the same energy.

(highest) valleys of the conduction (valence) band.

The band configuration at atmospheric pressure is drawn in solid lines and the high-pressure configuration is presented in the dashed line. In GaAs, the energy of the Γ valley centred at $k = 0$ increases, whereas the energy of the X valleys decreases. At 17 kbar, the highest pressure used in our experiments, the energy separation between the X and Γ valleys is reduced from 0.36 eV to 0.16 eV approximately. This is not quite enough to allow any noticeable transfer of electrons from the Γ to the X valleys at very low temperature. Electrons remain in the central valley and therefore retain the same characteristics whatever the pressure: no drastic band-structure effect is expected to occur.

GaSb on the other hand constitutes a more interesting case. The electronic states of lowest energy are located in the Γ valley centred at $k = 0$ and in four (ellipsoidal) valleys in the (111) directions. At atmospheric pressure the gap is a direct one and the energy difference between both types of valleys $\Delta E = E_L - E_\Gamma$ is about 0.1 eV. When pressure is applied, all the valleys go towards higher energy but the Γ valley does so at a much faster rate than the L valleys, in such a way that the Γ valley catches up with the L valleys. Above 10 kbar, the Γ valley is located at higher energy than the L valleys: the gap has become "indirect". The "gap inversion" occurs around 10 kbar. Electrons have been progressively transferred from the Γ to the L valleys, where their behaviour is quite different. The effective mass is lower in the Γ states than in the L ones: $m_L^* \gtrsim 2\, m_\Gamma^*$.

Whereas at low pressure the band structure is analogous to that of GaAs, at higher pressure it is quite similar to the Ge one. Tunnel measurements not only confirm this analogy but provide a very sensitive method of observing the progressive changes around the gap inversion.

Now that we have established the framework of our tunnel studies, let us turn to a few experimental details.

Experimental methods

Metal-semiconductor contacts were made by cleaving a bar of semiconductor under ultra-high vacuum conditions in a vapour stream of the evaporating metal. During evaporation, the residual pressure in the vacuum system is kept below 5×10^{-9} torr and a mask punched with holes is placed in front of the clean surface immediately after it has been cleaved. Contacts on the metal "dots" are subsequently secured by bonding a thin wire with a droplet of silver paste. The ohmic contact on the semiconductor side is realized by a classical alloying method before insertion in the vacuum system.

The pressure apparatus consists essentially of a high-pressure "bomb" filled with isopentane as a transmitting fluid. The samples are positioned on an obturator equipped with seven electrical feedthroughs. The bomb is connected via a capillary tube to the main press where the pressure is measured with a manganin gauge. As the bomb cools down, the isopentane freezes at a temperature which depends on pressure (approximately 145 K at 3 kbar and 260 K at 16 kbar). The value of the pressure P is controlled during this operation.

The junction characteristics are studied by conventional tunnelling-circuit electronics, namely by applying a d.c. current and a small modulation current of constant amplitude and a frequency of 500 Hz. The fundamental voltage, proportional to the incremental resistance $\mathrm{d}V/\mathrm{d}I$, is processed in a lock-in phase detector and is plotted against the applied voltage. The second harmonic signal is also detected and can be directly related to $\mathrm{d}^2I/\mathrm{d}V^2$, the derivative of the incremental conductance with respect to the applied voltage. The qualitative features of the tunnelling characteristics may be understood by simple qualitative arguments. It is easy to verify that the width of the potential barrier for the electrons of the highest energy involved in tunnelling is a maximum at zero bias. It decreases for both positive and negative bias, which leads to an increase of the associated barrier-transmission probability. This results in the decrease of the incremental resistance at high applied voltage of either polarity. The resistance presents a maximum value at a low voltage ($V \lesssim 50$ mV).

Influence of pressure on metal/*N*-GaAs contacts

When pressure is applied to a metal/*N*-GaAs contact, as discussed previously, no drastic change of the band structure of GaAs can be expected, but only a reasonable increase of the width of the energy gap. This increase results in an enhancement of the barrier height and of the effective mass. On the other hand we can neglect the pressure dependence of the dielectric constant, the carrier density and the area of the contact.

The incremental resistance of the contact is shown in *fig. 3a* at atmospheric pressure and 16.4 kbar. Both

[4] See for example J. S. Blakemore, Solid State Physics, Saunders, Philadelphia 1970.

[5] P. Guétin and G. Schréder, Solid State Comm. 9, 591, 1971 (No. 9).

[6] P. Guétin and G. Schréder, Phys. Rev. Letters 27, 326, 1971 (No. 6).

[7] The electron transfer responsible for the Gunn effect occurs at a higher electric field and temperature than those used in the tunnel experiments.

curves have been normalized at $V = 0$ and it clearly appears that the shape has only undergone minor changes. We have plotted in fig. 3*b* the quantitative variations of the resistance $R(0)$ at $V = 0$ with the applied pressure on a semilogarithmic scale. This resistance increases exponentially with pressure by a factor of 10 for each 15 kbar. Such variations can be considered as small for tunnel standards and are essentially related to the increase of the energy gap and, subsequently, of the barrier height V_b.

The relatively small values of the changes produced in the semiconductor material make these contacts very suitable to the study of the metal properties under pressure. As an example, the superconducting behaviour of lead has been investigated in Pb/*N*-GaAs contacts. In the superconducting state [8] electrons enter into a strong interaction in pairs through the intermediary of the lattice vibrations. Unlike the normal state, the superconducting state is characterized by a gap — a forbidden zone — of width 2Δ in the spectrum of possible energy values of the electrons around E_F — the Fermi energy in the metal. Since tunnelling involves electrons which come from an occupied state and go into an empty available state, it is clear that, as long as the absolute value of the applied voltage is smaller than Δ — half of the superconducting gap —, no current should flow through the contact at infinitely small temperature. The insert of *fig. 4* illustrates this particular feature of the tunnel current at 4.2 K and shows that it can be used directly as a measure of the superconducting gap. Fig. 4 describes the results of such measurements in the case of lead. The gap width 2Δ decreases linearly with pressure from 2.7 meV at atmospheric pressure to 2.3 meV at 15 kbar. These results compare very well with data obtained at lower pressure on metal-oxide-metal (M-O-M) junctions. In M-O-M structures, the barrier is provided by an oxide layer (about 30 Å) which must not present any pinholes. Applying pressure to such junctions without breaking through this ultra-thin layer is something of a *tour de force* and the probability of success has been shown to become prohibitively small when the pressure is higher than about 10 kbar. Metal semiconductor contacts have been found much stronger than M-O-M contacts and in this respect also constitute the best available tunnelling tool.The same kind of study can be carried out with all kinds of superconducting metals and alloys. More generally, one can state that metals presenting an anomalous density of states in the vicinity of the Fermi energy can be conveniently studied under pressure in a metal-semiconductor contact made on a semiconductor which, like GaAs, experiences only smooth and small changes in the corresponding pressure range. We think here particularly of the semimetal-semiconductor transition where the opening up of a forbidden energy gap should manifest itself as a resistance peak about $V = 0$. On the semiconductor side, there is of course some interest in determining the pressure coefficients of the phonon energies which can be deduced from the shift of the structures displayed by the second derivative curves.

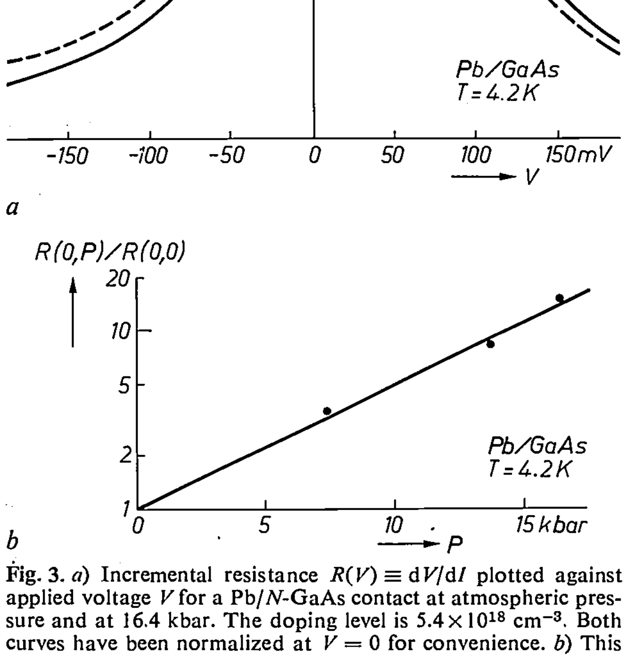


Fig. 3. *a*) Incremental resistance $R(V) \equiv dV/dI$ plotted against applied voltage V for a Pb/*N*-GaAs contact at atmospheric pressure and at 16.4 kbar. The doping level is 5.4×10^{18} cm^{-3}. Both curves have been normalized at $V = 0$ for convenience. *b*) This plot (semilogarithmic scale) illustrates the change with pressure of $R(0)$, the resistance at zero voltage, for a Pb/*N*-GaAs contact.

Now that we have discussed an example of the use of the technique in investigating the metal properties, we shall turn to the observation of band-structure effects in the semiconductor.

Influence of pressure on metal/*N*-GaSb contacts

In this section we shall outline the main differences between tunnelling in the direct- and indirect-gap configurations and emphasize the critical behaviour which shows up around the "gap inversion". We first discuss the main features of the incremental resistance at various pressures and show how they lead to valuable information on the band structure.

Fig. 5 gives the zero-voltage resistance $R(0)$ as a function of pressure for a typical Pb/*N*-GaSb sample containing 3.5×10^{18} impurities per cm^3. The points correspond to experimental data and the solid line gives the theoretical prediction. Comparison with fig. 3*b* shows that the "sensitivity" to pressure changes of the contact is roughly 10^4 times larger for GaSb than GaAs! Another feature which does not exist in M/*N*-GaAs contacts is the sharp break where $R(0)$ jumps by more than two orders of magnitude. The pressure $P_1 = 12$ kbar at which this discontinuity occurs is slightly higher than the expected "inversion" pressure P_0. Both of these features are closely related to the fact that the energy gap changes from a direct to an indirect one and that the electrons in the semiconductor are transferred from the Γ to the L valleys.

The existence of the resistance discontinuity is clearly explained if we recall the approximate form of the transmission coefficient:

$$D \approx \exp\,[-2 \int_{x_1}^{x_2} \mathrm{k}(x)\mathrm{d}x],$$

with

$$\mathrm{k}(x) \approx \left\{\frac{2m^*}{\hbar^2}\,[V(x) - \omega + \varepsilon_{||}]\right\}^{\frac{1}{2}} \approx v(m^*)^{\frac{1}{2}}.$$

The lower the effective mass, the larger the transmission coefficient. Below 10 kbar, tunnelling is essentially due to the electrons which are located in the low-mass Γ valley ($m_\Gamma \approx 5.10^{-2}\,m_0$) and are relatively easily transmitted through the barrier. Above 12 kbar the gap is indirect and near $V = 0$ tunnelling electrons come from the L valleys which present a high mass in the direction of tunnelling ($m_L{}^* \gtrsim 10^{-1}\,m_0$) and only give poor transmission.

Since the argument of the exponential factor is typically 15 to 20, the relatively small difference in the mass results in a high resistance change.

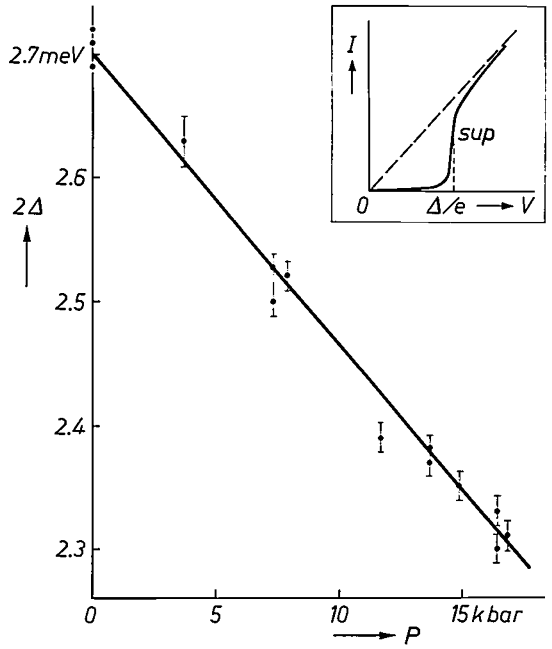


Fig. 4. Variations in the "superconducting gap" 2Δ of lead with pressure. The insert illustrates the shape of the *I*-*V* characteristics of the tunnel contact. The dashed and the solid line correspond to the metal in the normal state and in the superconducting state respectively. As long as the voltage does not reach the value Δ/e, electrons in the semiconductor find no available energy state in the metal to go into and the current is practically zero. Above $V = \Delta/e$ the current rises sharply and rapidly tends to the normal-state value.

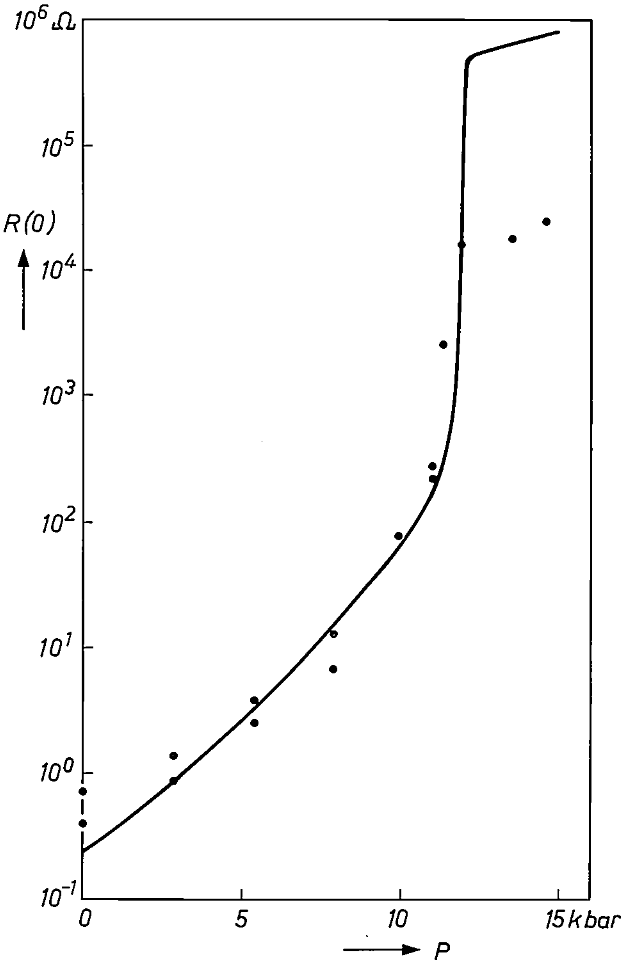


Fig. 5. Incremental resistance at $V = 0$ plotted against the pressure P for a Pb/*N*-GaSb contact. The points represent experimental data and the solid curve illustrates the results of a numerical computation. The sharp resistance discontinuity occurs at $P = P_1$, i.e. where the Γ valley becomes totally empty. The change of the effective mass of the relevant tunnelling electrons is responsible for this discontinuity. At $P \lesssim 10$ kbar the gap is direct and for $P > P_1 \approx 12$ kbar the gap is indirect.

Two aspects may be discussed in more detail: the agreement between experiments and theory in both configurations, and the position of the discontinuity of $R(0)$.

In the direct-band configuration, the computed and the observed resistance values are in very good agreement and therefore we can conclude that the current coming from the Γ valley is the direct current, which is clearly understood (fig. 1, path *3*). In the indirect configuration, the observed resistance is roughly 75 times less than the computed resistance and this suggests that a new parallel tunnel path adds to the usual direct path. This discrepancy may be traced to the diffusion of electrons on the ionized impurities and will not be considered further in this article.

[8] J. Volger, Philips tech. Rev. 29, 1, 1968.

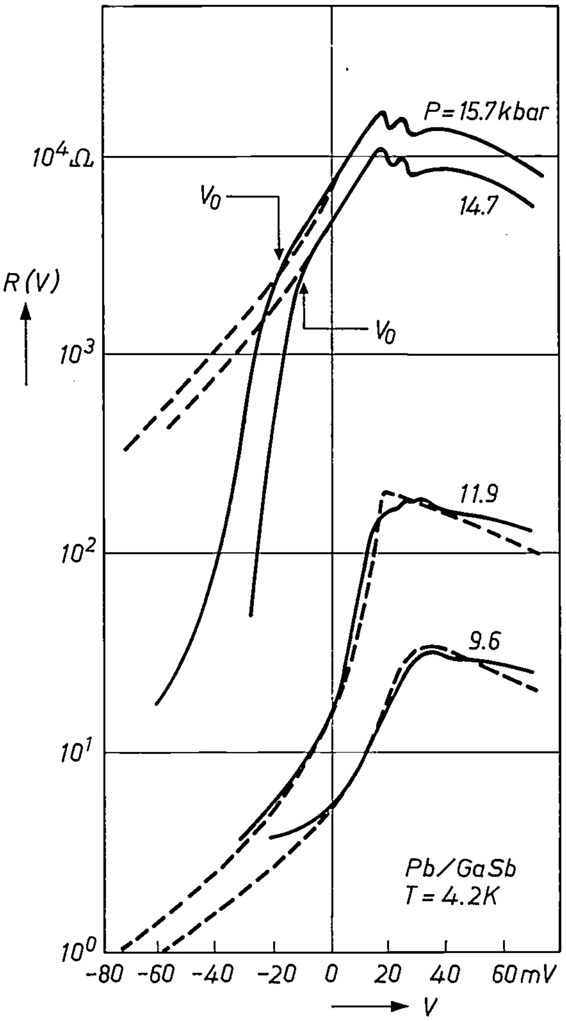


Fig. 6. Characteristics of the incremental resistance $R(V) = \mathrm{d}V/\mathrm{d}I$ plotted against the applied voltage V at different pressures. Dotted curves illustrate the results of the computation and have been normalized to the experimental curves at $V = 0$. The two lower curves ($P = 9.6$ and 11.9 kbar) correspond to pressure smaller than P_1 and the resistance is therefore determined mainly by the Γ valley. The shape of these two curves is in good agreement with theory.

The high-pressure curves ($P = 14.7$ and 15.7 kbars) are drawn for $P > P_1$ where the gap is indirect. Near $V = 0$ the experimental data agree very well with the computation in which only the L valleys are considered.

Two striking features appear on these curves: the incremental resistance drops sharply at voltages more negative than a pressure-dependent threshold V_0 — the "resistance kink" —, and at forward bias two resistance breaks appear at voltages which are quite independent of the pressure. These two aspects suggest that new conductance mechanisms add their contributions to the common direct one which only involves the L valleys.

Since the Γ valley has a lower mass value, it provides the majority of the current as long as electrons fill up some of its states. The sharp discontinuity appears at the pressure which corresponds to the total emptying of the Γ valley: all electrons have been transferred to the L minima. Pressure P_1 is then slightly higher than P_0: $P_1 \approx P_0 + 1.5$ kbar.

Now that we have acquired some understanding of the absolute value of the incremental resistance at zero voltage, let us consider in greater detail its spectroscopic aspect, in other words its variations with the applied voltage V. *Fig. 6* shows four typical curves. The two lower curves have been chosen in the "direct" configuration and the higher curves correspond to the "indirect" gap. Dotted lines describe the results of the computation when fitted at $V = 0$ for convenience. Whereas the low-pressure curves again present a good agreement with theory, the high pressure curves show two important features which may be related to the band structure of GaSb. These are: a resistance "kink" that appears at voltages more negative than V_0, marked with an arrow — the threshold V_0 depends on the pressure — two clear down steps in resistance located at voltages which hardly vary at all with pressure. We shall now discuss both features in turn.

Resistance kink at reverse bias

Above $P = P_1$ the lowest valleys are the L valleys and the dotted parts of the two higher curves represent what can be expected at reverse bias from direct tunnelling into the L states. The agreement with experiment is satisfactory between $V = 0$ and $V = V_0$. The sharp resistance drop beyond the V_0 threshold indicates clearly that another conduction mechanism shows up in parallel with the direct tunnelling into the L states. The higher the pressure, the more negative is the value V_0 of the threshold. This resistance kink may be interpreted by considering the aspect of the band structure near the interface (*fig. 7*). At $P > P_1$, the Γ valley is located at higher energy than the L valleys and is empty. Near zero bias, the conductance is due to the current I_1 into the L states, which are the only ones available at such energy. At voltages more negative than a certain threshold V_0, empty Γ states become available at the energy of the Fermi level of the metal and electrons can flow directly into the Γ states. This results in a current I_2 in parallel with I_1. The corresponding barrier shown in dashed lines is slightly higher and thicker than it is for the L states. Altogether these parameters are such that $I_2 \gg I_1$ and therefore the resistance drops quickly for $V < V_0$. This effect constitutes a very direct method of measuring ΔE, i.e. the separation between L and Γ valleys. We have:

$$eV_0(P) = \Delta E(P) - E_F^{(L)}.$$

The degenerate Fermi energy in the L valleys, $E_F^{(L)}$, can be calculated exactly and is quite independent of the pressure above $P = P_1$. In fact the sharp resistance drop when $V = V_0$ still appears (at forward bias) at a

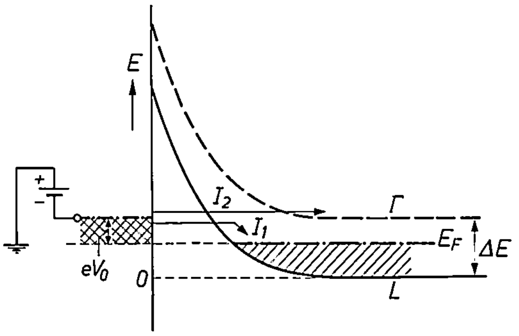


Fig. 7. Schematic view of the band structure in the contact region for $P > P_1$. The Γ valley is located at a higher energy than the L valleys, and the L states are all empty. For voltages more negative than the threshold V_0, electrons can be injected directly into the Γ valley, giving rise to a marked resistance drop because of the much lower effective mass $m_\Gamma^* \ll m_\Gamma^*$. Current I_2 adds to the L valley current I_1. The threshold V_0 is a direct measure of the energy separation between the Γ and the L valleys.

pressure slightly lower than P_1. The results of these measurements are collected in *fig. 8.* V_0 varies linearly with pressure and since the slow dependence of $E_F^{(L)}$ can be readily computed, tunnel characteristics also provide a direct measure of $\Delta E(P)$ that we can extrapolate to $P = 0$:

$$\Delta E(P) = \Delta E(0) - \alpha P,$$

with

$$\alpha = 9.6 \text{ meV/kbar and } \Delta E(0) = 100 \pm 5 \text{ meV}.$$

The rate of variation with pressure is in very good agreement with the latest determinations reported in the literature. On the other hand, the atmospheric-pressure value of ΔE appears to be somewhat higher than those reported. Values of $\Delta E(0)$ reported in the past range from 80 to 95 meV according to the authors and the method used. It is to be noted that apart from tunnelling, all the experimental techniques are very indirect and depend on the exact knowledge of numerous parameters. The tunnel effect represents at this point the most direct and accurate method of measuring such quantities. Whereas the value of the threshold is independent of m^* and E_F, it is clear that the value of the resistance for $V < V_0$ is sensitive to such quantities. This technique can therefore be considered to measure the effective mass.

Phonon-assisted tunnelling

The second remarkable feature which appears when the gap is indirect (fig. 6, top curves) consists of a pair of breaks in the incremental resistance at forward bias. They occur at voltages which do not depend much on the applied pressure and can best be studied by twice differentiating the current with respect to the voltage. The derivative d^2I/dV^2 of the incremental conductance with respect to the voltage is shown in *fig. 9a* for two pressures. The lower curve corresponds to atmospheric pressure when the gap is direct and the upper one has been drawn for one of the highest pressures available. Peaks (dips) correspond to conductance step increases (decreases). The lower curve presents essentially two dips labelled *TA*, *LA* (lead) which are characteristic of the superconducting lead electrode [9]. At high pressure, on the other hand, four extra structures can be clearly observed with a highly pressure-dependent amplitude. These have been labelled *TA*, *LA*, *TO* and *LO*. In order to extract the actual amplitude and shape of the *TA* peak, it is necessary to subtract the background curve, which is to some extent perturbed by the presence of the "superconductive" anomalies.

Before going into the physical interpretation of these "anomalous" structures, it is instructive to look at the dependence on pressure of the amplitude of the two main peaks (*LA*, *TO*). Fig. 9*b* displays the relative change $\Delta R/R$ of the resistance drop corresponding to the two main structures. We shall come back later to this figure and attempt to interpret the observed behaviour. In the meantime it is of great interest to note that the effect again seems to be closely related to the configuration of the bands since the amplitude jump appears very close to the critical pressure P_1 for total emptying of the Γ valley. The interpretation of these effects can be found if it is assumed that electrons in a solid are not exactly free but interact with their surroundings, particularly with the lattice vibrations. This is strongly suggested by the experimental fact that the second-derivative peaks occur at voltages which correspond closely to characteristic phonon frequencies.

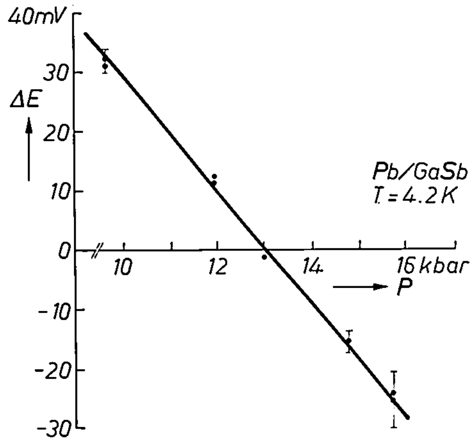


Fig. 8. This figure shows the collected results of the measurements of V_0. Since $E_F^{(L)}$ is approximately pressure-independent, this curve yields a direct measurement of the separation between L and Γ valleys, $\Delta E(P)$, plotted against pressure.

[9] They are related to well-known anomalies of the superconductive density of states and are fairly independent of the pressure. These two dips ought to appear in the same way at any pressure. Apart from these two structures, the second derivative taken at atmospheric pressure does not give any other appreciable feature.

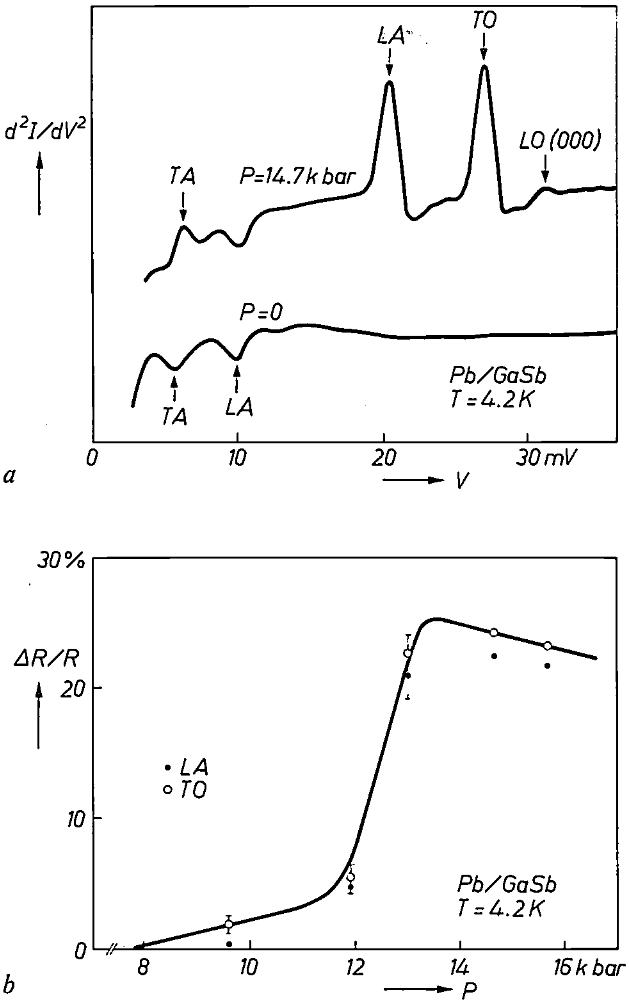


Fig. 9. *a*) Second-derivative characteristics plotted against applied voltage (forward bias) of Pb/N-GaSb contacts at two different values of the pressure. The lower curve corresponds to atmospheric pressure and shows two dips related to the superconducting properties of the lead electrode. The higher curve corresponds to $P = 14.7$ kbar ($> P_1$) and has four extra peaks which correspond to characteristic phonon energies in the GaSb. The two largest peaks labelled *LA* and *TO* correspond to the two resistance breaks appearing at forward bias on fig. 6. *b*) Amplitude of the relative change in the incremental resistance $\Delta R/R$ at the two main resistance breaks (fig. 6) plotted against the applied pressure. The amplitude rises sharply at $P \approx P_1$.

Consider the common features of the dispersion relation, i.e. the energy $\hbar\omega_q$ as a function of the wave vector q, of the phonons for a particular direction of high symmetry in the crystal (*fig. 10*). This information is usually obtained by experiments on inelastic scattering of neutrons. The different "branches" correspond to the various types of vibrations of the atoms of the crystal unit cell with respect to one another. Longitudinal (*LO*, *LA*) and transverse phonons (*TO*, *TA*) are related to compression and shear waves respectively. Phonons are called "acoustical" (*LA* and *TA*) or "optical" (*LO* and *TO*) according to whether the atoms in the crystal unit cell vibrate in phase or with the opposite phase. The maximum wave vector q_0 is related to the value a of the lattice constant in the relevant direction: $q_0 = \pi/2a$.

In materials such as germanium and gallium antimonide, it has been shown that the average wave vector of the L electronic states is equal to q_0 in the same (111) crystal direction (fig. 2):

$$k_L(111) = q_0(111).$$

The vibration $q_0(111)$ is called a "zone-edge" phonon. In fact the voltages of the second derivative peaks correspond closely to the energies of zone-edge phonons in the (111) direction. This suggests that the conduction mechanism which gives rise to the sharp resistance breaks is related to "inter-valley" processes where tunnelling electrons jump from the L to the Γ valleys. The momentum change is provided by the emission of a zone-edge phonon:

$$k_L(111) - k_\Gamma(0) = q_0(111).$$

Now that we have acquired a feeling for the relevant effect leading to the anomalies at forward bias, we wish to explain its influence on the tunnelling current more precisely.

Fig. 11 may assist the understanding of the physical process. In this picture we represent schematically the "constant-energy surface" in the momentum space. This momentum space has been chosen only two-dimensional, for simplicity, with k_x corresponding to the direction of the tunnel current and $k_{||}$ representing

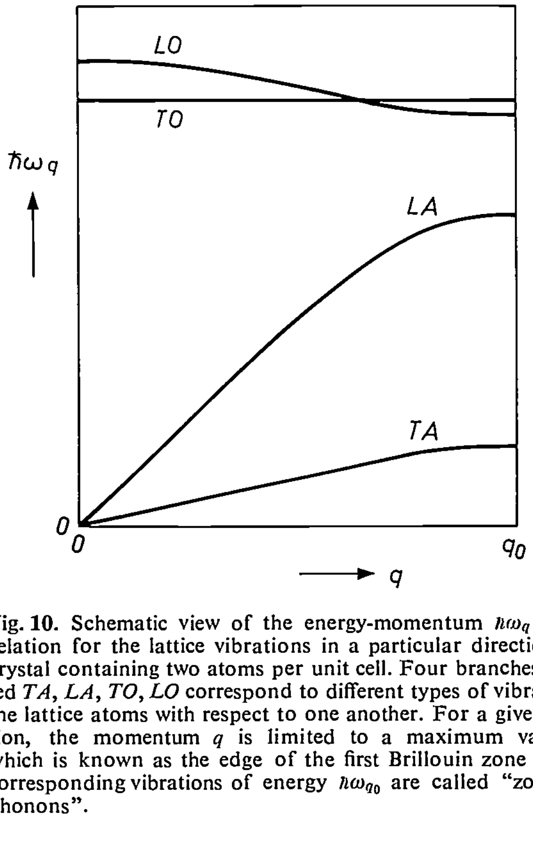


Fig. 10. Schematic view of the energy-momentum $\hbar\omega_q = f(q)$ relation for the lattice vibrations in a particular direction of a crystal containing two atoms per unit cell. Four branches, labelled *TA*, *LA*, *TO*, *LO* correspond to different types of vibration of the lattice atoms with respect to one another. For a given direction, the momentum q is limited to a maximum value q_0, which is known as the edge of the first Brillouin zone and the corresponding vibrations of energy $\hbar\omega_{q_0}$ are called "zone-edge phonons".

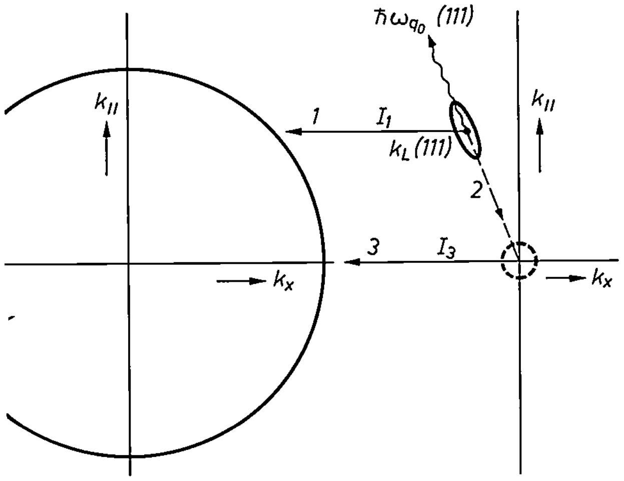


Fig. 11. This figure provides a schematic view of the two-step tunnelling process which involves the emission of a zone-edge phonon in indirect-gap semiconductors. We represent for each material the "constant-energy" surface of the electronic states in the momentum space which has been symbolically reduced to two dimensions. The metal (semiconductor) is shown on the left (right) of the vertical partition. k_X corresponds to the direction of the tunnel current and $k_{\|}$ is the momentum parallel to the junction plane. The reduction to two dimensions is made by replacing the two coordinates k_y and k_z in this plane by a single coordinate: $k_{\|}$. The metal is isotropic and the wave vector is very large at the energies involved in tunnelling. In the indirect-gap semiconductor, at the same energies, the available electronic states are all concentrated in the ellipsoidal L valleys. For simplicity a single valley only is shown in the picture. The dashed circle around $k = 0$ in the semiconductor expresses the fact that, if no steady electronic state exists at the relevant energies, the low-mass Γ valley is located at a slightly higher energy.

When a forward bias is applied, direct tunnelling (path *1*) coming from the L valleys gives rise to a current I_1. A two-step mechanism (path *2-3*) is also possible. In the first step the electrons in the L state emit a zone-edge phonon of energy $\hbar\omega_{q_0}$ (111) and are sent to the low-effective mass region at $k \approx 0$ (path *2*). The next step consists in the actual tunnelling mechanism towards the metal (path *3*).

the momentum parallel to the junction plane. The metal (semiconductor) is represented to the left (right) of the dotted partition. The "constant energy" surface limits the regions of the momentum space in both materials where available electronic states exist at the energies involved in tunnelling. In the semiconductor, at $P > P_1$ the gap is indirect and only the L valleys possess available states at the energies considered. We have further simplified the drawing by showing a single L valley. At the same energies, the constant-energy surface in the metal is much larger than in the semiconductor. Just as in optics, direct (specular) transmission implies a conservation of $k_{\|}$, the momentum parallel to the junction plane. The direct current I_1 coming from the L valleys at $P > P_1$ is represented by path *1*. The corresponding incremental resistance of the contact can be calculated and the results have been represented by dashed lines on the upper curves of fig. 6. The transmission probability corresponding to path *1* only involves the characteristics of the L valleys and is relatively low because of the high effective mass m_L*.

There is, however, a region of the momentum space which corresponds to a higher transmission probability. This is the region near the centre of the zone ($k \approx 0$) which is limited by a dashed circle. Indeed, although there is no available stationary state of momentum $k \approx 0$ at the relevant energy, we note that the low-mass Γ valley is located at energies which are only a little higher. As a matter of fact, *at reverse bias*, electrons are injected into the semiconductor at an energy which increases with the absolute value of the voltage (fig. 7); we have seen that, beyond a threshold V_0, the conductance of the direct tunnelling into the Γ valley predominates. In our particular problem, we look at *forward bias* and direct tunnelling from this empty valley is forbidden. However the proximity of the Γ valley with a low mass provides a high transmission probability, along path *3*, to electrons which might be sent from the occupied L states by some kind of mechanism. This mechanism must typically provide the electrons with the momentum k_L(111) [10]. The experimental features which we are presently discussing correspond to this two-step mechanism in which the momentum is provided by a zone-edge phonon. At very low temperature the equilibrium lattice vibrations are extremely ineffective and phonon absorption is very unlikely to occur. Since such a phonon has a finite energy $\hbar\omega_{q_0}$, it is only when the applied voltage V is larger than $\hbar\omega_{q_0}/e$ that electrons do possess enough "spare energy" to give up the required energy to the phonon. The current "assisted" by the phonons presents a threshold: $|eV_{\text{ph}}| = \hbar\omega_{q_0}$ and leads therefore to resistance drops at the voltage corresponding to the energies of the various phonon branches. Although the electrons have a low probability of emitting zone-edge phonons, the transmission along path *3* is so much easier than along path *1* that the total conductance related to the two-step process is sufficiently high to yield measurable effects.

This two-step tunnel mechanism has allowed us to measure accurately for the first time the energies of the relevant phonons in GaSb. In fact, a more thorough tunnelling study indicates that only three second-derivative peaks (fig. 9*a*) can be ascribed to zone-edge phonons (*TA*, *LA*, *TO*) and this strongly suggests that both optical branches are practically degenerate at the edge of the zone. The fourth structure labelled *LO*(000) corresponds to the $q \approx 0$ longitudinal optical phonons and will not be discussed here.

We are now able to explain qualitatively the dependence of amplitude on pressure (fig. 9*b*). At pressures

[10] It is to be noted that the absence of an available state at $k \approx 0$ at the energy of the tunnelling electrons does not forbid this two-step process, since the intermediate step has a short lifetime and consequently an ill-defined energy.

lower than P_1, the Γ valley contains electrons, in other words Γ states are available at the tunnelling energy for all values of the voltage. Consequently direct tunnelling to and from the Γ valley is allowed. This one-step mechanism obviously has a greater probability than the two-step mechanism since there is no need for the Γ electron to emit a phonon. The relative strength of the two-step process goes sharply to zero when pressure decreases from the value P_1. Beyond P_1, the slow decrease in amplitude must be attributed to the progressive shift of the Γ valley towards higher energy above the L states. We have confirmed this fact by a set of experiments under pressure on N-type germanium, which is an indirect-gap material at atmospheric pressure.

The study of the variation with pressure of the position and amplitude of the phonon conductance "anomalies" therefore provides interesting information on the electronic band structure, on the vibronic properties of the lattice atoms and also on the coupling between electrons and phonons.

Although atmospheric pressure tunnelling yields some valuable information on the properties of the materials and on the characteristics of the metal-semiconductor contact itself, it is clear that the external control of a crucial parameter such as the pressure constitutes a fruitful additional technique. In this article we have shown how this method can be used to investigate the band structure of the semiconductor, its vibronic properties and superconductivity in the metal electrode. This study has led us to a very direct determination of the energy separation of different valleys of the conduction band and to a first measurement of zone-edge phonon energies in GaSb. This technique may prove most valuable in the study of new materials such as small-energy-gap or magnetic semiconductors and ternary compounds used in optoelectronic devices, which present a direct or an indirect gap according to their composition. Metal-semiconductor contacts appear quite appropriate to this kind of experiment since they can be made in a very reproducible and reliable manner by the cleavage technique and under such conditions their behaviour can be understood quantitatively [11].

[11] P. Guétin and G. Schréder, J. appl. Phys. 42, 5689, 1971 (No. 13).

The lateral skin effect in a flat conductor

V. Belevitch

Introduction

The classical skin effect is a well-known phenomenon, a convenient reference being the recent survey by H. B. G. Casimir and J. Ubbink [1]. In a wire of circular cross-section, the radial distribution of the current density is a Bessel function of argument proportional to the square root of the frequency At d.c., the density is uniform, whereas at high frequency the current is approximately concentrated in a peripheral sheet of thickness

$$\delta = \surd(2/\mu\varrho\omega) \qquad (1)$$

(μ = permeability, ϱ = conductivity, ω = angular frequency; practical electromagnetic units), called the *skin depth*. Correspondingly, the resistance R per unit length increases from the d.c. value R_0, at first in proportion to the square of the frequency, but ultimately it tends to infinity as the inverse of the skin depth, and hence as the square root of the frequency.

Qualitatively, the preceding results hold for massive conductors of any cross-section with a smooth boundary, the only difference being that the tendency of the current density to concentrate towards the surface is more marked at the points where the curvature is greatest. For instance, in a conductor of elliptic cross-section, the density at the ends of the major axis will ultimately be larger than at the ends of the minor axis.

The effect is particularly marked when the eccentricity of the ellipse is very large, and an interesting problem is the limiting case of a very thin elliptic cylinder (*fig. 1a*), which almost reduces to a flat strip of width $2a$. The problem of major practical importance is, of course, the thin rectangular strip (fig. 1*b*) widely used in printed circuitry and in microwave applications. It so happens, however, that the elliptic case is amenable to analytical treatment whereas the rectangular case can only be treated numerically. Consequently both cases deserve to be considered, with the hope of deducing from the theory of the thin ellipse some results which hold, at least qualitatively, for the thin rectangle.

Because the skin-effect problem in *thin* conductors involves two dimensions of different orders of magnitude, a thickness $2b$, and a lateral dimension $2a$, with

$$b \ll a, \qquad (2)$$

it automatically splits into two almost unrelated problems. At frequencies such that

$$b \ll \delta \qquad (3)$$

the current density is still uniform along the thickness coordinate, so that the only problem is the lateral distribution of the linear current density (A/cm), to be written as $i(x)$, along the width coordinate: $i(x)\mathrm{d}x$ thus designates the total current between x and $x + \mathrm{d}x$. The problem corresponding to the restriction (3) characterizes the *lateral skin effect*, and is the only one treated in this paper, because the depth penetration, occurring at much higher frequencies, is well known.

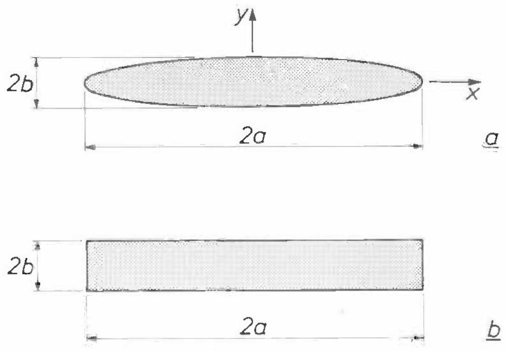


Fig. 1. *a*) Cylindrical conductor with elliptic cross-section. *b*) Cylindrical conductor with rectangular cross-section.

The plan of the article is as follows. A first section is devoted to a qualitative physical discussion of the lateral skin effect in thin conductors and the resulting increase in resistance with frequency. Although all the mathematical derivations are concentrated in the last section, some general remarks of a mathematical nature are necessary at the beginning and are included in the second section. A summary of the results is then presented: complete analytical results are given for the thin ellipse and are all original, to the best of our knowledge; the relatively meagre existing information on the thin rectangle is reviewed and a minor addition is made. In the following section, the impedance (both for the ellipse and the rectangle) is characterized by its poles and zeros, which brings deeper additional information

Prof. Dr. V. Belevitch is the Director of MBLE Laboratoire de Recherches, Brussels.

[1] H. B. G. Casimir and J. Ubbink, The skin effect, Philips tech. Rev. **28**, 271-283, 300-315, 366-381, 1967.

on its behaviour. Finally, an approximate treatment of the high-frequency behaviour of the impedance for the rectangle is given in the Appendix.

Lateral skin effect

As mentioned in the introduction, the skin effect in thin conductors can be separated into a lateral problem and a depth-penetration problem. The lateral problem is, however, of a different nature for conductors with no sharp edges (such as a hollow elliptic cylinder) on the one hand, and for conductors with sharp edges, on the other. Since condition (2) for a thin strip does lead to sharp edges (points of infinite curvature) at $x = \pm a$, both for the thin ellipse and the thin rectangle, the current concentration towards the ends is more important, which results in a different law of increase of resistance. In this section we treat successively (a) thin conductors in general, (b) thin conductors with no sharp edges, (c) thin conductors with sharp edges.

Instead of the three dimensions a, b and δ appearing in (2) and (3), it is convenient to introduce the two dimensionless ratios

$$\frac{\sqrt{(ab)}}{\delta} \quad \text{and} \quad \frac{a}{b}\,; \tag{4}$$

the lateral problem is obtained when b/a is made strictly zero, as the limit of (2). On the other hand, at frequencies where $\sqrt{(ab)}/\delta$ becomes large compared to unity, the lateral skin effect reaches its asymptotic behaviour. This means that the linear density $i(x)$ and the external electromagnetic fields have high-frequency limiting values which are independent of frequency. The corresponding behaviour will be called asymptotic lateral, and thus assumes a frequency range such that

$$b \ll \delta \ll \sqrt{(ab)}, \tag{5}$$

which is of course compatible with (2).

When the frequency increases further, so that (3) no longer holds, the true current density (A/cm^2) begins to vary along the thickness coordinate, but the linear density (A/cm) and the external fields keep their lateral asymptotic values, because of the separation noted earlier of the lateral problem from the depth-penetration problem.

For thin conductors, the pattern of the variation of resistance with frequency is markedly different from that for massive conductors. There are in fact two distinct phases of increase (the lateral effect and the depth effect), obeying different laws and separated by a large frequency interval corresponding to (5), where the lateral effect has already reached its asymptotic state while the depth penetration has not yet come into play. Moreover, the linear density and the external fields reach their asymptotic behaviour in the first phase and remain unaltered during the second phase.

For any thin conductor with no sharp edges, such as a thin hollow elliptic cylinder of moderate eccentricity, the asymptotic linear density is finite at every point, and so is the lateral asymptotic resistance. The law of resistance increase therefore has the form shown in *fig. 2*: the first lateral phase AB is followed by a long stationary interval BC corresponding to (5), where the resistance keeps its constant lateral asymptotic value, the second phase (depth penetration) is CD and the resistance ultimately increases as the square root of the frequency.

Although the problem of the thin hollow elliptic cylinder is rather academic, its lateral asymptotic behaviour (in the phase BC) is so elementary and illuminating that it deserves a short discussion. Since the external magnetic field satisfies Laplace's equation and has no component normal to the ellipse, the lines of force are homofocal ellipses. The linear current density (along the periphery of the ellipse) is given directly by the discontinuity of the tangential component of the magnetic field and is thus inversely proportional to the distance along the normal between two adjacent ellipses of the family. In particular, the linear density is independent of the thickness $2h$ of the cylinder, even if the latter is variable. By contrast, the asymptotic resistance (the constant ordinate of BC of fig. 2) depends on the thickness, because the true density $i/2h$ (A/cm^2) in an element of area $2h\mathrm{d}s$ ($\mathrm{d}s$ = element of length along the boundary) produces a dissipated power

$$\int (i/2h)^2\, 2h\, \mathrm{d}s \tag{6}$$

involving h. In particular, if the thickness is chosen proportional to the lateral asymptotic density, $i/2h$ is a constant in (6), just as at d.c., and the asymptotic value of R/R_0 is 1. Since the resistance is a non-decreasing function of frequency, it must remain constant in the whole lateral phase, for the law of thickness variation adopted. For the hollow elliptic cylinder, the corresponding law defines the conductor as the interior between two homothetic ellipses, which means that the ratios of the major and minor axes are equal. In such a conductor, the increase of linear current density towards the ends of the major axis is exactly compensated by the increased thickness, so that the true current density remains uniform. As a trivial particular case, there is no lateral skin effect in a hollow circular cylinder of constant thickness.

The above discussion, leading to the resistance behaviour of fig. 2, was specifically restricted to thin conductors with no sharp edges, and does not apply to the flat strip, whether rectangular or elliptic. This is because the lateral asymptotic linear current density becomes infinite at the edges ($x = \pm a$) of the strip, so that the dissipation (6), and hence the resistance, is infinite for any thickness law $h(x)$, unless h also becomes infinite at the edges, which is inconsistent with the assumption of a thin conductor. In fact the asymptotic linear density for a strip of any thickness is

$$\frac{i(x)}{i_0} = \frac{2}{\pi \sqrt{(1 - x^2/a^2)}}, \qquad (7)$$

where

$$i_0 = I/2a \qquad (8)$$

is the average linear density, and I the total current. Expression (7) is well known [2] but will be derived again in this article. For a rectangular section, the linear density thus varies from the uniform density (8) at d.c. to the asymptotic density (7) at high frequencies. For the elliptic section, however, it is the true density $i/2h$ which is uniform at d.c. Since the conductor of fig. 1 has the variable thickness $2h(x)$, with

$$h = b \sqrt{(1 - x^2/a^2)}, \qquad (9)$$

the linear density at d.c. is not uniform, but is given by

$$\frac{i(x)}{i_0} = \frac{4}{\pi} \sqrt{(1 - x^2/a^2)}, \qquad (10)$$

so that the variation from d.c. to high frequencies in the elliptic case is much stronger.

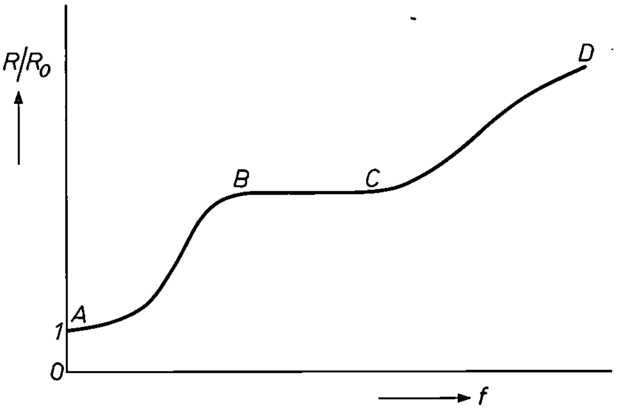


Fig. 2. The resistance R as a function of the frequency f for a thin hollow elliptic cylinder: AB is the lateral phase, BC the lateral asymptotic value, CD the phase of depth penetration. R_0 is the value of R for zero frequency. The curves are only qualitative.

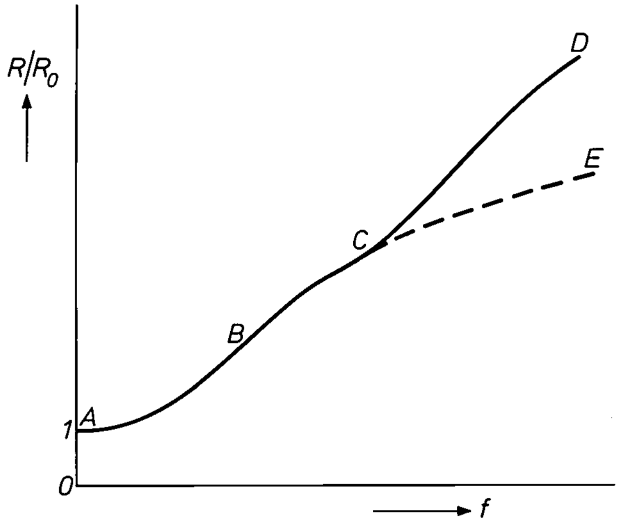


Fig. 3. As fig. 2 but now for a thin flat conductor. Here the curve BCE represents the asymptotic behaviour.

Since the lateral asymptotic resistance of a flat strip is infinite, the law of resistance increase must be of the form qualitatively shown in *fig. 3*. At the end of the lateral phase AB, the resistance reaches its asymptotic behaviour BCE, from which it deviates in accordance with CD when depth penetration comes into play. Curve BCE tends to infinity in accordance with a law still to be discovered, but certainly more slowly than the square root of the frequency since it is dominated by the latter behaviour at the end of phase CD. Finally, the lateral law AB, and its asymptotic behaviour BCE, are different for a thin rectangle and a thin ellipse, whereas the ultimate square-root law (at the end of phase CD) is the same in both cases.

Aspects of the mathematical treatment

In order to avoid certain duplications in the analysis of the elliptic and rectangular strips, we denote the thickness of the strip at the abscissa value x by $2h(x)$, so that h is the constant b in the rectangular case, and the variable (9) in the elliptic case. To permit a coherent normalized frequency to be used, both for the elliptic and the rectangular sections, we introduce the variable

$$k = \frac{j\omega\mu}{4\mu R_0} = j\Omega, \qquad (11)$$

which is proportional to the square of the first parameter (4). We thus have the following notations:

		ellipse	rectangle	
D.c. resistance per unit length	R_0	$\frac{1}{\pi\varrho ab}$	$\frac{1}{4\varrho ab}$	(12)
Normalized frequency	k	$\frac{j\omega\mu\varrho ab}{4}$	$\frac{j\omega\mu\varrho ab}{\pi}$	

The classical skin-effect equation for a massive conductor is

$$\Delta u + \lambda u = 0, \qquad (13)$$

with

$$\lambda = -\varrho\mu j\omega, \qquad (14)$$

[2] H. Kaden, Arch. Elektrotechnik **28**, 818, 1934.

where u stands for any axial component (electric field, current density, or vector potential) and where Δ denotes the 2-dimensional Laplace operator. In the case of the circular cylinder, the magnetic lines of force are concentric circles at all frequencies, and there is no exchange of flux between the conductor and the surrounding dielectric. As a consequence, the internal current distribution can be studied alone, as a solution of (13) with circular symmetry, and the external field is usually disregarded. By contrast, for elliptic and rectangular sections, the magnetic field has a non-zero normal component penetrating into the conductor (except at high frequencies), so that the internal problem is not separable from the external one. Since $\varrho = 0$ in the dielectric, the external problem satisfies (13) with $\lambda = 0$, which is a Laplace equation, and the external and internal solutions are connected by boundary conditions. For thin conductors, only the Laplace equation and the boundary conditions remain, which produces a considerable simplification.

At a large distance D from its centre, a conductor carrying a total current I produces a magnetic field of tangential component $I/2\pi D$, and hence a magnetic flux per unit length proportional to $\ln D$, which tends to infinity with D. Since the external problem depends on the position of the return conductor, which cannot be relegated to infinity, because of the preceding difficulty a coaxial return conductor is generally assumed of large, but finite, radius D, concentric with the go conductor. This makes negligible the proximity effect of the return conductor, but the arbitrary constant D appears in all inductance expressions.

Any skin-effect impedance $Z = R + \mathrm{j}\omega L$ has the nature of the impedance of a (distributed) RL circuit. In particular, it is known from circuit theory that the inductance L is a monotonically decreasing function of the frequency and takes its minimum value L_∞ at infinity. Since there is no internal magnetic field at high frequencies, L_∞ is the only natural definition of the external inductance. Moreover, $Z - \mathrm{j}\omega L_\infty$ is then a minimum-reactance impedance, and the constant D disappears in this difference. In the following, we always evaluate the reduced impedance

$$z = \frac{Z - \mathrm{j}\omega L_\infty}{R_0}, \tag{15}$$

where R_0 is the d.c. resistance (12). The external inductance L_∞ is the one related to the capacitance C per unit length of the pair of conductors by $CL_\infty = 1/c^2$, where c is the velocity of light. For a thin strip (ellipse or rectangle) we have:

$$L_\infty = \frac{\mu}{4\pi} \ln \frac{2D}{a}. \tag{16}$$

Summary of results

Thin elliptic strip

For the thin elliptic strip, the normalized impedance (15) is

$$z = \frac{J_k(k)}{J_k'(k)}, \tag{17}$$

where $J_k(k)$ is the Bessel function of the first kind of order and argument k given by (11). In the denominator of (17), J' denotes the derivative with respect to the argument (and not to the order), and hence the value of $\mathrm{d}J_k(s)/\mathrm{d}s$ at $s = k$.

In terms of the auxiliary variable

$$u = \arccos x/a, \tag{18}$$

the linear current density is given by any of the three following equivalent expressions:

$$\frac{i}{i_0} = \frac{2}{\pi|\sin u|}\left[1 - \frac{2}{kJ_k'(k)}\,\Sigma\, nJ_{n+k}(k)\cos 2nu\right], \tag{19}$$

$$\frac{i}{i_0} = \frac{4|\sin u|}{\pi J_k'(k)}\left[J_k(k) + 2\,\Sigma\, J_{n+k}(k)\cos 2nu\right], \tag{20}$$

$$\frac{i}{i_0} = \frac{4}{\pi J_k'(k)}\,\Sigma\left[J_{k+n-1}(k) - J_{k+n}(k)\right]\sin(2n-1)u, \tag{21}$$

where all sums are for $n = 1, 2, \ldots \infty$ and where i_0 is (8). For k infinite, (19) reduces to its first term and gives (7), by (18). The series (19) is, however, divergent at the end-points $x = \pm a$ corresponding to $u = 0$ or π. By contrast, (20) and (21) are convergent, and (20) reduces to (10) for $k = 0$.

A continued-fraction expansion of (17) is

$$z = 1 + \cfrac{1}{\cfrac{2}{k} + \cfrac{1}{1 + \cfrac{1}{\cfrac{4}{k} + \ldots}}} \tag{22}$$

where the successive denominators are alternately 1 and $2n/k$ ($n = 1, 2, \ldots$). The impedance (17) is thus represented by the equivalent circuit of *fig. 4*. The approximation of (22) limited to order k^2 is

$$z = 1 + \frac{k}{2} - \frac{k^2}{4}. \tag{23}$$

At high frequencies, the known asymptotic expression [3] of (17) is

$$z = C\,k^{1/3}, \tag{24}$$

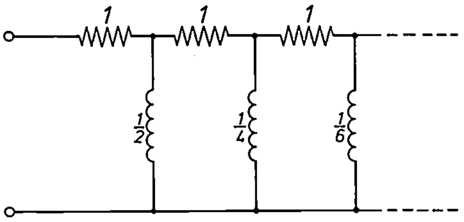


Fig. 4. Equivalent circuit for the impedance of eq. (17) based on the expansion (22).

where the numerical factor C is given by:

$$C = \frac{\Gamma(1/3)}{2^{1/3}\, 3^{1/3}\, \Gamma(2/3)} = 1.088 \ldots . \qquad (25)$$

With the definition (11) of Ω, we thus obtain:

$$z = (0.942 \ldots + \mathrm{j}\, 0.544 \ldots)\Omega^{1/3}. \qquad (26)$$

The real and imaginary parts of the function (17) are shown in *fig. 5* and compared with the tangents resulting from (23) and with the asymptotic expression (26). The numerical computation of (17) was based on the expansion (22) with truncations corresponding to 20 or 30 *RL* sections in the equivalent circuit of fig. 4; this produced no significant difference in the results, in the range $|k| < 10$.

Thin rectangular strip

In comparison with the full analytical results just summarized for the thin ellipse, very little is known for the thin rectangle. From a numerical study of the integral equation for the linear current density, V. Belevitch *et al.* [4] have obtained low-frequency approximations for the impedance. On the other hand, a purely numerical treatment of the problem (by different methods) has led P. Silvester [5] and C. Beccari and C. Ronca [6] to a resistance law confirming the earlier measurements of A. E. Kennelly and H. A. Affel [7]. From all these results it appears that the relative resistance increase Re $(z-1)$ for the ellipse is about twice that for the rectangle. The only new result obtained in this article is an analytic expression of the rectangle resistance as the ratio of two infinite determinants:

$$z = \frac{\Delta_{11}}{\Delta}, \qquad (27)$$

where Δ is the determinant of the symmetric matrix

$$\left[\begin{array}{c:cccc} 1 & -\frac{2}{1\times3} & -\frac{2}{3\times5} & -\frac{2}{5\times7} & \cdots \\ \hdashline -\frac{2}{1\times3} & 2\left(1+\frac{1}{k}-\frac{1}{3\times5}\right) & -2\left(\frac{1}{1\times3}+\frac{1}{5\times7}\right) & -2\left(\frac{1}{3\times5}+\frac{1}{7\times9}\right) & \cdots \\ -\frac{2}{3\times5} & -2\left(\frac{1}{1\times3}+\frac{1}{5\times7}\right) & 2\left(1+\frac{2}{k}-\frac{1}{7\times9}\right) & -2\left(\frac{1}{1\times3}+\frac{1}{9\times11}\right) & \cdots \\ -\frac{2}{5\times7} & -2\left(\frac{1}{3\times5}+\frac{1}{7\times9}\right) & -2\left(\frac{1}{1\times3}+\frac{1}{9\times11}\right) & 2\left(1+\frac{3}{k}-\frac{1}{11\times13}\right) & \cdots \\ \cdots & & & & \end{array}\right], \qquad (28)$$

and Δ_{11} the principal minor separated by dotted lines. In the matrix elements, the first term is constant on a parallel to the main diagonal and the second on a parallel to the second diagonal. Results (27)-(28) are discussed further in the next section and in the Appendix.

Poles and zeros

An impedance $z(k)$ as defined by (15) satisfies

$$z(0) = 1, \qquad (29)$$

and is an *RL* impedance, so that its zeros $k_1, k_2, \ldots$ and its poles $k_1', k_2', \ldots$ are negative real and separate each other:

$$0 < -k_1 < -k_1' < -k_2 < -k_2' < \ldots . \qquad (30)$$

The present section is devoted to an analytical and numerical study of the distribution of the poles and zeros, both for the ellipse and for the rectangle, with the idea of obtaining additional information on the impedance behaviour of the rectangle and, more particularly, on its asymptotic behaviour.

Physically, poles and zeros characterize transient modes of decay (with different boundary conditions)

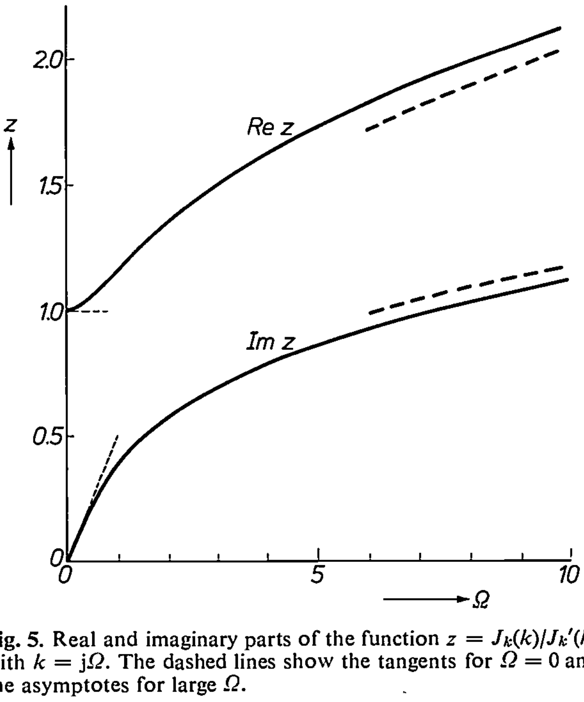


Fig. 5. Real and imaginary parts of the function $z = J_k(k)/J_k'(k)$ with $k = \mathrm{j}\Omega$. The dashed lines show the tangents for $\Omega = 0$ and the asymptotes for large Ω.

[3] M. Abramowitz and I. A. Stegun, Handbook of mathematical functions, Nat. Bur. Stand., Washington 1964, p. 368.

[4] V. Belevitch, P. Guéret and J. C. Liénard, Rev. HF 5, 109, 1962.

[5] P. Silvester, Proc. IEEE 54, 1147, 1966.

[6] C. Beccari and C. Ronca, Elettrotecnica 56, 607, 1969.

[7] A. E. Kennelly and H. A. Affel, Proc. I.R.E. 4, 523, 1916.

and correspond to free solutions of the skin-effect equation, hence to eigenvalues λ of the Laplace operator in (13). The asymptotic distribution of these eigenvalues has been studied in mathematical physics in connection with various problems of statistical thermodynamics; in particular, the nth eigenvalue λ_n of (13), for a two-dimensional domain of area S, and for quite general boundary conditions, is known [8] to be asymptotic to $4\pi n/S$. From (14), (11) and $R_0 = 1/\varrho S$, we thus obtain $-n$ as the asymptotic value of the nth pole and zero, which means:

$$k_n = -n - \alpha_n; k_n' = -n - \beta_n\, (n = 1,2,\dots\infty), \quad (31)$$

with α_n/n and β_n/n tending to zero for large n. Moreover, the alternation (30) of poles and zeros restricts the deviations of (31) to

$$0 < \beta_n - \alpha_n < 1. \quad (32)$$

Although the asymptotic distribution (31) is independent of the shape of the section so that it also holds for a circular cross-section of radius a, not all natural frequencies are excited by the forced current in this case (by circular symmetry), so that only a small subset of the eigenvalues (corresponding to a one-dimensional problem where λ_n is asymptotic to $n^2\pi^2/a^2$) appear as poles and zeros of the impedance.

Since α_n/n and β_n/n certainly tend to zero if α_n and β_n tend to constant values, it is interesting to discuss the case where α_n and β_n are rigorously constant in (31). The function satisfying (29) is then

$$\frac{\Gamma(\alpha + 1)\, \Gamma(k + \beta + 1)}{\Gamma(\beta + 1)\, \Gamma(k + \alpha + 1)} \quad (33)$$

and is asymptotic to

$$\frac{\Gamma(\alpha + 1)}{\Gamma(\beta + 1)}\, k^{\beta - \alpha} \quad (34)$$

for large $|k|$, by Stirling's approximation for the gamma function.

If α_n and β_n are not constant but tend sufficiently quickly to constant values α and β, the asymptotic expression of the impedance is still of the form $Ck^{\beta-\alpha}$, but with a coefficient C different from that of (34), because the latter was imposed on (33) by condition (29). The asymptotic law may, however, become quite different when the variation of α_n or β_n is very slow, as shown by the example of the logarithmic derivative $\Psi(k) = \Gamma'(k)/\Gamma(k)$ of the gamma function. For the function $\Psi(k + 1)$ we have $\beta_n = 0$, whereas α_n tends to zero [9] as $-1/\ln n$. Although (34) gives the value 1 for $\beta - \alpha = 0$, the Ψ function tends to infinity as $\ln k$.

For the ellipse impedance (17), it is known [10] that the nth zero of $J_k(k)$ is asymptotic to $-n + \frac{1}{6}$. Since the exponent $\beta - \alpha$ of (34) is known to be $\frac{1}{3}$ in (24), one thus expects the nth pole to be asymptotic to $-n - \frac{1}{6}$. This is confirmed by *Table I* based on a numerical computation for the equivalent circuit of fig. 4 with 20 and 30 sections.

Table I. Zeros (k_n) and poles (k_n') of the function $J_k(k)/J_k'(k)$.

n	k_n	k_n'
1	−0.83752	−1.11712
2	−1.83490	−2.13294
3	−2.83422	−3.14014
4	−3.83393	−4.14441
5	−4.83377	−5.14728
6	−5.83368	−6.14937

Table II. Zeros (k_n) and poles (k_n') of the impedance of a thin rectangular conductor.

n	k_n	k_n'
1	− 0.87914	− 0.9716
2	− 2.06221	− 2.1591
3	− 3.10107	− 3.2156
4	− 4.12081	− 4.2496
5	− 5.13360	− 5.2742
6	− 6.14298	− 6.2935
7	− 7.1504	− 7.3093
8	− 8.1565	− 8.3227
9	− 9.1618	− 9.3341
10	−10.1664	−10.344
11	−11.1705	−11.352
12	−12.1743	−12.358
13	−13.1777	−13.360
14	−14.1809	
15	−15.1838	

For the rectangle, the zeros and poles have been computed from (27)-(28) on matrices truncated at orders 30 and 40, and are given in *Table II* which confirms the asymptotic behaviour (31). This behaviour was also apparent in Silvester's results [5] in spite of differences in normalization in (11) and (15): Silvester's rough tabulation gives the zeros (but not the poles) of $z + k \ln 2$ in terms of the frequency variable $4k$, but is coherent with our results.

Although the natural frequencies have similar asymptotic distributions for the rectangle and the ellipse, the deviations α_n and β_n show a markedly different small-scale behaviour, as it appears in *fig. 6*. In contrast with the rapid convergence to the asymptotes for the ellipse, the deviations show a logarithmic drift for the rect-

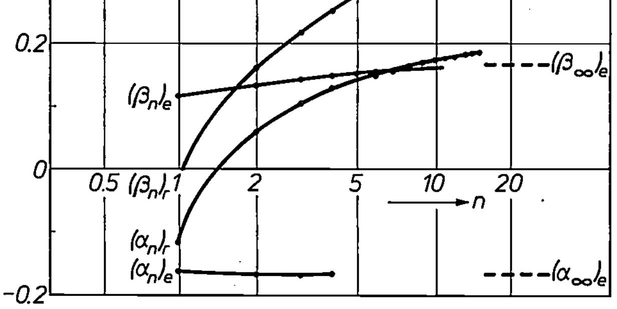


Fig. 6. $\alpha_n = k_n - n$ and $\beta_n = k_n' - n$. the deviations from n for the zeros and poles of the normalized impedance of a conductor with elliptical and rectangular cross-section (subscripts e and r). The dashed lines give the asymptotic values $(\alpha_\infty)_e$ and $(\beta_\infty)_e$.

angle, with an increasing difference $\beta_n - \alpha_n$. Since (32) prevents such an increase from continuing indefinitely, we are far from having reached the asymptotic behaviour of the deviations. Since, however, a continuation of the logarithmic drifts with the slopes resulting from fig. 6 is compatible with (32) up to about $n = 10^8$, there is little hope of obtaining the true asymptotic behaviour by numerical computations.

Our success in obtaining the asymptotic expression (24) of the ellipse impedance is due to the existence of the closed-form expression (17) and to the availability of relatively advanced Bessel-function data. If it were possible to establish (24) directly, without using the closed form (17), a similar approach might succeed for the rectangle where such a form is not available. Such a direct method is described in the next paragraphs for the ellipse. For the rectangle, the relevant mathematics are much more complicated and the treatment is given in the Appendix.

The asymptotic impedance of the ellipse will now be obtained by simple physical considerations of the network of fig. 4, which are, of course, equivalent to mathematical considerations of the corresponding continued fraction (22). A finite approximation of degree n of the network is obtained either by short-circuiting the $(n+1)$th shunt inductance or by opening the $(n+1)$th series resistance, and the corresponding impedances will be called z_S and z_0. At high frequencies, the network reduces to n unit resistances in series in the first case, so that the approximate impedance is

$$z_S = n. \tag{35}$$

In the second case, however, the impedance of the last inductance dominates the last resistance at high frequencies, and the resistance can be neglected; the last inductance thus combines in parallel with the preceding one, and the reasoning applies again. Ultimately, the network reduces to the parallel combination of the first n inductances. Since the total susceptance is $2(1+2+3+\ldots+n)$, which is approximately n^2 for large n, we obtain the impedance

$$z_0 = k/n^2. \tag{36}$$

In classical network and line theory, the input impedances z_0 and z_S of a dissipative 2-port opened or shorted at its output converge to a common value z (the characteristic impedance) when the network attenuation, or the line length, becomes infinite. By contrast, the divergent behaviour of (35)-(36) is due to the essential singularity at infinity of the function (17), resulting from its asymptotic behaviour (24). Although all three impedances z_0, z_S and z thus diverge for large $|k|$, there is some hope of obtaining the asymptotic expression of z by imposing a common asymptotic behaviour on z_0 and z_S. As first attempt, if $z_0/z_S = 1$ is forced into (35)-(36), the resulting constraint

$$n^3/k = 1 \tag{37}$$

imposes the common value $k^{1/3}$ on z_0 and z_S, so that the asymptotic law (24) is confirmed, except for a small difference in the coefficient C, whose correct value (25) is replaced by 1. The discrepancy is due to the fact that the principal values (35)-(36) of z_S and z_0 have been computed by making k large for a fixed value of n, without considering the constraint (37) which was only found *a posteriori*, and as a first approximation. This suggests that an iterative process might lead to an improvement both of the constraint (37) and of the resulting value of the coefficient C of (24). The mathematical justification of this process is based on the inequalities $z_S < z < z_0$, holding for any positive n and k because of potentiometric effects, which impose a common asymptotic behaviour on all three impedances if z_S/z_0 is forced to tend to 1. Owing to the divergent values (35)-(36) of z_0 and z_S for large k, fixed n, the ratio z_S/z_0 can only tend to 1 if n and k tend simultaneously to infinity, in accordance with some (as yet unknown) constraint. Because the constraint is unknown, it can only be reached by successive approximations, which lead to improved estimates for the asymptotic expression of z because the margins resulting from the potentiometric inequalities are decreased at every step. Although the process has not been proved convergent, the margins obtained in one or two steps are already sufficiently small for all practical purposes.

The second approximation replacing (35)-(36)-(37) is obtained as follows. When a unit current is injected in the network of impedance (35), the input voltage is n and the voltage at the ith mode is proportional to $n - i$. The total magnetic energy in the shunt inductances is thus

$$\frac{1}{2}\sum_{i=1}^{n} 2i(n-i)^2 \approx \frac{n^4}{12},$$

and must be equated to the energy $v^2/2L$ in an equivalent inductance L shunting the resistance (35). Since $v = n$, we obtain $L = 6/n^2$, and (35) is replaced more accurately by

$$z_S = \frac{n}{1+n^3/(6k)}\,. \tag{38}$$

By a similar reasoning we are led to evaluate the total dissipation in the network of the initially reactive impedance (36) and to represent it as a series resistance, which is found to be $8n/15$, so that (36) is replaced by

$$z_0 = \frac{k}{n^2}\left(1+\frac{8n^3}{15k}\right). \tag{39}$$

It is not legitimate to equate (38) and (39) rigorously, for the re-

[8] R. Courant and D. Hilbert, Methods of mathematical physics, Interscience, New York 1953, p. 432.

[9] See Abramowitz and Stegun [3], p. 259.

[10] E. Jahnke, F. Emde and F. Lösch, Tables of higher functions, 6th edn., Teubner, Stuttgart 1960, p. 154.

sulting second-degree equation contains terms in $(n^3/k)^2$ that are still neglected in (38)-(39). One must therefore linearize the equation around the first-order approximation (37), by replacing n^3/k by 1 in the correction factors. This reduces (38) to $6n/7$ and (39) to $23k/15n^2$. By equating the last two expressions we obtain for both the value (24) with $C = 1.044\ldots$.

Mathematical derivation

The mathematical formulation of the lateral skin-effect problem for a thin conductor will now be derived simultaneously for the elliptic section (fig. 1*a*) and for the rectangular section (fig. 1*b*) with the common notation $2h$ for the thickness, h being given by (9) in the first case and equal to the constant b in the second. The linear current density is the discontinuity of the tangential component H_x of the magnetic field. By symmetry we have for a right-handed coordinate system:

$$i = -2H_x\Big|_{y=+0}. \tag{40}$$

On the other hand, the true current density (A/cm²) is

$$\frac{i}{2h} = \varrho\, E_z\,. \tag{41}$$

Finally, Lenz's law yields:

$$\frac{\partial E_z}{\partial x} = \mu \mathrm{j}\omega H_y\,. \tag{42}$$

It is convenient to introduce the vector potential A which has only a z-component. We then have:

$$H_x = \frac{1}{\mu}\frac{\partial A_z}{\partial y}\,;\; H_y = -\frac{1}{\mu}\frac{\partial A_z}{\partial x}\,, \tag{43}$$

whereas the voltage drop ZI along the conductor is

$$-\frac{\partial V}{\partial z} = E_z + \mathrm{j}\omega A_z\,. \tag{44}$$

In (44), E_z is expressed in terms of A_z by (41), (40) and the first equation (43); this yields:

$$ZI = \mathrm{j}\omega A_z - \frac{1}{\mu\varrho h}\frac{\partial A_z}{\partial y}\Big|_{y=+0}. \tag{45}$$

On the other hand, by elimination from (40)-(41)-(42), we obtain the Biot boundary condition [11]:

$$\frac{\partial}{\partial x}\left(\frac{H_x}{\varrho h}\Big|_{y=+0}\right) = -\mu\mathrm{j}\omega H_y\,. \tag{46}$$

When the magnetic fields are eliminated from (46) by (43), the resulting relation shows that the derivative of (45) with respect to x is zero. Biot's condition is thus equivalent to saying that the impedance computed by (45) is independent of x. The form (45) of the boundary condition is to be preferred to (46), since it yields the impedance without additional effort. Finally, the problem amounts to solving the Laplace equation for A_z, with the condition (45) on the conductor (i.e. for $y = 0$, $|x| < a$), and the prescription of the value

$$A_z = -\frac{I\mu}{2\pi}\ln\frac{\sqrt{(x^2+y^2)}}{D} \tag{47}$$

at large distance, corresponding to the vector potential of a filament of current I at the origin, with a coaxial return of large radius D.

We introduce the conformal transformation

$$x + \mathrm{j}\,y = a\cos(u + \mathrm{j}v), \tag{48}$$

yielding

$$x = a\cosh v\cos u;\;\; y = a\sinh v\sin u. \tag{49}$$

The transformation is one-to-one with the restrictions

$$-\pi \leqslant u \leqslant \pi;\;\; v \geqslant 0.$$

In the (x,y)-plane, the curves of constant v are homofocal ellipses (*fig. 7*); the segment $y = 0$, $|x| \leqslant a$ is the infinitely flat ellipse $v = 0$, and v increases outwards to infinity. The curves of constant u are the hyperbolae of fig. 7 but there is a cut along the segment $v = 0$, so that u is positive in the upper half-plane Re $y > 0$ and negative in the lower half-plane. The semi-infinite segment $y = 0$, $x \geqslant a$ corresponds to $u = 0$, whereas the segment $y = 0$, $x \leqslant -a$ corresponds to $u = \pm\pi$.

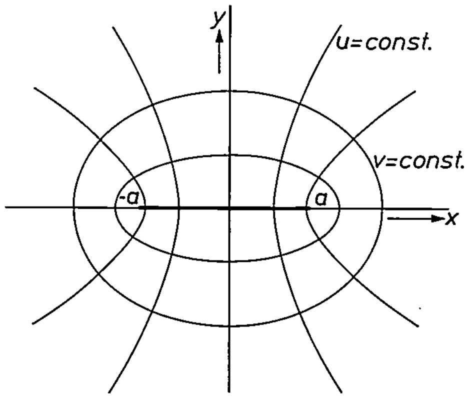


Fig. 7. The conformal representation $x + \mathrm{j}y = a\cos(u + \mathrm{j}v)$.

The expression of A_z is of the form

$$A_z = \frac{I\mu}{2\pi}\left[\ln\frac{2D}{a} - v + \Sigma A_n\,\mathrm{e}^{-2nv}\cos 2nu\right], \tag{50}$$

where the first two terms give the principal value (47), because (48) yields

$$\frac{\sqrt{(x^2+y^2)}}{D} = \frac{a\cosh v}{D} \approx \frac{a\mathrm{e}^v}{2D}\,,$$

whereas the sum (extending from $n = 1$ to ∞) with undetermined coefficients A_n is the general harmonic function, vanishing at infinity, and having the appropriate quadrantal symmetry.

For $y = +0$, and hence $v = 0$, $u > 0$, then by (49):

$$\left.\frac{\partial A_z}{\partial y}\right|_{y=+0} = \frac{1}{a|\sin u|}\left.\frac{\partial A_z}{\partial v}\right|_{v=0},$$

so that (45) becomes:

$$Z = \frac{\mathrm{j}\omega\mu}{2\pi}\left(\ln\frac{2D}{a} + \Sigma\, A_n \cos 2nu\right) + \\ + \frac{1}{2\pi\varrho ha|\sin u|}\left(1 + 2\,\Sigma\, n\, A_n \cos 2nu\right). \tag{51}$$

The coefficients A_n must now be determined so that (51) takes a constant value for all u $(-\pi \leqslant u \leqslant \pi)$.

For the elliptic section, the first relation (49) reduces to (18) for $v = 0$, and (9) becomes

$$h = b\,|\sin u|, \tag{52}$$

so that the denominator in (51) simplifies to

$$h|\sin u| = b \sin^2 u = \frac{b}{2}(1 - \cos 2u). \tag{53}$$

With the notations (15)-(16) and (12), and with the substitution

$$B_n = -\,k\, A_n, \tag{54}$$

(51) multiplied by (53) becomes

$$(z + 2\,\Sigma\, B_n \cos 2nu)\,(1 - \cos 2u) + \\ + \frac{2}{k}\,\Sigma\, n\, B_n \cos 2nu = 1\,. \tag{55}$$

Replacing the product of cosines occurring in the left-hand side of (55) by cosines of sums and differences, one obtains a Fourier series, and its identification with 1 yields the infinite system

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0\ldots \\ -1 & 2(1+1/k) & -1 & 0 & 0\ldots \\ 0 & -1 & 2(1+2/k) & -1 & 0\ldots \\ 0 & 0 & -1 & 2(1+3/k) & -1\ldots \\ \ldots & & & & \end{bmatrix} \begin{bmatrix} z \\ B_1 \\ B_2 \\ B_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \tag{56}$$

of linear equations in z and the unknown coefficients B_n.

Disregarding temporarily the first two equations contained in (56), one obtains the three-term recurrence relation

$$B_{n-1} + B_{n+1} = 2(1 + n/k)B_n \quad (n = 2, 3, \ldots), \tag{57}$$

which is very similar to the recurrence relation

$$J_{v-1}(s) + J_{v+1}(s) = \frac{2v}{s} J_v(s) \tag{58}$$

for Bessel functions. For $v = n + k$, $s = k$, (58) shows that

$$B_n = C\, J_{n+k}(k) \tag{59}$$

satisfies (57), with C an arbitrary constant. In fact (59) is the solution of the second-order difference equation (57) in our case, because the other linearly independent solution (involving a Bessel function of the second kind) is excluded on physical grounds since it makes all coefficients B_n infinite at d.c. Since (57) holds down to $n = 2$ and thus involves B_1, (59) holds down to $n = 1$ and only two unknowns remain: the common factor C of (59), and the impedance z. These are determined by the first two equations (56) which have been disregarded. By solving these equations, and making use of the known expression for the derivative of a Bessel function:

$$J_v'(s) = \frac{v}{s} J_v(s) - J_{v+1}(s)\,, \tag{60}$$

(used for $v = s = k$), we obtain

$$C = \frac{1}{J_k'(k)}, \tag{61}$$

and (17).

From the known expression (50) for A_z, where A_n is deduced from (54), (59) and (61), we can compute H_x by (43) and i by (40). This gives (19). The other form (20) is obtained when i is computed by (41) with the value of E_z deduced from (44) where $-\,\partial V/\partial z$ is ZI. This completes the proof of all the basic results, (17) to (21), for the ellipse.

The expansion (22) of (17) can be deduced from the three-term recurrence relations (57). In the equivalent network of fig. 4, they correspond to the Kirchhoff relations between the currents in branches incident to a common node, and the following electrical proof is therefore equivalent to a mathematical discussion of (57). Consider the ladder network of *fig. 8*, where the series admittances are denoted $Y_1, Y_2, \ldots$ and the shunt

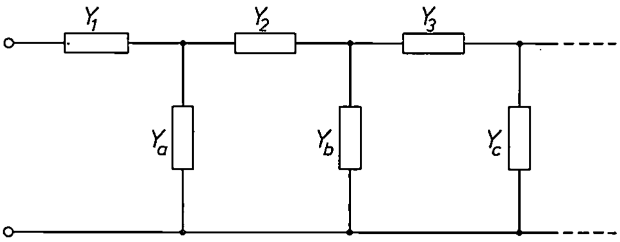


Fig. 8. A ladder network.

[11] M. Biot, Ann. Soc. Sci. Brux. **51**, 94, 1931.

admittances $Y_a, Y_b, \ldots$. Denote by V_i the node potentials with respect to ground, as indicated. If a unit current is injected at the input, elementary node analysis yields the linear system

$$\begin{bmatrix} Y_1 & -Y_1 & 0 & 0 & \ldots \\ -Y_1 & Y_1+Y_a+Y_2 & -Y_2 & 0 & \ldots \\ 0 & -Y_2 & Y_2+Y_b+Y_3 & -Y_3 & \ldots \\ \ldots & & & & \end{bmatrix} \begin{bmatrix} V_0 \\ V_1 \\ V_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \tag{62}$$

for the node voltages, and the solution V_0 is the input impedance. From a comparison of (56) and (62), it is found that the solution z of (56) is the input impedance of the ladder network of fig. 4 where the elements are normalized (k is taken as the complex frequency). Since the input impedance of fig. 4 is (22), we have indirectly obtained the continued-fraction expansion of the function (17).

For the rectangular section, h is constant in the denominator of (51). After multiplication by the known Fourier series

$$|\sin u| = \frac{4}{\pi}\left(\frac{1}{2} - \frac{\cos 2u}{1 \times 3} - \frac{\cos 4u}{3 \times 5} - \ldots\right) \tag{63}$$

and substitution of (54), (51) becomes

$$(z + 2\Sigma B_n \cos 2nu)\left(1 - \frac{2}{1 \times 3}\cos 2u - \frac{2}{3 \times 5}\cos 4u - \right.$$
$$\left. - \ldots\right) + \frac{2}{k}\, \Sigma\, n B_n \cos 2nu = 1\,. \tag{64}$$

The linear system resulting from (64) is analogous to (56) except that the matrix is now (28). This establishes (27).

Appendix: High-frequency impedance of the rectangular strip

At the end of the section on poles and zeros, we succeeded in obtaining a good approximation of the asymptotic impedance of the ellipse, without using its closed-form expression (17). The result was (24) but, instead of the correct coefficient 1.088 . . . of (25), we obtained 1 and 1.044 . . . by successive approximations. In this Appendix, we apply the same method to the rectangle. The analysis is, however, much more difficult, essentially because the matrix (28) is a full matrix whereas the one (56) for the ellipse was tridiagonal, so that only the first approximation will now be worked out. Consequently, the accuracy of the result cannot be assessed. Also, Silvester's numerical results cover too narrow a frequency range to provide an adequate verification. In spite of its limited significance, our result is the only one presently available; it can be improved by further research, and has been obtained by a method having its own mathematical interest.

As for the ellipse, the following analysis is based on the equivalent circuit, but could be translated into purely mathematical terms. Because the matrix (28) is not tridiagonal, the circuit is not a ladder network, but a general RL network with an infinite number of nodes. Since, however, terms in k^{-1} only occur on the diagonal, inductances only connect each node to ground and the interconnections between non-ground nodes are purely resistive. Finally, for $k = \infty$, the sum of all elements in any row of (28) is zero, on account of (63) for $u = 0$. This means that the direct conductance from any node to ground is zero, so that the branch connecting node n to ground is the inductance of value $1/2n$ alone, as in fig. 4.

When the network is truncated at n nodes, by opening all resistances leading to further nodes, what remains is an n-node resistive network with an inductance from each node to ground, but with no resistive path from the first node to ground. The impedance from node 0 to ground is thus infinite at high frequency and its principal value is due to the inductances alone, so that all resistances can equally be short-circuited. The resulting impedance is the parallel combination of the first n inductances, and this yields (36), as in the elliptic case.

In the complementary method of truncation at n nodes, all further nodes are shorted to ground. The resistances leading to further nodes then produce a resistive path from node 0 to ground, so that the impedance is resistive at high frequency, as in (35), and can be evaluated by open-circuiting all the inductances. This amounts to computing (27) for the matrix (28) truncated at order n and for $k = \infty$. Note that the sum of the elements in each row is no longer zero, owing to the truncation, so that the matrix is non-singular. Finally, the evaluation of the truncated impedance is equivalent to solving (64) for z with

$$k = \infty, B_{n+1} = B_{n+2} = \ldots = 0\,. \tag{65}$$

Note that the truncation of the Fourier series of coefficients B_i corresponds to the shorting of the higher nodes, whereas the Fourier series (63) appearing in (64) is not truncated, for this corresponds to the preserved resistive connections to higher nodes. By (63), the form of (64) modified by (65) is thus

$$z_s + 2 \sum_{i=1}^{n} B_i \cos 2iu = \frac{2}{\pi|\sin u|}\,. \tag{66}$$

In the original (non-truncated) form (64), originating from the boundary condition (45), the coefficients B_i and the impedance z were determined by making the latter independent of u. Because of the truncation, this becomes impossible rigorously, and z_s can only be made constant at n points (the number of undetermined coefficients) and, owing to the quadrantal symmetry of (66), these may be all chosen in one quadrant, say the first. The quadrantal symmetry is preserved by choosing n equidistant points with intervals $\pi/2n$ from each other and half that interval from the ends. The interpolation points are thus

$$u = \frac{\pi}{4n}, \frac{3\pi}{4n}, \frac{5\pi}{4n}, \ldots, \frac{(2n-1)\pi}{4n}.$$

For this classical trigonometric interpolation, the "d.c. component" z_s of (66) is simply the mean value of the second term of (66) at the interpolation points:

$$z_s = \frac{2}{n\pi} \sum_{i=0}^{n-1} \frac{1}{\sin(2i+1)\pi/4n}\,. \tag{67}$$

Since the above derivation of (67) is rather indirect, we check that the same derivation yields the known value (35) in the elliptic case. Equation (55) with the reductions (65) then yields

$$z_s + 2 \sum_{i=1}^{n} B_i \cos 2iu = \frac{1}{1 - \cos 2u} = \frac{1}{2 \sin^2 u},$$

and (67) is replaced by

$$z_s = \frac{1}{2n} \sum_{i=0}^{n-1} \frac{1}{\sin^2 (2i+1)\pi/4n}, \tag{68}$$

which is indeed (35), by a known identity [12].

No closed-form expression is available for (67), but, for large n, it can be approximately evaluated with the help of the Euler-MacLaurin summation formula applied to the function

$$\frac{1}{\sin x} - \frac{1}{x}$$

so as to extract the singularity at $x = 0$. In this way, (67) is approximately obtained as

$$z_s = \frac{4}{\pi^2} \ln \frac{16n\gamma}{\pi}, \tag{69}$$

where $\gamma = 1.781\ldots$ is Euler's constant.

By identifying (36) and (69), and eliminating n, one obtains:

$$k = z\left(\frac{\pi}{16\gamma}\right)^2 e^{\pi^2 z/2}, \tag{70}$$

a relation defining implicitly z as a function of k. By separating the real and imaginary parts of the logarithm of (70) one establishes that the imaginary part of z remains finite for $k = j\Omega$, whereas its real part r tends to infinity. From the modulus of (70), one then deduces:

$$\Omega = r\left(\frac{\pi}{16\gamma}\right)^2 e^{\pi^2 r/2}. \tag{71}$$

The author is grateful to Prof. C. Manneback who suggested the study of the thin ellipse in the system of coordinates (48), to Prof. C. J. Bouwkamp and to A. C. Vliegenthart for help in obtaining expression (69), to C. Wellekens and J. Praet for computer results, and to Dr. Y. Genin for many useful discussions.

[12] I. S. Gradshteyn and I. M. Ryzhik, Table of integrals, series, and products, 4th edn., Academic Press, New York 1965, section 1.382, formula (1) with $x = 0$.

Deutsche Luftbild K.G. (Freigabe-Nr. D.L.A. Hbg: 2859/69).

Philips Forschungslaboratorium Hamburg

Magnetoacoustic effects in bismuth

K. Walther

Introduction

Bismuth is a semimetal which has long been of interest in scientific investigation because of its exceptional electronic properties. In general, solids can be divided into three classes: metals, semimetals and semiconductors. The energy states of electrons in solids are grouped into various permitted energy bands, which are separated by forbidden energy gaps. The occupation of these energy bands is described by Fermi-Dirac statistics. At low temperatures all states up to a certain energy E_F, called the Fermi energy, are occupied, and states with higher energies are empty. The motion of charge carriers in solids can be described in terms of an effective mass m^*, which is generally different from the mass m_0 of the free electron because of interaction with the lattice atoms.

In a *metal* the number of valence electrons is such that partially filled energy bands exist. Electrical conduction is possible even at low temperatures, since no energy gap separates the occupied and empty states within an energy band. The concentration of charge carriers taking part in electrical conduction is comparable to the concentration of lattice atoms and is about 10^{22} to 10^{23} cm^{-3}.

The situation is quite different in a *semiconductor*, where a finite energy gap exists between valence and conduction bands. In most cases, the valence band is filled completely at low temperatures, and the conduction band is empty. Electrical conduction is possible only at higher temperatures, either by thermally exciting electrons from the valence band into the conduction band (intrinsic conduction) or by doping the semiconductor, thus creating a certain number of electrons and holes in the conduction and valence bands (extrinsic conduction). In semiconductors the concentration of charge carriers contributing to electrical conduction (10^{14} to 10^{19} cm^{-3}) is much smaller than the concentration of lattice atoms.

A *semimetal* such as bismuth is characterized by the fact that the top of the valence band is located at an energy slightly higher than the bottom of the conduction band (*fig. 1*). The position of the Fermi level, separating occupied and empty states, is such that a small concentration p of unoccupied states, which act as holes in the conduction process, is present in the valence band, and a corresponding electron concentration n exists in the conduction band. The concentration of charge carriers in semimetals is much smaller than the concentration of lattice atoms. Electrical conduction is possible at all temperatures since there is no energy gap between occupied and empty states in a band.

The exceptional electronic properties of bismuth, which have made this semimetal an interesting subject of research, are most pronounced at temperatures T in the liquid helium range, where the relaxation times τ, describing scattering of carriers by the lattice, become very long ($\tau \approx 10^{-9}$s). The effective masses in bismuth are very small ($m^*/m_0 \approx 0.01$ for electrons), resulting in very high carrier mobilities $\mu = e\tau/m^*$ (where e is the electronic charge) at low temperature: $\mu \approx 10^8$ cm^2/Vs for electrons. In the undoped state bismuth is a "compensated" semimetal with equal electron and hole concentrations: $n = p \approx 3\times10^{17}$ cm^{-3} at $T = 4.2$ K.

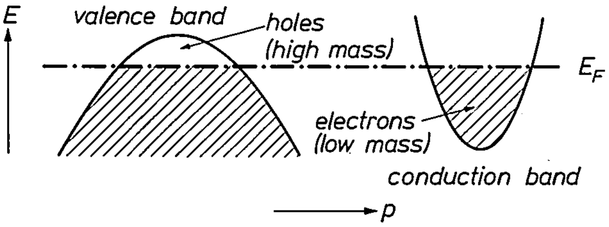


Fig. 1. Energy-band structure for a semimetal such as bismuth. Schematic plot of the relation between energy E and momentum p of the charge carriers for the valence band and the conduction band. E_F Fermi energy. Both bands are filled with electrons up to the Fermi level. The effective mass m^* of the holes in the valence band is high, for the electrons in the conduction band m^* is only a small fraction of the mass m_0 of a free electron.

These electronic properties lead to the unusual galvanomagnetic effects of bismuth, which were discovered at the end of the 19th century. The magnetoresistance in bismuth is several orders of magnitude larger than in other materials, since a transverse magnetic field deflects both electrons and holes to the same side of the sample, thus avoiding the build-up of space charge.

Dr. K. Walther is with Philips Forschungslaboratorium Hamburg GmbH (PFH), Hamburg, Germany.

Owing to the small carrier concentrations and effective masses bismuth became the first material in which oscillatory quantum effects as a function of magnetic field strength were discovered. In 1930 L. Schubnikow and W. J. de Haas [1] observed characteristic oscillations in the magnetoresistance of bismuth. De Haas and P. M. van Alphen [2] found similar oscillations in the magnetic susceptibility. These effects were subsequently interpreted as being due to the quantized motion of charge carriers perpendicular to the magnetic field. Later the oscillatory quantum effects became a powerful tool for investigating the electronic band structure of many solids.

The interaction between ultrasonic waves and charge carriers in bismuth in the presence of a magnetic field was investigated first by D. H. Reneker [3]. An acoustic wave of either longitudinal or transverse polarization creates a mechanical strain ε within the crystal, which varies sinusoidally and periodically shifts the energies of the charge carriers by an amount δE, proportional to the strain: $\delta E = C\varepsilon$. The proportionality factor C is called the "deformation potential" of the appropriate energy band. The situation in metals and extrinsic semiconductors with only one type of charge carrier is in general rather complicated, since the motion of the charge carriers causes electric space-charge fields, which partially screen out the primary force due to the deformation potential. This difficulty is avoided in a compensated semimetal like bismuth, where no space-charge field can develop, because the acoustic wave simultaneously displaces electrons and holes. Magnetoacoustic investigations are therefore a suitable method for determining the unscreened deformation potentials for electrons and holes in bismuth [4] [5]. This is the first subject to be discussed in the present article.

The second theme to be covered in this article is that of the acoustodynamic effects in the presence of a d.c. field, which causes the carriers to drift with a velocity v_d. Interest in this subject has been stimulated by the possibility of technical applications similar to those in a travelling-wave amplifier, where a microwave signal propagating along a slow-wave structure interacts with an electron beam. Analogously, in a solid an acoustic wave gains energy from the drifting carriers, when the drift velocity in the direction of acoustic propagation exceeds the velocity of sound v_s, i.e. an ultrasonic wave can be amplified by applying an electrical drift field. This effect was first demonstrated in the piezoelectric semiconductor CdS by A. R. Hutson, J. H. McFee and D. L. White [6]. An acoustodynamic effect in bismuth was first observed by L. Esaki [7], who found a kink in the current-voltage curve of a bismuth sample in a transverse magnetic field at a drift velocity v_d equal to v_s. This kink effect was interpreted in the following way [8]: under conditions of ultrasonic amplification, $v_d > v_s$, the drifting carriers generate ultrasonic noise in a broad frequency range. This statistically varying acoustic wave in turn drags along an additional component of d.c. current I_{ae} (called the acoustoelectric current), which appears in the current-voltage characteristic of the sample. Direct observation of the ultrasonic noise in bismuth has been reported by the author [9]. The first theoretical investigation of ultrasonic amplification in semimetals was published by W. P. Dumke and R. R. Haering [10], and experimental evidence for ultrasonic amplification in bismuth has been reported by A. M. Toxen and S. Tansal [11]. Subsequently the influence of an electrical drift field on the magnetoacoustic quantum oscillations in bismuth has been studied experimentally [12]. The maximum ultrasonic amplification [13] attainable in bismuth, is limited mainly by the interaction with the ultrasonic noise, which is generated simultaneously.

Band structure and deformation potential in bismuth

The energy-band structure in solids describes the relation between the energy E and the momentum p of the charge carriers, taking into account interactions with the lattice atoms, which are arranged periodically in a crystal lattice. For an electron in free space this relation is simply:

$$E = \frac{p^2}{2\,m_0} = \frac{p_1{}^2 + p_2{}^2 + p_3{}^2}{2\,m_0}, \tag{1}$$

where p_1, p_2 and p_3 are the components of momentum with respect to three orthogonal directions in "momentum space". For charge carriers in a solid the free-electron mass m_0 has to be replaced by an effective mass, which generally is anisotropic in momentum space (components $m_1{}^*$, $m_2{}^*$, $m_3{}^*$), yielding the result:

$$E = \tfrac{1}{2}\left(\frac{p_1{}^2}{m_1{}^*} + \frac{p_2{}^2}{m_2{}^*} + \frac{p_3{}^2}{m_3{}^*}\right). \tag{2}$$

In this equation energy and momentum are referred to the bottom of the conduction band for electrons and to the top of the valence band for holes (see fig. 1). The relation between energy and momentum in eq. (2) yields parabolae for various directions in momentum space (see *fig. 2*), which are weakly curved in a direction p_1 of high mass $m_1{}^*$ and exhibit a strong curvature in a direction p_2 of low mass $m_2{}^*$. The charge carriers change their energy and momentum states under the influence of a periodic sound wave due to the deformation-potential coupling. Such transitions are possible only in the vicinity of the boundary between occupied and empty states, i.e. near the Fermi energy E_F. The quantity of interest for acoustic effects is therefore the surface of constant energy $E = E_F$ in momentum space, which is called the Fermi surface. It can be seen from eq. (2) that this surface is an ellipsoid with a long axis parallel to the large-mass direction p_1 and a short axis along the low-mass direction p_2 (see fig. 2). Such an energy band is referred to as "ellipsoidal parabolic".

The crystal structure of bismuth has trigonal symmetry. The location of the ellipsoidal Fermi surfaces for electrons and holes is described with respect to the following crystallographic axes: x = binary axis, y = bisectrix, z = trigonal axis (*fig. 3*). The Fermi surface of electrons consists of three ellipsoids, labelled by the band indices $l = 1, 2, 3$, which are arranged with trigonal symmetry (120° angular separation) around the z-axis. The large-mass direction of the electron ellipsoids is tilted through a small angle ($\approx 6°$) with respect to the y-axis. The Fermi surface of holes consists of a single ellipsoid of revolution about the z-axis, labelled by the band index $l = 4$. Its large-mass direction is parallel to the z-axis.

Since the principal axes of the constant-energy ellipsoids in bismuth do not coincide with the crystallographic axes, eq. (2) has to be rewritten in a more general form:

$$E = \tfrac{1}{2} \sum_{i,k=1}^{3} \alpha_{ik} p_i p_k \,, \tag{3}$$

for each band $l = 1, \ldots 4$. Here the anisotropy of the reciprocal effective mass is described by introducing the components α_{ik}, which can be arranged in the form of a matrix. The following shape is obtained for the electron ellipsoid $l = 1$:

$$(\alpha_{ik})_{l=1} = \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & \alpha_4 \\ 0 & \alpha_4 & \alpha_3 \end{pmatrix}, \tag{4}$$

The component α_4 accounts for the tilt with respect to the y-axis. The electron bands $l = 2, 3$ are obtained by rotating this ellipsoid through ± 120° about the trigonal axis.

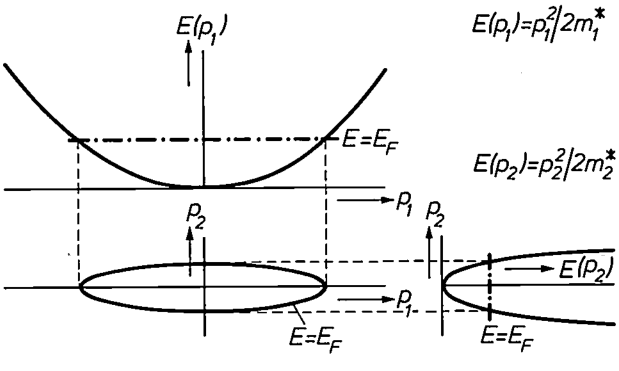


Fig. 2. Ellipsoidal-parabolic energy band. The relations $E(p)$ are parabolic, the surfaces E = const. in momentum-space are ellipsoids. The effective mass m_1* is high, the mass m_2* is low.

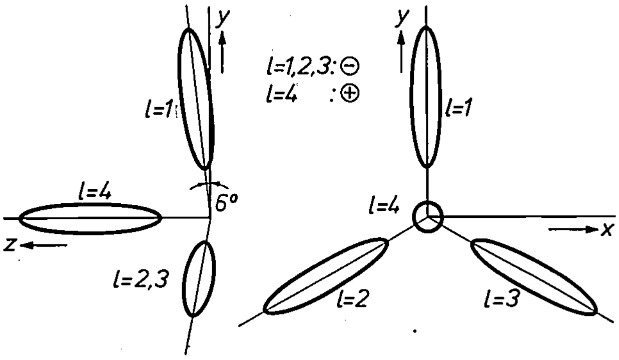


Fig. 3. Location of ellipsoids of constant energy for electrons (l = 1, 2, 3) and holes (l = 4) in bismuth. The crystal structure has trigonal symmetry: x is the binary axis, y the bisectrix and z the trigonal axis.

The matrix for the hole ellipsoid (l = 4) is given by:

$$(\alpha_{ik})_{l=4} = \begin{pmatrix} \beta_1 & 0 & 0 \\ 0 & \beta_1 & 0 \\ 0 & 0 & \beta_2 \end{pmatrix}, \tag{5}$$

where the appearance of two equal components β_1 describes rotational symmetry around the trigonal axis.

The energy change δE of the charge carriers as a function of mechanical strain ε:

$$\delta E = C\varepsilon \tag{6}$$

(C = deformation potential) in general yields an anisotropic relationship for each energy band $l = 1, \ldots 4$. This can be accounted for by introducing an anisotropic deformation potential with components $(C_{ik})_l$ which has the same symmetry as the reciprocal effective mass of the respective band. These components can be obtained from equations (4) and (5) by writing a_i and b_i instead of α_i and β_i. Elasticity theory shows that the mechanical strain in a solid is also an anisotropic quantity containing spatial derivatives $\partial e_i/\partial x_k$ of the displacement components e_i with respect to the

[1] L. Schubnikow and W. J. de Haas, Proc. Sec. Sci. Kon. Akad. Wetensch. Amsterdam **33**, 130, 350, 363, 418 and 433, 1930 (Comm. Phys. Lab. Univ. Leiden **19**, Nos. 207a, 207c, 207d, 210a and 210b).

[2] W. J. de Haas and P. M. van Alphen, Proc. Sec. Sci. Kon. Akad. Wetensch. Amsterdam **33**, 680 and 1106, 1930, and **35**, 454, 1932 (Comm. Phys. Lab./Kamerlingh Onnes Lab. Univ. Leiden **19** and **20**, Nos. 208d, 212a and 220d).

[3] D. H. Reneker, Phys. Rev. **115**, 303, 1959.

[4] S. Inoue and M. Tsuji, J. Phys. Soc. Japan **22**, 1191, 1967.

[5] K. Walther, Phys. Rev. **174**, 782, 1968.

[6] A. R. Hutson, J. H. McFee and D. L. White, Phys. Rev. Letters **7**, 237, 1961.

[7] L. Esaki, Phys. Rev. Letters **8**, 4, 1962.

[8] A. R. Hutson, Phys. Rev. Letters **9**, 296, 1962.

[9] K. Walther, Phys. Rev. Letters **15**, 706, 1965.

[10] W. P. Dumke and R. R. Haering, Phys. Rev. **126**, 1974, 1962.

[11] A. M. Toxen and S. Tansal, Phys. Rev. Letters **10**, 481, 1963.

[12] K. Walther, Phys. Rev. Letters **16**, 642, 1966, and Z. Naturf. **21a**, 1443, 1966.

[13] K. Walther, Solid State Comm. **4**, 341, 1966.

coordinate axes x_k. In a plane acoustic wave of angular frequency ω and wave-vector components q_k the displacement is proportional to

$$e_i \exp\left[\mathrm{i}\left(\sum_{k=1}^{3} q_k x_k - \omega t\right)\right] \text{ and } \partial e_i/\partial x_k \propto e_i q_k .$$

The general anisotropic form replacing eq. (6) therefore reads:

$$\delta E = \sum_{i,k=1}^{3} C_{ik} e_i q_k \equiv (e \cdot \tilde{C} \cdot q) . \qquad (7)$$

As a specific example for the interaction, let us consider a shear wave propagating along the x-axis. Here $q_1 \neq 0$ and $e_1 = 0$ because of the transverse polarization. Coupling is given by the elements C_{21} and C_{31}, which are zero for the electron band $l = 1$ and the hole band $l = 4$ (see eqs. (4) and (5)). Therefore this mode only couples to the electron bands $l = 2$ and 3 (see also fig. 6).

To measure all the components C_{ik} of the deformation potential for electrons and holes in bismuth a number of ultrasonic-absorption experiments with known directions of the wave vector $\boldsymbol{q}$ and the polarization vector e (using longitudinal and transverse modes) is carried out, which yields various linear combinations of the unknown coefficients C_{ik}, from which the latter can be calculated. One difficulty occurring in such experiments is the fact that, in general, contributions from all the energy bands $l = 1, \ldots 4$ enter into the ultrasonic attenuation simultaneously. A suitable method of separating these different contributions is the application of a quantizing magnetic field. This will be explained in the following section.

Quantum oscillations of the magnetoacoustic attenuation in bismuth

The presence of a magnetic field of flux density $\boldsymbol{B}$ drastically modifies the energy-momentum relation for the charge carriers. If the motion is described in semiclassical terms, the Lorentz force $e\boldsymbol{v}\times\boldsymbol{B}$ acting perpendicular to $\boldsymbol{B}$ causes the carriers to move in an orbital plane perpendicular to the magnetic field, while the component $p_{\mathfrak{b}}$ of momentum parallel to $\boldsymbol{B}$ remains unaffected. In momentum space the carriers move along an ellipsoidal curve perpendicular to $\boldsymbol{B}$ on the Fermi ellipsoid $E = E_{\mathrm{F}}$ (*fig. 4*). This orbital motion occurs periodically with an angular frequency equal to the cyclotron frequency

$$\omega_{\mathrm{c}} = \frac{e}{m_{\mathrm{c}}} B . \qquad (8)$$

Here m_{c} is the cyclotron mass of the orbit, which is equal to the geometric mean of the extremal masses on the orbital ellipse. In general m_{c} changes as the direction of $\boldsymbol{B}$ is varied with respect to the Fermi ellipsoid. Quantum theory shows that the energy of the cyclotron motion $\hbar\omega_{\mathrm{c}}$ is quantized with an orbital quantum number $n = 0, 1, 2, \ldots$, and the energy levels of a charge carrier in a magnetic field are given by:

$$E(n,k_{\mathfrak{b}}) = (n + \tfrac{1}{2})\,\hbar\omega_{\mathrm{c}} + \hbar^2 k^2/(2\,m_{\mathfrak{b}}) , \qquad (9)$$

where $\hbar$ is Planck's constant h divided by 2π. The second term is the energy of the translational motion parallel to $\boldsymbol{B}$ with effective mass $m_{\mathfrak{b}}$. In the wave-mechanical description of the charge carriers the wave number $k_{\mathfrak{b}}$ is introduced in the momentum component $p_{\mathfrak{b}} = \hbar k_{\mathfrak{b}}$. The energy levels, given by eq. (9), are called Landau levels and exhibit a parabolic $E(p_{\mathfrak{b}})$-dependence (*fig. 5*). The energy difference between adjacent Landau levels is $\hbar\omega_{\mathrm{c}}$.

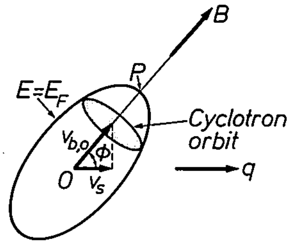


Fig. 4. Cyclotron orbit of a charge carrier on the Fermi ellipsoid in a plane (shaded) perpendicular to the magnetic field B. The direction of sound propagation is given by the wave vector $\boldsymbol{q}$.

The presence of the magnetic field strongly influences the distribution of energy states in momentum space. With no field the allowed states are distributed uniformly over the p-space volume, and the density of states $N(E)$ with energies between E and $E + \Delta E$ increases in proportion to $\sqrt{E}$; see fig. 5. In a magnetic field the allowed states are concentrated close to the quantized cyclotron orbits, which are given by eq. (9), and the number of states in the remaining parts of momentum space is depleted. This leads to the singularities in the density of states near the bottom of the Landau levels, shown in fig. 5.

Let us explain the elementary process of acoustic absorption in quantum-mechanical terms: an electron in the initial state ν with energy $E_\nu = E(n,k_{\mathfrak{b},\mathrm{o}})$ and wave number $k_{\mathfrak{b},\mathrm{o}}$ absorbs an acoustic quantum (phonon) with energy $\hbar\omega$ and wave number $q_{\mathfrak{b}}$ parallel to $\boldsymbol{B}$ and is transferred into a final state ν' with energy $E_{\nu'} = E(n'k'_{\mathfrak{b},\mathrm{o}})$ and wave number $k'_{\mathfrak{b},\mathrm{o}}$. The conditions for conservation of energy and momentum yield:

$$\begin{aligned} E_{\nu'} - E_\nu &= \hbar\omega , \\ k'_{\mathfrak{b},\mathrm{o}} &= k_{\mathfrak{b},\mathrm{o}} + q_{\mathfrak{b}} . \end{aligned} \qquad (10)$$

In a sufficiently high magnetic field the frequency ω of the phonon is much smaller than the cyclotron frequency ω_c, e.g. we have $\omega/\omega_c \approx 2\times10^{-4}$ for $f = 60$ MHz and electrons in bismuth with $m_c = 0.01\, m_0$ at $B = 1$ kG. Therefore in most cases the phonon energy is insufficient to induce electron transitions between Landau levels with different n. The latter transitions are observed in cyclotron-resonance experiments in the microwave range. For ultrasonic investigations the states ν and ν' are located at the same Landau level (see fig. 5). Introducing the selection rule $n - n' = 0$ into eqs. (9) and (10) yields:

$$E_{\nu'} - E_\nu = \hbar\omega \approx \frac{\hbar^2 k_{b,o} q_b}{m_b}, \tag{11}$$

where a small term proportional to $q_b{}^2$ has been neglected. This equation can be rewritten in the form

$$v_s = v_{b,o} \cos \phi, \tag{12}$$

by introducing the velocity component parallel to $\boldsymbol{B}$, $v_{b,o} = p_{b,o}/m_b = \hbar k_{b,o}/m_b$, the velocity of sound $v_s = \omega/q$ and writing $q_b = q \cos \phi$, where ϕ is the angle between the directions of acoustic propagation and magnetic field (see. fig. 4). Equation (12) states that the projection of the carrier velocity parallel to $\boldsymbol{B}$ along the direction of acoustic propagation must be equal to the velocity of sound ("surf-riding" condition for the absorption process). From all possible orbital planes perpendicular to the magnetic field this condition selects one effective plane for each angle ϕ. The distance $v_{b,o}$ from the centre O of the Fermi ellipsoid increases with the angle ϕ and reaches its maximum possible value, the Fermi velocity v_F, at the critical angle $\phi_c = \text{arc} \cos(v_s/v_F)$, where the orbital plane degenerates to the point P on the Fermi ellipsoid (see fig. 4). For electrons in bismuth v_s/v_F is 10^{-2} to 10^{-3}, and the critical angle is very close to $\pi/2$, with $\pi/2 - \phi_c \approx 0.5°$ to $0.05°$ [2]. The condition for the conservation of momentum, eq. (12), cannot be satisfied for $\phi > \phi_c$. In this region Reneker [3] has measured a sharp drop in the magnetoacoustic attenuation of bismuth ("tilt effect"), from which the Fermi velocity could be evaluated. V. E. Henrich [14] has measured nonextremal cross-sections of the Fermi surface with finite $v_{b,o}$ in the magnetoacoustic attenuation of bismuth when ϕ was tilted towards $\pi/2$. In most cases for nonperpendicular field directions the tilt effect can be neglected and $v_{b,o}/v_F$ is extremely small. Therefore in magnetoacoustic absorption orbital planes very near to the extremal cross-section of the Fermi ellipsoid perpendicular to $\boldsymbol{B}$ are observed.

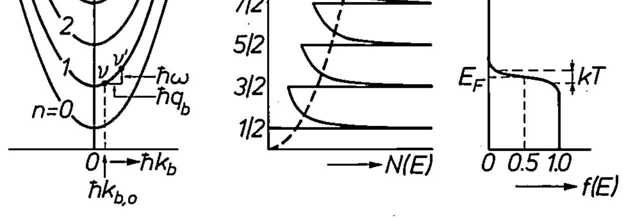


Fig. 5. Landau levels with orbital quantum number $n = 0, \ldots 3$ in the presence of a magnetic field, the density of states $N(E)$ and the Fermi-Dirac distribution function f(E). The points ν and ν' represent the initial and final states of an electron for absorption of a phonon with frequency ω and wave-vector component q_b. For this transition $\Delta n = 0$.

The probability for occupation of energy states is given by the Fermi-Dirac function

$$f(E) = \frac{1}{\exp\,[(E - E_F)/kT] + 1}, \tag{13}$$

which is unity below the Fermi energy and zero above E_F (see fig. 5); the width of the transition region is determined by the thermal energy kT, where k is Boltzmann's constant. The net acoustic absorption is proportional to the difference in occupation probabilities of the electronic states ν' and ν, which is a maximum at $E = E_F$ (see fig. 5) and can be approximated as:

$$f_{\nu'} - f_\nu \approx \left(\frac{\partial f}{\partial E}\right)_\nu (E_{\nu'} - E_\nu) \propto \frac{\hbar\omega}{kT} \cosh^{-2}\left(\frac{E_\nu - E_F}{2\,kT}\right), \tag{14}$$

using eqs. (10) and (13).

In the range of quantizing magnetic fields the ultrasonic attenuation $\alpha(B)$ can be calculated as [15]:

$$\alpha(B) \propto \frac{\omega}{T} B(\hat{e} \cdot \tilde{C} \cdot \hat{q})^2 \frac{m_b}{v_s{}^2 \cos \phi} \times$$

$$\times \sum_n \cosh^{-2}\left(\frac{E(n, k_{b,o}) - E_F}{2\,kT}\right). \tag{15}$$

The coupling factor with the deformation potential enters in accordance with eq. (7), where $\hat{e}$ and $\hat{q}$ are the unit vectors in the directions of acoustic polarization and propagation. The factor B in eq. (15) is a consequence of the fact that the size of permitted cyclotron orbits in momentum space expands with increasing magnetic field and a smaller number of orbits fits into the volume of the Fermi ellipsoid $E = E_F$ (see fig. 5). Up to the point where the last Landau level passes the Fermi energy E_F the number of charge carriers within the Fermi ellipsoid does not change appreciably, i.e. when the number of Landau levels below E_F decreases with increasing B, the population of these levels must increase, thus explaining the factor B. The quantity $m_b/\cos \phi$ in eq. (15) results from integration over eq. (11), the condition for the conservation of energy, and can be calculated from the known band-structure parameters in bismuth.

Physically the acoustic-absorption process may be described as follows. A sharp spike in the ultrasonic attenuation occurs when a singularity in the density of states passes the Fermi level (see fig. 5). The absorption peaks are periodic in $1/B$, and the period can be calculated from the condition $E(n,k_{b,o}) = E_F$, using eqs. (8) and (9):

$$\Delta\left(\frac{1}{B}\right) = \frac{1}{B_{n+1}} - \frac{1}{B_n} = \frac{e\hbar}{m_c E_F} = \frac{e\hbar}{S_{\rm extr}}, \quad (16)$$

neglecting the tilt effect. Here $S_{\rm extr}$ is the extremal cross-section of the Fermi ellipsoid perpendicular to $\boldsymbol{B}$ in momentum space. The different types of charge carriers in bismuth (bands $l = 1, \ldots 4$) can be identified by their characteristic oscillation periods. The spikes in the magnetoacoustic attenuation are much sharper than the oscillations observed in the De Haas-Van Alphen effect, because the orbital plane in acoustic absorption is determined by the sharp selection rule, eq. (12), whereas in the De Haas-Van Alphen effect many orbital planes of different cross-sections and oscillation periods contribute.Most oscillatory contributions interfere destructively, only the extremal cross-sections with $\mathrm{d}S/\mathrm{d}k_b = 0$ yield a net oscillatory effect.

Fig. 6 shows measured quantum oscillations in the magnetoacoustic attenuation for the slow shear wave with $\boldsymbol{q}$ parallel to the x-axis in bismuth, in which the contributions due to the electron bands $l = 2$ and 3 can be separated clearly. The double absorption peak at 13-14 kG is due to the influence of the electron spin. In a magnetic field each Landau level splits into two sublevels (spin "up" and spin "down"). However spin splitting will not be discussed here in further detail.

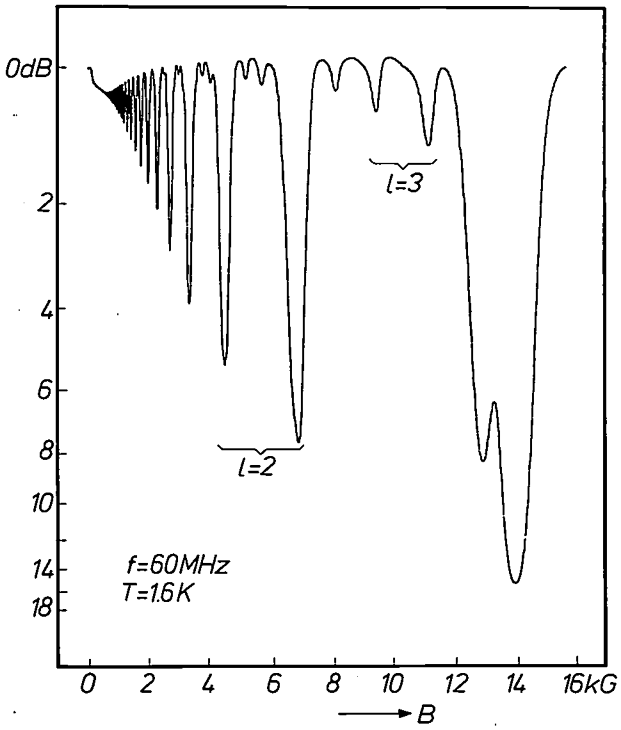


Fig. 6. Transmitted-signal amplitude for the slow shear wave propagating along the x-direction at $f = 60$ MHz, $T = 1.6$ K in bismuth as a function of the magnetic field B, showing quantum oscillations due to the electron bands $l = 2$ and 3.

Determination of deformation potentials in bismuth

Experimental investigations on the magnetoacoustic quantum oscillations in bismuth have been published by several authors [3] [16-20]. In most of these cases the interest was in determining cyclotron masses, extremal cross-sections of the Fermi surface and spin-splitting factors from the oscillation periods $\Delta(1/B)$. The deformation potentials for electrons and holes in bismuth can be determined from the peak values $\alpha_p(B)$ of the magnetoacoustic attenuation [4] [5], which occur at $\cosh^{-2}(\ldots) = 1$ in eq. (15).

This requires careful consideration of the effect of collisions, which in general reduce the amplitude and broaden the width of the absorption peaks. Numerous theoretical articles on this subject have been published [15] [21] [22] which demonstrated that the effect of scattering depends on the product $q_b l_b$, where $q_b = q \cos\phi$ and l_b is the mean free path of the charge carriers in the direction of the magnetic field. In this article collisions are taken into account by multiplying the measured attenuation with a scattering correction factor $K(q_b l_b) \geqslant 1$, which is unity for very high $q_b l_b$ values and increases with decreasing $q_b l_b$. The dependence $K = K(q_b l_b)$ can be determined experimentally.

In order to calculate the deformation potentials from the measured peak attenuation $\alpha_p(B)$, a reduced attenuation is defined:

$$\frac{\alpha_p'(B)}{B} = K\,\frac{\alpha_p(B)}{B}\cdot\frac{v_s^2\cos\phi}{m_b}. \quad (17)$$

From eq. (15) linear combinations between various deformation potential components can be obtained by calculating ratios of reduced attenuations at constant ω and T for two different cases r and s, which may differ in ultrasonic mode, band index or field direction ϕ:

$$\pm\sqrt{\frac{(\alpha_p'(B)/B)_r}{(\alpha_p'(B)/B)_s}} = \frac{(\hat{e}\cdot\tilde{C}\cdot\hat{q})_r}{(\hat{e}\cdot\tilde{C}\cdot\hat{q})_s}. \quad (18)$$

Because of the square root that appears in eq. (18) the deformation potentials can be determined from magnetoacoustic measurements only to a single ambiguity of sign. S. Inoue and M. Tsuji [4] obtained this sign by combining information from magnetoacoustic quantum oscillations with piezogalvanomagnetic measurements in bismuth as a function of static strain: the measurements were carried out by A. L. Jain and

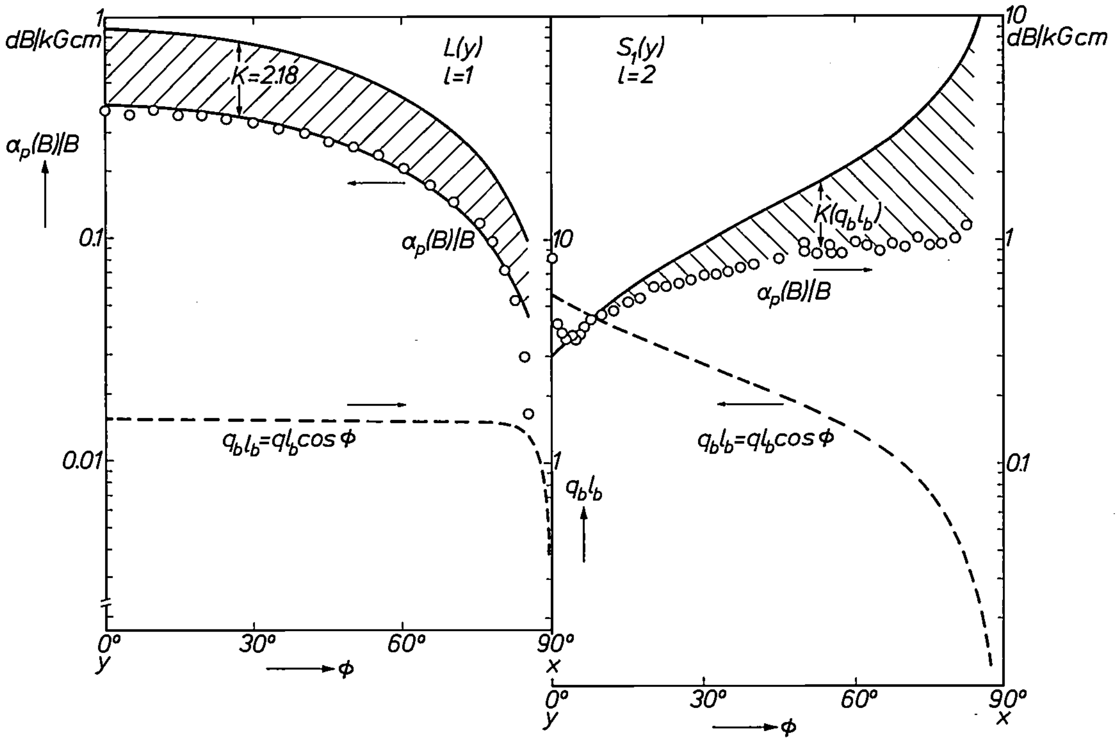
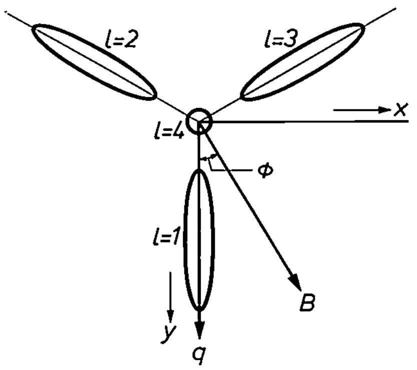


Fig. 7. Attenuation $\alpha_p(B)/B$ for longitudinal wave $L(y)$ (electron band $l = 1$) and shear wave $S_1(y)$ ($e//x$, electron band $l = 2$) in bismuth, propagating along the y-direction, as a function of magnetic field orientation ϕ. The quantity $q_b l_b$ was calculated from eq. (20).



R. Jaggi [23]. The angular dependence of the ultrasonic attenuation in eq. (15) is given by:

$$\frac{\alpha_p(B)}{B} \propto \frac{m_b}{K(q_b l_b)\cos\phi}, \qquad (19)$$

where the mean free path l_b in the direction of the magnetic field is calculated from the velocity component v_b:

$$l_b = \tau v_b = \tau\sqrt{\frac{2E_F}{m_b}},$$

$$q_b l_b = q\cos\phi\cdot\tau\sqrt{\frac{2E_F}{m_b}}. \qquad (20)$$

In bismuth samples in which the purity is not extremely high (ratio of electrical resistivity $\varrho(300\text{ K})/\varrho(4.2\text{ K})$ below 200) the anisotropy of relaxation time τ is small [24] in comparison with the anisotropy of the effective mass. (see however [25]). Therefore in the large-mass direction of greatest attenuation given by eq. (19) the scattering correction factor K is large, because $q_b l_b$ from eq. (20) is small. In the low-mass direction the attenuation becomes small and scattering correction can be neglected; $K \approx 1$, since $q_b l_b$ is large.

This behaviour will be explained for two specific cases in bismuth: *figs. 7a* and *7b* show the angular dependence of the peak attenuation $\alpha_p(B)/B$ for the longitudinal wave and the shear wave with $\hat{e}//x$-axis, respectively, propagating in the y-direction. The open circles are the experimental points at $T = 1.6$ K and $f = 60$ MHz, the solid lines give the calculated angular dependence proportional to $m_b/\cos\phi$, which would be

[14] V. E. Henrich, Phys. Rev. Letters **26**, 891, 1971.
[15] V. L. Gurevich, V. G. Skobov and Y. A. Firsov, Zh. eksper. teor. Fiz. **40**, 786, 1961.
[16] A. P. Korolyuk, Fiz. tverd. Tela **5**, 3323, 1963, and Zh. eksper. teor. Fiz. **51**, 697, 1966.
[17] A. M. Toxen and S. Tansal, Phys. Rev. **137**, A211, 1965.
[18] S. Mase, Y. Fujimori and H. Mori, J. Phys. Soc. Japan **21**, 1744, 1966.
[19] Y. Sawada, E. Burstein and L. Testardi, J. Phys. Soc. Japan **21**, Suppl., 760, 1966.
[20] T. Sakai, Y. Matsumoto and S. Mase, J. Phys. Soc. Japan **27**, 862, 1969.
[21] S. H. Liu and A. M. Toxen, Phys. Rev. **138**, A487, 1965.
[22] A good review is given by Y. Shapira, in: Physical Acoustics, editor W. P. Mason, Vol. V, Academic Press, New York 1968, page 1.
[23] A. L. Jain and R. Jaggi, Phys. Rev. **135**, A708, 1964.
[24] R. N. Zitter, Phys. Rev. **127**, 1471, 1962.
[25] R. Hartman, Phys. Rev. **181**, 1070, 1969.

valid for $K = 1$, and the dashed lines are the quantities $q_b l_b$, calculated from eq. (20), using an average value of isotropic relaxation time $\tau = 1.9 \times 10^{-10}$s [24] and a Fermi energy $E_F = 17.7$ meV for electrons. In fig. 7*b* the coupling to the electron band $l = 2$ increases, and both cos ϕ and the mean free path l_b decrease with increasing angle ϕ. The resulting marked decrease of $q_b l_b$ corresponds to an increasing correction factor K, which is shown as the difference (hatched area) between the experimental points and the solid theoretical curve, fitting both curves at $\phi = 0°$ with $K = 1$. In this way the function $K(q_b l_b)$ can be determined experimentally. In fig. 7*a* the coupling to the electron band $l = 1$ decreases as a function of ϕ. The corresponding increase in the mean free path l_b is compensated by the decrease of cos ϕ, resulting in a constant value $q_b l_b = 1.55$ up to $\phi = 80°$. By comparison with fig. 7*b* a constant scattering correction factor $K = 2.18$ is obtained, which explains the good agreement of the angular dependence between the theoretical curve and the experimental points: to obtain the corrected attenuation the experimental curve has to be displaced by a constant amount along the logarithmic scale.

By taking ratios of reduced attenuations for various ultrasonic modes propagating in the x- and y-directions, linear relations between the deformation potentials a_1, a_2 and a_4 for electrons are obtained, using eqs. (4), (7), (15) and (18). These are plotted as straight lines in a coordinate system with axes a_1/a_2 and a_4/a_2 in *fig. 8*. The two case numbers on each line refer to different combinations of ultrasonic mode and band index and are explained in *fig. 9*. The average coordinates of the crossing region yield the values $a_1/a_2 = - 0.37$ and $a_4/a_2 = + 0.25$. Using these values, the ratio $a_3/a_2 = - 0.29$ can be calculated from measurements with acoustic propagation along the z-direction and the bisector between y and z. So far, the deformation potentials of electrons are known only relative to the value a_2, which can be determined from a fit to the absolute value of the attenuation given by eq. (15). The result is $a_2 = \pm 5.9$ eV. The anisotropy of the measured magnetoacoustic attenuation for the 14 cases of fig. 9 can be described theoretically with an accuracy of about 14% using four deformation-potential components for the electrons. In a similar way the deformation potential for holes can be determined as $b_2/b_1 = -1.03$ and $b_2 = \pm 1.2$ eV. The ultrasonic method provides a very direct way of measuring the deformation potential with considerable redundancy. The sign ambiguity, which is inherent in this method, can be resolved by a combination with piezogalvanomagnetic measurements [23]. In this way Inoue and Tsuji [4] have demonstrated that both signs of a_2 and b_2 are negative.

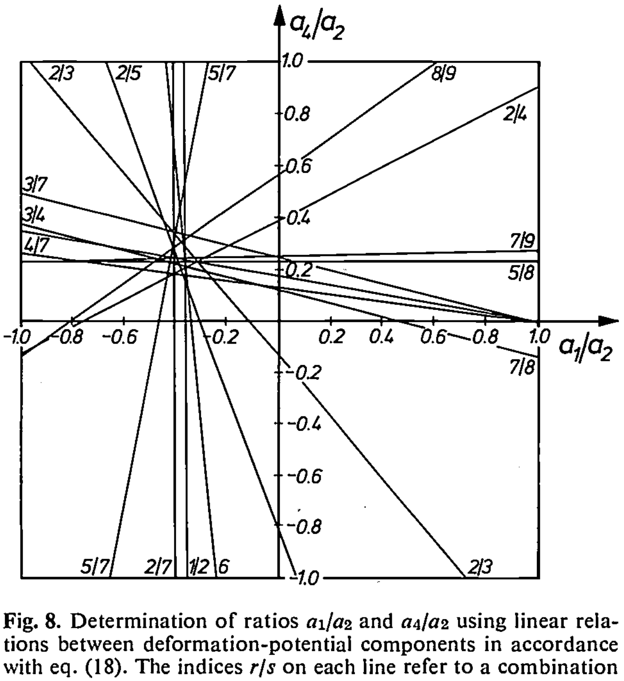


Fig. 8. Determination of ratios a_1/a_2 and a_4/a_2 using linear relations between deformation-potential components in accordance with eq. (18). The indices r/s on each line refer to a combination of case numbers, which are explained in fig. 9. The lines intersect in a region around $a_1/a_2 = -0.37$ and $a_4/a_2 \doteq +0.25$.

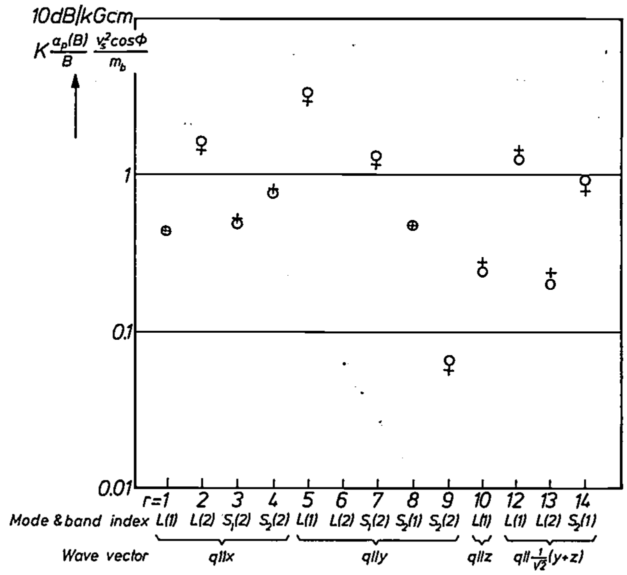


Fig. 9. Reduced ultrasonic attenuation $\alpha_p'(B)/B = K\alpha_p(B)v_s^2 \cos\phi/Bm_b$ in bismuth for 14 different combinations of ultrasonic mode and band index. o measured values, + theoretical fit according to eq. (15), using $a_2 = \pm 5.9$ eV.

Effect of electric drift field: "kink" effect and ultrasonic noise in bismuth

The discussion above on the interaction between acoustic waves and charge carriers in bismuth was confined to the case of zero drift velocity $\boldsymbol{v}_d$. To investigate acoustodynamic effects in the presence of a pulsed electric drift field $\boldsymbol{E}$, it is advantageous to use a sample geometry with a strong transverse magnetic field $\boldsymbol{B}$ [7], shown in fig. 10. Owing to the high magnetoresistance of bismuth at liquid-helium temperatures the current densities required to realize drift velocities larger than the sound velocity v_s, can be reduced by a factor of 10^6 with respect to the case of zero magnetic field. The origin of the high magnetoresistance of bismuth can be seen from the vector diagram of *fig. 11:* a strong magnetic field rotates the current vectors for electrons and holes in opposite senses from the direction of the electric field. The angles of rotation, called the Hall angles $\Theta_{H,e}$ and $\Theta_{H,h}$ are nearly 90°, since the ratio of current components parallel and perpendicular to E is given by $I_{\parallel}/I_{\perp} = (\mu B)^{-1} \approx 10^{-3}$ - 10^{-4} at $B = 10$ kG with carrier mobilities $\mu = 10^7$ - 10^8 cm^2/Vs. Since the current vectors for electrons and holes are deflected in opposite directions, the limitation in deflection angle, which occurs in unipolar conductors because of the presence of the transverse Hall field, is largely absent in semimetals.

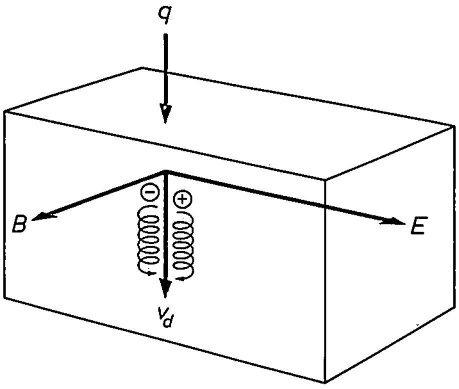


Fig. 10. Motion of electrons and holes in bismuth in crossed electric and magnetic fields. The average drift velocity $v_d = E\times B/B^2$. The direction of sound propagation is again denoted by q.

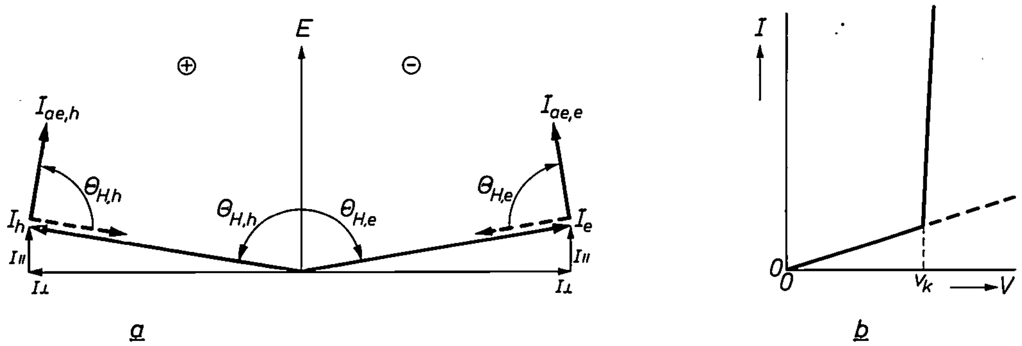


Fig. 11. *a*) Vector diagram showing the ohmic current components I and the acoustoelectric current components I_{ae} for electrons (subscript e) and holes (subscript h) in the presence of a strong magnetic field, perpendicular to the plane of the drawing. The arrow E denotes the direction of the electric field. The Hall angles Θ_H are nearly 90°. *b*) Kink effect in bismuth. With increasing voltage V, at a certain value V_k the increase of the total current becomes much more marked due to the build-up of the acoustoelectric current component. The dashed line represents the ohmic characteristic observed before the build-up of the acoustoelectric current.

The motion of the drifting charge carriers will be described in "real space" (see fig. 10): In the presence of crossed electric and magnetic fields electrons and holes perform cyclotron motions with opposite senses of rotation, where the centre of the cyclotron orbits is displaced perpendicular to E and B with an average d.c. drift velocity $v_d = E\times B/B^2$, using the "free-carrier" model. The anisotropy of the effective mass modifies this expression, and S. G. Eckstein [26] has shown that for certain directions of E and B very high drift velocities $v_d = 7$ - $8\ E/B$ can be obtained. To obtain a drift velocity E/B comparable with the velocity of sound ($\approx 10^5$ cm/s) at $B = 10$ kG, an electric drift field E of 10 V/cm is required.

Because of the deformation potential interaction (see eq. 7) an acoustic wave propagating along say the x-direction generates a sinusoidally varying potential $\Phi_s(x,t)$ with an associated electric field-strength

$$E_s(x,t) = \frac{1}{e}\,F_s = -\,\frac{1}{e}\,\frac{\partial\Phi_s}{\partial x}$$

acting on the charge carriers (see *fig. 12*). The majority of the charge carriers move with the d.c. drift velocity v_d. However the acoustic wave carries along with it a certain concentration

$$n_s = |n_s|\ \exp\,[i(qx - \omega t - \Psi)]$$

of charge carriers at the velocity of sound because of the electric field

$$E_s = |E_s|\ \exp\,[i(qx - \omega t)],$$

where Ψ is the phase angle between E_s and n_s. The carrier concentration n_s reaches a maximum when the carriers move at the velocity of sound, and decreases with increasing velocity difference between carriers and the acoustic wave. An ultrasonic wave generates a d.c. component of electric current, called the "acoustoelectric" current $I_{ae} \propto |n_s|\ |E_s|\ \cos\Psi$, which is proportional to the square of the mechanical strain.

The energy balance between the drifting charge carriers and the sound wave is governed by the average interaction force

$$F_s = \langle F_s(x)\, n_s(x) \rangle / \langle | n_s(x) | \rangle, \qquad (21)$$

where $\langle\ \rangle$ is the average value over one acoustic wavelength. Depending on the projection of the drift velocity along the direction of acoustic propagation $v_{d,q} = \boldsymbol{v}_d \cdot \hat{q}$, three different cases have to be discussed (see fig. 12):

1) No interaction for $v_{d,q} = v_s$: the maxima of the carrier concentration n_s coincide with the minima of the potential Φ_s, and the average interaction force vanishes.
2) Ultrasonic amplification for $v_{d,q} > v_s$: the maxima of the carrier concentration n_s are displaced from the potential minimum, so that an average interaction force opposed to the drift motion exists, which slows down the carriers. The wave therefore gains energy from the carriers and is amplified.
3) Ultrasonic attenuation for $v_{d,q} < v_s$: in comparison with the previous case the sign of the average interaction force is reversed: the charge carriers are accelerated by the electric field E_s and receive this energy from the acoustic wave, which is attenuated.

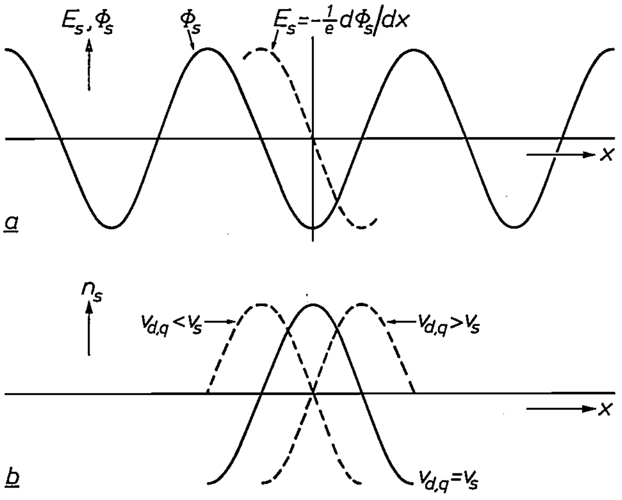


Fig. 12. *a*) Electrical potential $\Phi_s(x,t)$ and electrical field-strength $E_s(x,t)$ caused by a sound wave, plotted as a function of x. *b*) The same for the charge-carrier concentration $n_s(x,t)$ for various drift velocities $v_{d,q}$. If $v_{d,q}$ is equal to the velocity of sound v_s there is no interaction. If $v_{d,q} > v_s$ the sound wave is amplified, if $v_{d,q} < v_s$ the sound wave is attenuated.

When the condition $v_{d,q} > v_s$ for ultrasonic amplification is satisfied, the drifting carriers generate ultrasonic noise in a broad frequency range, even when no external acoustic wave is coupled into the crystal.
This ultrasonic noise is excited in the direction of the carrier drift and drags along a d.c. component of acoustoelectric current I_{ae}. Esaki [7] has observed a kink in the current-voltage characteristic of a bismuth sample at $T = 2$ K in a strong transverse magnetic field at a voltage V_k, where the drift velocity exceeded the velocity of sound (see fig. 11*b*). The ratio of differential resistivities below and above the kink voltage was higher than 50. The kink effect may be explained by means of fig. 11*a*: owing to the presence of ultrasonic noise above V_k acoustoelectric currents $I_{ae,e}$ and $I_{ae,h}$ for electrons and holes are generated primarily in opposite directions (dashed arrows) to the ohmic currents I_e and I_h and are subsequently rotated through the Hall angles $\Theta_{H,e}$ and $\Theta_{H,h}$, thus creating a strong increase of current flow in the direction of the electric field. The author [9] has observed the build-up of ultrasonic noise associated with the kink effect, using a shear quartz transducer (*Y*-cut), attached to a surface perpendicular to the *x*-direction of a bismuth sample. *Figs. 13a* and 13*b* show the ultrasonic noise in the frequency range around 60 MHz (upper trace), the drift current (centre trace), and the drift voltage (lower trace) at $T = 1.8$ K and $B = 16$ kG. In fig. 13*a* the directions of electric and magnetic fields are chosen such that the drift velocity $\boldsymbol{v}_b = \boldsymbol{E} \times \boldsymbol{B}/B^2$ points towards the quartz transducer, in fig. 13*b* the magnetic field direction is reversed, and the carrier drift is directed away from the transducer. The acoustoelectric current builds up with an incubation time of approximately 4 μs (centre trace), a corresponding decrease in the drift voltage is observed (lower trace), and the build-up of ultrasonic noise is detected at a time 5 μs after the leading edge of the drift pulse (upper trace in fig. 13*a*). On field reversal, the ultrasonic noise is detected 7 μs later, and its amplitude is reduced by 20 dB; the sensitivity of the oscilloscope was increased by a factor of 10 (fig. 13*b*). This behaviour demonstrates that the ultrasonic noise is excited in the direction of the carrier drift velocity: in fig. 13*b* the noise travels in a direction away from the transducer and reaches the latter only after being reflected from the sample surface perpendicular to x. The additional delay time corresponds approximately to the sample width divided by the sound velocity for the slow shear wave ($v_s = 0.89 \times 10^5$ cm/s), propagating in the *x*-direction. The ultrasonic noise was also observed in a broad frequency range up to 220 MHz. Using drift pulses with very sharp leading edges, T. Yamada [27] has measured the current-voltage characteristics of a bismuth sample at a time interval prior to the build-up of ultrasonic noise and acoustoelectric current and found a purely ohmic behaviour, even for voltages far above the kink point, e.g. $V = 5\ V_k$ (see fig. 11*b*).

[26] S. G. Eckstein, Physics Letters **13**, 30, 1964.
[27] T. Yamada, J. Phys. Soc. Japan **20**, 1647, 1965.
[28] J. H. McFee, J. appl. Phys. **34**, 1548, 1963.

Ultrasonic amplification and inverted quantum oscillations in bismuth

In this section the effect of an electric drift field on the attenuation of an ultrasonic wave will be discussed. A. M. Toxen and S. Tansal [11] have reported an increase of 14 dB/cm in the amplitude of a 15 MHz shear wave in bismuth on application of a drift field. *Fig. 14* shows the pulse amplitudes for three ultrasonic modes (upper traces) at $f = 60$ MHz and $T = 4.2$ K in bismuth as a function of electric drift-pulse amplitude [12] (lower traces). The first pulse *Tr* on the upper traces corresponds to electrical break-through of the transmitter pulse; the ultrasonic pulses, labelled L, S_1 and S_2, refer to the longitudinal mode and to the fast and slow shear waves respectively. The pulse $3L$ is due to the triple-path signal of the longitudinal mode. A magnetic field $B = 13$ kG was chosen, which corresponds to a maximum in the magnetoacoustic attenuation, similar to fig. 6. In figs. 14*a*–*e* the carrier drift motion is in the same direction as the ultrasonic propagation. The ultrasonic amplitude increases with the drift field. The largest amplitude changes, up to 27 dB/cm, are observed for the slow shear wave S_2. Ultrasonic noise is generated beyond the kink point (see figs. 14*d* and 14*e*). In fig. 14*f* the carrier drift motion is reversed with respect to the signal propagation, which is evident from the delayed arrival of the noise pulse, and the drift field causes only small changes in the signal amplitude. For signal and noise propagation in the same direction no further increase of signal amplitude is observed when the noise amplitude reaches measurable values (figs. 14*d* and 14*e*). On the other hand the theory of ultrasonic amplification predicts a further increase in signal amplitude on increasing drift field beyond the kink voltage. The limitation of ultrasonic amplification observed experimentally is evidently caused by nonlinear interaction between the ultrasonic noise and the ultrasonic signal. The large-amplitude statistical electrical potential due to the noise competes with the weak sinusoidal modulation of the electrical potential due to the signal (see fig. 12). This effect is well known for the piezoelectric semiconductor CdS [28], where the incubation time between the leading edge of the drift pulse and the build-up of ultrasonic noise may be appreciable. By shifting the time interval of the signal pulse with respect to the noise pulse, McFee [28] has demonstrated that the signal amplification of 45 MHz shear waves in CdS can be reduced up to 23 dB/cm due to the presence of ultrasonic noise. In the case of bismuth the amplitude of the generated ultrasonic noise strongly depends on the direction of the magnetic field. Extensive investigations [12] have revealed the fact that maximum signal amplification is obtained under conditions of minimum noise generation and *vice versa*.

The effect of an electric drift field on the magnetoacoustic quantum oscillations in bismuth is demonstrated in *fig. 15*, which shows the amplitude of the slow shear wave S_2 as a function of magnetic field. In curve *a*, relating to a frequency of 60 MHz, no electric drift field is present. Quantum oscillations in the magnetoacoustic attenuation similar to those in fig. 6 are observed. Curve *b* refers to a frequency of 188 MHz, and an electric drift field above the kink point is present. In comparison with curve *a* the quantum oscillations under amplifying conditions are inverted: when a singularity in the density of states crosses the Fermi

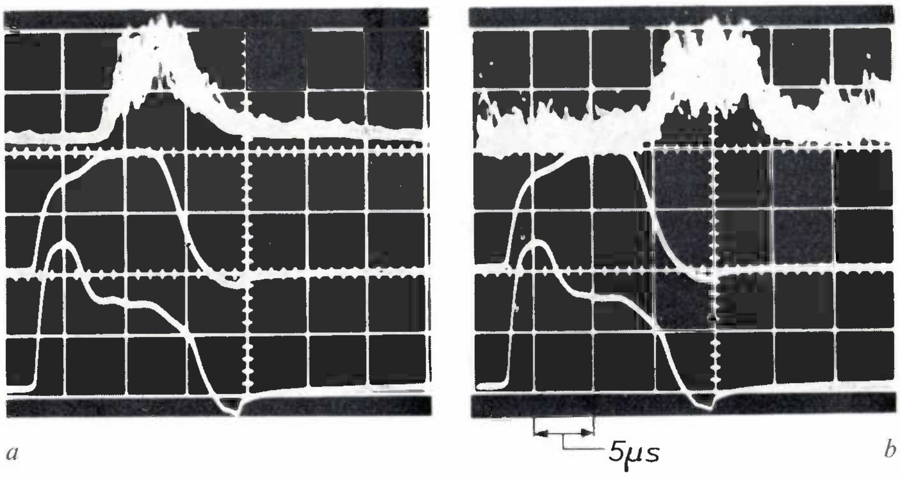


Fig. 13. The upper traces show the build-up of ultrasonic noise (frequency range around 60 MHz) along the x-direction in a bismuth sample, $B = 16$ kG, $T = 1.8$ K. The centre traces show the drift current (calibration 50 A per major division) and the lower traces show the drift voltage (calibration 20 V per major division). *a*) Drift velocity directed towards transducer. *b*) Drift velocity reversed.

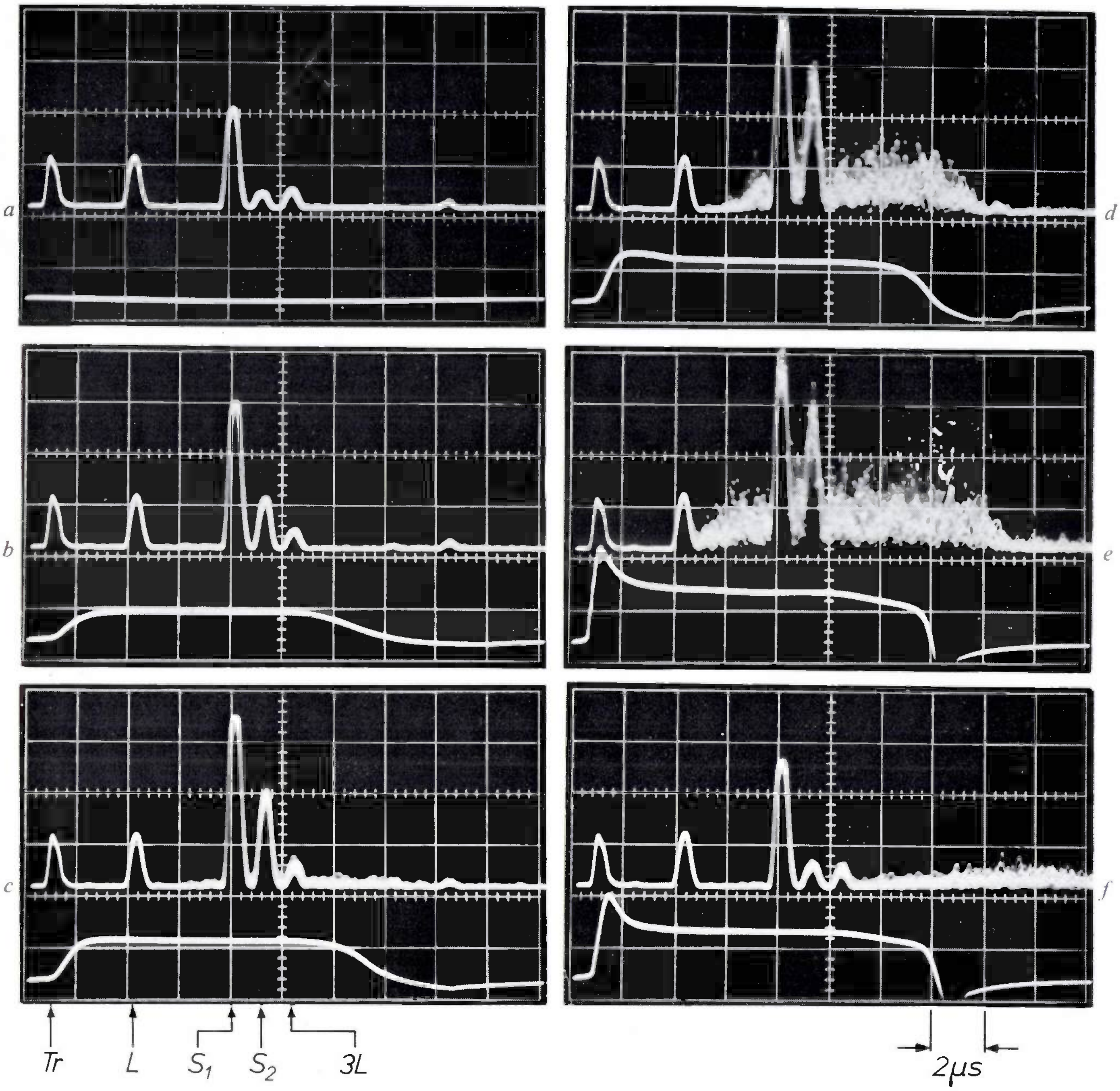


Fig. 14. Ultrasonic signal amplitudes and ultrasonic noise at $f = 60$ MHz (upper traces) and drift voltage (lower traces) in bismuth plotted against time at $B = 13$ kG, $T = 4.2$ K. *a)-e)* Drift velocity parallel to sound propagation. *f)* Drift velocity reversed. The symbols and the details of the curves are explained in the text.

level (see fig. 5) a maximum of ultrasonic attenuation is observed without a drift field (curve *a*), and at the same magnetic field a maximum in the ultrasonic amplification appears in the presence of an electric drift field E (curve *b*). The latter curve was recorded under conditions of optimum signal amplification above the kink point with a drift current nearly independent of magnetic field. At constant drift current theory predicts (1) for small magnetic fields (weakly developed kink effect): $E \propto B^2$, i.e. a drift velocity $\boldsymbol{v}_{\mathrm{d}} \approx E/B \propto B$, and (2) for large magnetic fields (strongly developed kink effect): $E \approx E_k \propto B$, i.e. a drift velocity $\boldsymbol{v}_{\mathrm{d}} \approx E/B \approx$ const. [7]. The drift velocity for curve *b*, plotted in the lower part of fig. 15, is nearly independent of the magnetic field between 4 and 16 kG and decreases below 4 kG. The inverted quantum oscillations under amplifying conditions are therefore portrayed as a function of magnetic field with nearly constant $\boldsymbol{v}_{\mathrm{d}}$.

The optimum results for ultrasonic amplification in bismuth were obtained [13] using the slow shear wave with $q//x$, the electric drift field was applied along the z-direction, the magnetic field was parallel to the bisector direction between the $-y$ and $+z$ axes. *Fig. 16* shows the change in ultrasonic attenuation $\alpha - \alpha_0$ for various frequencies between 61 and 267 MHz as a function of drift velocity at $B = 2.15$ kG, referred to the attenuation α_0 at $B = 0$. The directions of acoustic propagation and carrier drift are the same for positive values of the drift velocity. Negative values of $v_{\rm d}$ mean that the direction of the drift velocity is reversed by magnetic field reversal. The frequency dependence of maximum amplification with drift field and maximum attenuation without drift field is plotted in *fig. 17*. For $f = 267$ MHz a maximum amplification $(\alpha - \alpha_0)_{\rm ampl} = -53.5$ dB/cm was measured, and the attenuation without drift field was extrapolated to be $(\alpha - \alpha_0)_{\rm att} = 93.5$ dB/cm, such that the drift field caused a total change in attenuation of about 147 dB/cm. Fig. 17 also shows the extrapolated frequency dependence of the attenuation α_0 given by Reneker [3],

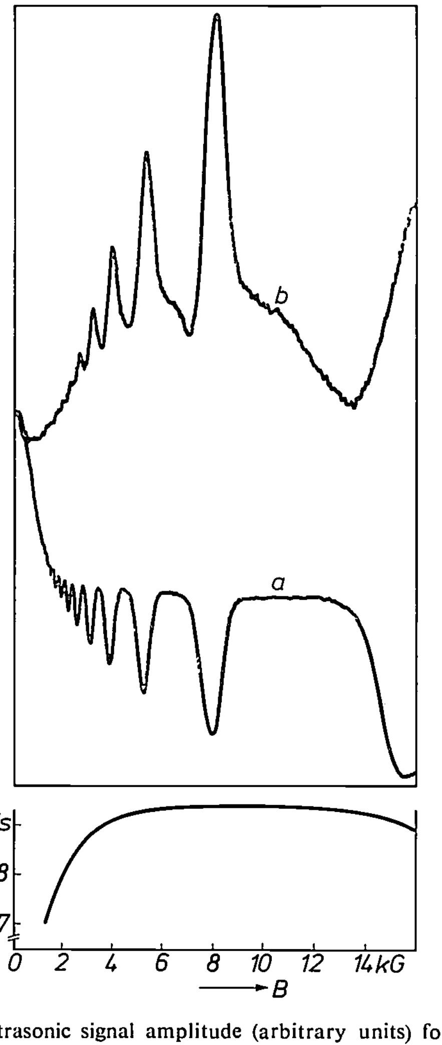


Fig. 15. Ultrasonic signal amplitude (arbitrary units) for slow shear wave (*above*) and drift velocity E/B (*below*) in bismuth as a function of magnetic field at $T = 4.2$ K. Curve *a*, quantum oscillations at $f = 60$ MHz without drift field. Curve *b*, "inverted" quantum oscillations at $f = 188$ MHz with drift field present.

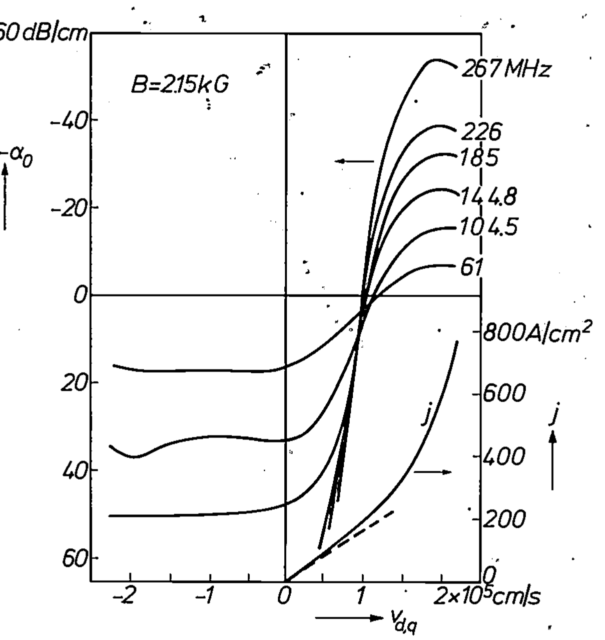


Fig. 16. Attenuation $\alpha - \alpha_0$ of slow shear wave $(q//x)$ and drift-current density j in bismuth as a function of the drift velocity $v_{\rm d,q} = 1.7\,E/B$ at $B = 2.15$ kG and $T = 4.2$ K.

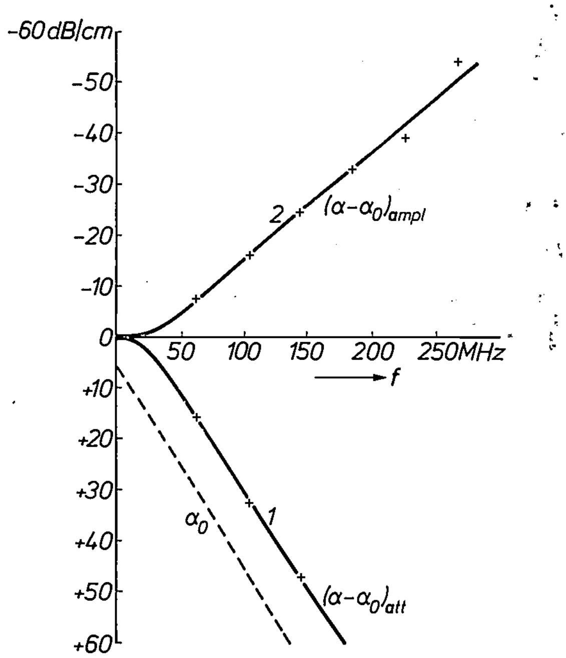


Fig. 17. Maximum changes in the attenuation without drift field (curve *1*) and amplification with drift field (curve *2*) for the slow shear wave $(q//x)$ in bismuth as a function of frequency f at $B = 2.15$ kG and $T = 4.2$ K. The dashed line shows the attenuation α_0 quoted by Reneker [3] at $T = 78$ K.

demonstrating that the measured values of ultrasonic amplification do not yield overall acoustic gain for the bismuth crystal. The situation is different in piezoelectric CdS, where "net gain" can be obtained by ultrasonic amplification [6]. High values of ultrasonic amplification in the case of bismuth (fig. 16) are favoured by the following factors:

The sound velocity for the slow shear wave is low: $v_\mathrm{s} = 0.89 \times 10^5$ cm/s.

The mass-anisotropy factor in the drift velocity [26] is high for the electron band $l = 2$, where $v_{\mathrm{d,q}} = 1.7\ E/B$.

The kink in the drift-current curve is quite small at low magnetic fields. A large drift velocity can therefore be obtained: $(v_{\mathrm{d,q}})_{\mathrm{max}}/v_\mathrm{s} \approx 2.5$.

The generation of ultrasonic noise is weak at low magnetic fields, and the incubation time for noise build-up is long [27]. The degradation of ultrasonic amplification due to nonlinear interaction with noise is consequently small.

Vortices

J. Volger

Many examples of vortices occur in nature: we find them in liquids, in galaxies, in superconductors. It is the aim of this survey to point out some analogies to the reader. In the first section some results of vector analysis are mentioned, together with a few principles of the theory of electricity. Then vortices in liquids, dislocations in crystals, fluxlines in superconductors and vortices in superfluid helium are discussed. Both generally-known aspects and some results of recent investigations are presented.

A physical situation is often described by a vector field. In each point of a certain space a vector $P(x,y,z)$ is given, for instance the velocity $v(x,y,z)$ of a streaming liquid. Two operations lead to important information on the spatial variation of the vector, viz.

$$\operatorname{div} \boldsymbol{P} = \frac{\partial P_x}{\partial x} + \frac{\partial P_y}{\partial y} + \frac{\partial P_z}{\partial z}\,, \tag{1}$$

$$\operatorname{curl} \boldsymbol{P} = \begin{vmatrix} 1_x & 1_y & 1_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ P_x & P_y & P_z \end{vmatrix}. \tag{2}$$

Gauss's theorem may be written as

$$\oint\!\!\oint_S \boldsymbol{P} \cdot \mathrm{d}\boldsymbol{S} = \iiint_V \operatorname{div} \boldsymbol{P}\, \mathrm{d}V. \tag{3}$$

It says that, in the velocity field of an incompressible fluid, fluid appears (from a source) or disappears (in a sink) at points where $\operatorname{div} \boldsymbol{v} \neq 0$.

The circulation of the vector field, expressed as the operation *curl*, also becomes obvious in the hydrodynamical picture. According to Stokes's theorem we have:

$$\oint_l \boldsymbol{P}\, \mathrm{d}\boldsymbol{l} = \oint\!\!\oint_S \operatorname{curl} \boldsymbol{P}\, \mathrm{d}\boldsymbol{S}, \tag{4}$$

i.e. the circulation, which is the line integral of the tangential component taken along a closed contour, is given by the total amount of curl enclosed by the contour. If somewhere $\operatorname{curl} \boldsymbol{P} \neq 0$ and we follow this vector going along its own direction we shall find that it does not vanish, because a general property of the vector curl is that its divergence is zero, $\operatorname{div} \operatorname{curl} \boldsymbol{P} \equiv 0$. Often in the physical situation a thin cylinder may be distinguished such that inside $\operatorname{curl} \boldsymbol{P} \neq 0$ and outside $\operatorname{curl} \boldsymbol{P} = 0$. This cylinder is called the core of a vortex. It ends at the surface of the field under consideration — for instance the wall of a vessel in which flowing liquid is confined — or it is ring-wise closed on itself.

A vector field is determined by its divergences and curls, in other words it can be calculated from its sources and its vorticity. An example of a vector field determined by its divergences is the electrostatic field of a system of fixed point charges e_i. The electrical field strength can be obtained from an electrostatic potential function ϕ, which is given by

$$\phi(x,y,z) = \sum \frac{e_i}{r_i(x,y,z)}\,, \tag{5}$$

where $r_i(x,y,z)$ is the distance between the point (x,y,z) and the point charge e_i. From (5) the electric field strength $\boldsymbol{E}$ follows by application of the gradient operator:

$$E = -\operatorname{grad} \phi = -\frac{\partial \phi}{\partial x} 1_x - \frac{\partial \phi}{\partial y} 1_y - \frac{\partial \phi}{\partial z} 1_z. \tag{6}$$

An example of a vector field determined by its curl is the magnetic field around a wire carrying a current. We have, for current density $\boldsymbol{i}$, Maxwell's equation $\operatorname{curl} \boldsymbol{H} = 4\pi \boldsymbol{i}/c$. The magnetic field can be obtained from a vector potential A which is given by

$$A(x,y,z) = \int \frac{\boldsymbol{i}\, \mathrm{d}V}{c\, r_v(x,y,z)}\,, \tag{7}$$

where $r_v(x,y,z)$ is the distance between the point (x,y,z) and the volume element $\mathrm{d}V$. The magnetic field $\boldsymbol{H}$ follows from (7) by application of the operator curl:

$$\boldsymbol{H} = \operatorname{curl} A. \tag{8}$$

Many a young student is puzzled by this vector potential when it is encountered for the first time [1]. The author remembers being very surprised when he checked the form of $\boldsymbol{H}$ and A in the following two cases. *Infinite straight wire*, carrying a current J; see *fig.1.*

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

[1] H. B. G. Casimir, Theorie der electriciteit, Het Kompas, Antwerp/De Spieghel, Amsterdam, 1936.

The field $\boldsymbol{H}$ is tangential to circles surrounding the wire. Its strength is

$$H = \frac{2J}{rc}.$$

$$\boldsymbol{H} = \operatorname{curl} \boldsymbol{A};\; \boldsymbol{A} = A_z \boldsymbol{1}_z;\; A_z = \frac{-2J}{c} \ln r + [c]. \qquad (9)$$

It seemed understandable that $\boldsymbol{H}$ would diminish at large distance from the wire, but alarming that $\boldsymbol{A}$ would go to infinity (although logarithmically). The question arises: what physical meaning does this vector potential $\boldsymbol{A}$ have?

Solenoidal magnetic field H_0 within an infinite straight solenoid, an axial counter current at the cylindrical surface providing for a zero net current in the axial direction. Outside the solenoid (see *fig. 2*), $H = 0$. The vector potential is tangential to circles around the solenoid; its strength is

$$A = 2\phi/r \quad \text{outside}$$

(although $H = 0$ outside) and (10)

$$A = 2\pi r\, H_0 \quad \text{inside,}$$

where ϕ is the magnetic flux of the solenoid:

$$\phi = \pi r_0^2\, H_0.$$

This remarkable result also leads to the provocative question: does A have any physical meaning?

Thus we have already started with the programme of this article: to make an excursion through some vector fields, taking the curl as the guiding element. We shall consider particularly such situations in which vortices form the main aspect. At the core of a vortex curl $\boldsymbol{P} \neq 0$ and outside curl $\boldsymbol{P} = 0$. Various phenomena will be related through this binding element. In four cases we shall look at the characteristics, the generation and the dynamics of vortices, and point out analogies. Of course, it is not intended to give rigorous proofs or comprehensive treatments.

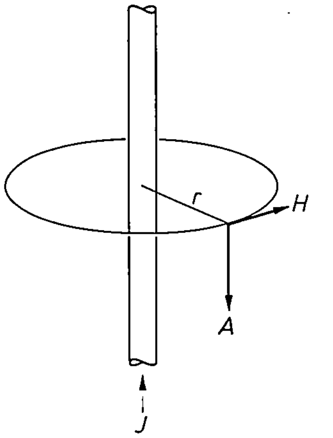


Fig. 1. The magnetic field $\boldsymbol{H}$ and the vector potential A in the space around a straight wire of infinite length.

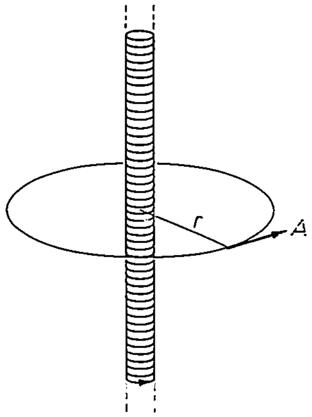


Fig. 2. The vector potential A in the space around a solenoid of infinite length.

Vortices in liquids

Hydrodynamics was developed substantially in the 19th century, especially the simpler theory, which has moreover been a model for the theory of electricity.

The most prominent characteristic of a fluid is its relatively small viscosity. Viscosity causes irreversible transfer of momentum from places where the fluid velocity is large to places where it is small, provided that the velocity field has a deforming action upon the fluid. When there is uniform, i.e. rigid rotation of the fluid as a whole (angular velocity ω), curl $\boldsymbol{v} = 2\omega$ holds for the local velocity everywhere and this does not give rise to viscous effects. Where momentum transfer occurs we may expect damping forces, internal friction, dissipation.

However, in the ideal fluid we ignore viscosity effects. If the gradient of the hydrostatic pressure is the driving force, the equation of Euler — akin to Newton's law — governs the velocity field:

$$\varrho \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\operatorname{grad} p, \qquad (11)$$

where $\mathrm{d}\boldsymbol{v}/\mathrm{d}t$ is the acceleration of a fluid element and ϱ its density. From this equation one derives Kelvin's theorem:

$$\frac{\mathrm{d}}{\mathrm{d}t} \oint \boldsymbol{v}\, \mathrm{d}\boldsymbol{l} = 0, \qquad (12)$$

which says that the circulation $\oint \boldsymbol{v}\, \mathrm{d}\boldsymbol{l}$, taken along a closed contour moving with the fluid, is constant in time. Once the velocity field of a fluid is irrotational, this remains so — at least in the approximation in which Euler's equation holds. On the other hand, once a vortex exists in the fluid, it will persist in undiminished strength.

The characteristics of a vortex in a fluid are simple. The curl of the velocity field is in the core. Assuming that the field is irrotational outside and that there is no potential flow, the velocity field is easily seen to be concentric around the vortex core, with

$$|\boldsymbol{v}| = \frac{[c]}{r}. \tag{13}$$

See *fig. 3*. The analogy with fig. 1 is obvious. The situation within the core is less simple. In certain cases there may be empty space.

An interesting case arises when the liquid is not isotropic and the set of hydrodynamic parameters (such as $\boldsymbol{v}$ and p) must include the so-called director $\boldsymbol{d}$. This is so in liquid crystals, where $\boldsymbol{d}$ may, for instance, be the local direction in which the long molecules of the substance are aligned. Geurst [2] has recently given an extension of Kelvin's theorem for this case. The extension is a consequence of the existence of a moment of inertia for a rotation about an axis perpendicular to $\boldsymbol{d}$. The circulation $\Gamma = \oint \boldsymbol{v}\,\mathrm{d}\boldsymbol{l}$ refers to the normal centre-of-gravity motions of the molecules. However, in director-space a circulation Γ' is defined by:

$$\Gamma' = \alpha \oint \frac{\mathrm{d}\boldsymbol{d}}{\mathrm{d}t}\,\mathrm{d}\boldsymbol{d},$$

α being the molecular moment of inertia per unit mass and Geurst has shown that $\Gamma + \Gamma' =$ constant.

How are vortices generated? Let us consider again the non-ideal fluid. Now, instead of Euler's equation of motion, the equation of Navier-Stokes holds. In the case of an incompressible viscous fluid it may be written as

$$\varrho \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -\mathrm{grad}\, p - \eta\, \mathrm{curl\, curl}\, \boldsymbol{v}, \tag{14}$$

where η is the dynamic viscosity coefficient of the fluid. It is the viscosity which gives rise, for instance, to the well-known parabolic Poiseuille profile of the laminar fluid flow in a tube with uniform circular cross-section:

$$v_x = \frac{r_0^2 - r^2}{4\eta}\frac{\mathrm{d}p}{\mathrm{d}x}, \tag{15}$$

as is immediately obtained from (14) by integration under the condition of constant flow.

Now suppose that the viscous force upon a fluid element, which is of the order $\eta v/r_0^2$, is small with respect to the inertial force that this fluid element would experience in the event of its moving perpendicular to the laminar flow and which is of the order $\varrho v^2/r_0$, where v is the average velocity in the pipe. From the analysis of fluctuation phenomena in the fluid flow it appears that if indeed the latter inertial force is larger than the former viscous one, the laminar Poiseuille flow is unstable. Turbulent motion sets in, with more or less distinguishable vortex structure. The decisive quantity is the Reynolds number, *Re*, which is just the ratio of the forces mentioned:

$$Re = \frac{r_0 v}{\eta/\varrho}.$$

In other geometries some other typical dimension, l_{macr}, is characteristic, and always

$$Re = \frac{l_{\mathrm{macr}} v}{\eta/\varrho}. \tag{16}$$

The transition laminar→turbulent occurs whenever $Re > Re_{\mathrm{crit}}$. For certain geometries $Re_{\mathrm{crit}} \simeq 10^3$.

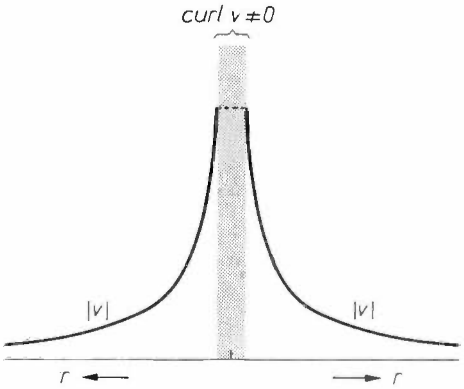


Fig. 3. The velocity field near a vortex core (shaded) in an ideal fluid. In the vortex core curl $\boldsymbol{v}$ is zero.

It may be useful to introduce the microscopic meaning of the viscosity. We may always write:

$$\eta = l_{\mathrm{micr}} v_{\mathrm{th}} \varrho, \tag{17}$$

where v_{th} is the thermal velocity of the atoms and l_{micr} a temperature-dependent quantity with the dimension of length, and having a value of the order of the interatomic distance. Now

$$\frac{v_{\mathrm{crit}}}{v_{\mathrm{th}}} = [c]\frac{l_{\mathrm{micr}}}{l_{\mathrm{macr}}}. \tag{18}$$

$[c]$ is a numerical constant about equal to the critical Reynolds number.

Relations like (18) will appear also in the sections that will follow. In the denominator we always encounter a fatal exertion which can destroy the essential order of

[2] J. A. Geurst, Physics Letters 36A, 63, 1971.

the system — in this case a velocity equal to the thermal molecular velocity. However, at a critical value that can be very small with respect to the fatal one, vortices are formed. The reduction factor is — apart from a numerical constant — the ratio between a microscopic length, which informs us on the situation at atomic scale, to a macroscopic length which is a typical dimension of the system seen at a macroscopic scale.

Now we return to Euler's equation and we consider the consequences of moving vortices in the ideal fluid. The velocity field may be not uniform, such that $\boldsymbol{v}$ changes as we proceed along a streamline. Then:

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{grad}\ \boldsymbol{v})\ \boldsymbol{v}. \qquad (19)$$

With the help of the identity

$$\mathrm{grad}\ \tfrac{1}{2}\, v^2 = \boldsymbol{v}\times \mathrm{curl}\ \boldsymbol{v} + (\boldsymbol{v}\cdot \boldsymbol{grad})\ \boldsymbol{v}, \qquad (20)$$

Euler's equation becomes

$$\frac{\partial \boldsymbol{v}}{\partial t} + \mathrm{grad}\left(\tfrac{1}{2} v^2 + \frac{p}{\varrho}\right) = \boldsymbol{v}\times \mathrm{curl}\ \boldsymbol{v}. \qquad (21)$$

For steady flow $\partial\boldsymbol{v}/\partial t = 0$. If we now consider the direction of flow, $\boldsymbol{v}$, we see from (21) that in that direction grad $(\frac{1}{2} v^2 + p/\varrho)$ has no component, because it is a vector perpendicular to $\boldsymbol{v}$, as indicated by the right-hand side. This is in fact Bernouilli's law: $\frac{1}{2} v^2 + p/\varrho$ is constant along a streamline.

It is, however, also interesting to consider a quasi-stationary situation, in which $\partial\boldsymbol{v}/\partial t \neq 0$, in particular when due to moving vortices in the fluid. To the extent that we take an average over a sufficiently long averaging period we may neglect $\partial\boldsymbol{v}/\partial t$. With this approximation we obtain, comparing two points A and B in the fluid:

$$\left(\tfrac{1}{2} v^2 + \frac{p}{\varrho}\right)_B - \left(\tfrac{1}{2} v^2 + \frac{p}{\varrho}\right)_A = \int_A^B [\boldsymbol{v}\times \mathrm{curl}\ \boldsymbol{v}]\, \mathrm{d}l. \qquad (22)$$

Suppose A and B are chosen in quiet regions where $v \cong 0$. It is seen from (22) that a pressure difference may exist between A and B if the right-hand side of (22) is a finite quantity, i.e. if the velocity field brings a net amount of vorticity per second over the path AB [3]; see *fig. 4*.

Equation (22) may be compared with similar ones in the case of electromagnetic induction, compare *fig. 5*. Between A and B a potential difference exists, equal to the average amount of flux transported per second over the path AB. In the situation depicted [4] we are dealing with a unipolar motor, but the amount of the flux handled per second is essential for both generators (dynamos) and motors.

Finally, we put the question whether the typical aspects of hydrodynamics are also found in heat conduction, particularly of solids. Generally speaking this will not be so. The most relevant phenomenon in this connection is heat conduction when phonon scattering at the surface of the crystal is dominant [5]. This regime requires very low temperature and almost perfect crystals. The particle character of the phonons then comes out very clearly, and a Poiseuille-like profile of the heat flow may be found, but no evidence has ever been found of the formation of heat vortices.

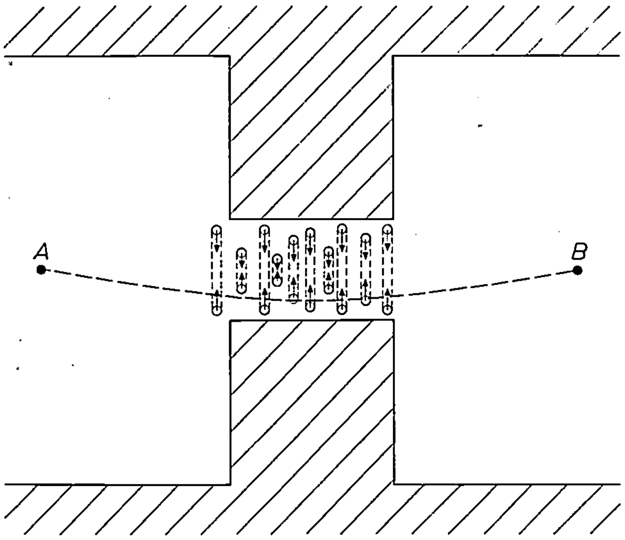


Fig. 4. Schematic representation of a current flowing in a pipe with a constriction. Here a steady generation of vortex rings is supposed, which shrink (see the arrows) and eventually vanish. Between A and B there is a pressure difference which corresponds to the steady passage of vortices over the line AB.

Dislocations

Having dwelt on rather classical grounds in the preceding paragraphs, we will now look around in a field that has come into being in this century — the field of dislocations in crystals. The vector fields now represent not a flow but a displacement or strain.

Dislocations cause profound changes in the organiza-

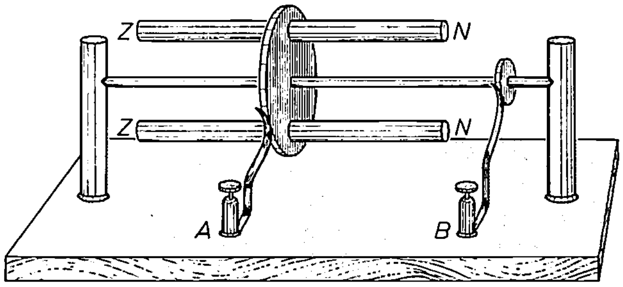


Fig. 5. Homopolar motor. The terminals A and B are connected through the brushes and the rotating copper disc in which two permanent magnets are mounted. A potential difference applied between A and B requires a steady transport of magnetic flux — i.e. areas of rotational current distribution in the magnetic rods — over path AB.

tion of a crystal. *Elastic* deformation involves only small displacement of the lattice points with respect to their position in the ideal crystal, and the displacement is a single-valued function of the coordinates. The displacement vector $\boldsymbol{t}$ is usually defined in the lattice points only, but one could perhaps even think of a continuous function $\boldsymbol{t}(x,y,z)$ if the wave-mechanical picture of the crystal with its nuclei and electrons is taken into consideration. At any site the symmetry of the lattice is only slightly disturbed. Everywhere a one-to-one correspondence between the atoms in the ideal crystal and those in the elastically deformed crystal is possible. This, however, is no longer so in the case of *plastic* deformation, when linear imperfections, the dislocations, occur. If we now select a closed path *1*, *2*, *3*, ... *N* in the real crystal and we trace it back in the ideal crystal — which should be possible since the atoms along the path are situated with respect to each other in an unambiguous way — we may find that the corresponding path *1'*, *2'*, *3'*, ... *N'* (see *fig. 6*) is not closed. In this case the circuit encircles a dislocation. We have to add a line element in order to close the circuit, and if only one dislocation is present, this line element is seen to be just one of the lattice vectors, irrespective of the circuit. It is called the Burgers vector, after J. M. Burgers, who treated dislocations in lattices theoretically long before direct experimental proof was available. We observe that the displacement vector defined in this way has a finite circulation.

When considering the deformation of a crystal we are not particularly interested in the actual displacement $\boldsymbol{t}$ (note that uniform translation or rotation does not lead to deformation). The really important quantities are its derivatives which follow from a comparison of the local actual unit cell with the ideal one and which characterizes the deformation. We have, for example,

$$\begin{aligned} &\text{the dilatation } e = \operatorname{div} \boldsymbol{t} \\ &\text{the rotation } \;\; \boldsymbol{\Omega} = \tfrac{1}{2} \operatorname{curl} \boldsymbol{t}. \end{aligned} \qquad (23)$$

These must be related to the internal body force (a force per unit of volume) and two proportionality constants must occur: the stiffness against compression and the stiffness against shear. The body force $\boldsymbol{f}$ can be derived from two potentials, like every vector field. We write:

$$\boldsymbol{f} = -\operatorname{grad} \phi + \operatorname{curl} \boldsymbol{\Psi} \qquad (24)$$

and find:

$$\begin{aligned} e &\propto \phi, \\ \boldsymbol{\Omega} &\propto \boldsymbol{\Psi}. \end{aligned} \qquad (25)$$

The displacement of the lattice points caused by the introduction of one edge dislocation (fig. 6) is considerable throughout the whole crystal. $\boldsymbol{t}$ has almost everywhere the direction of the Burgers vector $\boldsymbol{b}$. At large distances r from the dislocation line, t is proportional to ln b/r. The intricacies of the crystal symmetry which are related to the anisotropic binding forces that hold the crystal together, make the precise calculation of lattice deformation and internal forces around even a single edge dislocation rather complicated and in the case of a close dislocation network it is virtually impossible. The general trend is that the components of the deformation diminish inversely proportionally to the distance from a dislocation. This reminds us of the generating vector potential ($\propto$ ln $1/r$) and the magnetic field ($\propto 1/r$) at a current-carrying long wire.

Let us now discuss the generation and motion of dislocations, looking for physical phenomena that are similar to the ones we encounter in the physics of fluid vortices.

Dislocations will be set in motion when the local forces are increased to such an extent that atomic rearrangement must follow. Atoms are transferred to neighbouring sites and so are the dislocations, but the crystal as a whole does not yield. The process of plastic deformation is essentially due to the very existence or moving dislocations.

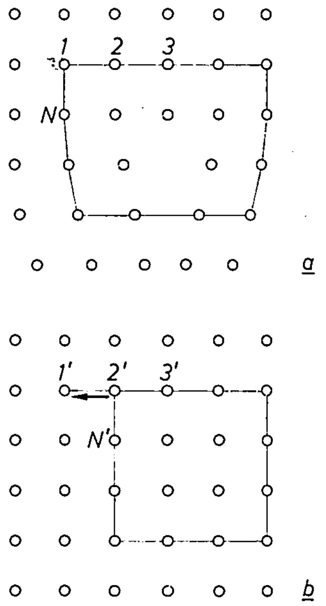


Fig. 6. The path encircling an edge dislocation in the disturbed crystal lattice (*a*), traced back in the ideal lattice (*b*). In the latter case the beginning and the end point do not coincide. The additional step which is required (arrow) is the Burgers vector.

Glasses exhibit a very high viscosity at low temperatures and they have no mobile dislocations as a means for plastic deformation; glasses therefore usually break rather than deform. We could perhaps distinguish two

[3] P. W. Anderson, Rev. mod. Phys. 38, 298, 1966.

[4] Gustav Wiedemann, Die Lehre von der Elektricität, 2nd edn., 1893/98.

[5] H. B. G. Casimir, Physica 5, 495, 1938.

kinds of solids in view of their behaviour under stress: solids of the first kind break without passing a plastic deformation phase, whereas solids of the second kind deform plastically due to the dislocation motions and only fracture if further action of dislocations is prohibited by secondary effects.

Dislocations multiply through the Frank-Read mechanism, starting from the initial dislocations that were formed during crystal growth. The dislocations form a network, dislocation elements being anchored in some way at their ends and bending out under shear stress in such a manner that eventually a new dislocation loop is formed and emitted [6]. Inclusion of dislocations is hardly avoidable during crystal growth [7].

Glide of a crystal without the interaction of moving dislocations would require a theoretical shear stress τ_{th} of the order of magnitude of the shear modulus. In actual fact the observed critical shear stress τ_{cr} is many orders of magnitude smaller, due to the favourable slip processes made possible by the dislocations. We have:

$$\frac{\tau_{\text{cr}}}{\tau_{\text{th}}} = [c]\,\frac{b}{l}\,, \tag{26}$$

where b is the Burgers vector and l the typical length of the basic dislocation elements in the Frank-Read sources mentioned; $[c]$ is a dimensionless constant of order unity, b/l is the ratio of a characteristic microscopic length to a macroscopic one just as we have met in (18).

Let us look somewhat more closely to the process of macroscopic deformation. Under steady-state creep conditions the glide velocity is determined by the concentration σ of mobile dislocations with Burgers vector b in the shear direction and their velocity v:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\Delta x}{x}\right) = b\sigma v. \tag{27}$$

The strain rate of a bar under stress is obtained from (27) by multiplication by a geometrical factor of the order unity, to account for the orientation of the slip planes with respect to the direction of macroscopic strain [8].

Now consider a special case — rather a *Gedanken*-experiment. Suppose that the shear occurs in a very orderly manner, such that only one dislocation at the same time is present in the glide region. Small dimensions of this weak region would probably favour the occurrence of this situation. Now a typical frequency f appears, viz. the number of times per second that a dislocation passes a given point. This frequency implies a kind of sonic modulation that is superimposed upon the linear shear motion and that should in principle be detectable with the help of fine mechanical or acoustical sensors. It seems likely that a tuned coupled resonator would be helpful. We would have:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Delta x) = f b. \tag{28}$$

One might also expect that application of a vibration would cause an anomaly in the stress *vs* strain rate curve just at the value of $\mathrm{d}(\Delta x)/\mathrm{d}t$ given by (28). A striking resemblance with the Josephson effect, which will be mentioned in the next section, may be observed.

The properties which crystals exhibit as a result of the existence of dislocations are manifold — but it is not intended to digress upon them. We should like, however, to make an exception for two recent investigations. The first one deals with the reflection of phonons at the surface of a crystal, a phenomenon that may govern the heat conduction of a crystal at low temperature, as has been mentioned earlier [5]. The reflection can be specular or diffuse. It turns out [9] that the diffuse variety can result from multiple inelastic scattering of the phonons in a thin layer just beneath the surface — some ten microns thick — if and when a considerable number of dislocations are present. One might prefer to speak in this case of absorption and thermalization of incident phonons, followed by re-emission from the layer.

Another investigation has shown that the thermal contact between liquid helium and a solid (e.g. the metal foil of a heat exchanger) is increased by the existence of such a layer of increased dislocation density just under the surface. The main point is the acoustical mismatch between both media — the phonon velocities and the densities differ by a factor of 20 and 50 respectively. Phonons therefore are subject to a large coefficient of reflection and, still worse, an unusually large proportion of the phonons suffer total reflection. But phonons that would be totally reflected, still penetrate the adjacent medium over a small distance as the so-called evanescent waves and may then yet be scattered by the dislocations present there — to the effect that a considerable reduction of net reflection obtains. This provides a reasonable explanation of the differences in thermal resistance of the surface (Kapitza resistance), found with different materials. In *fig. 7* recent calculations of the effect by H. Haug and K. Weiss are shown [10].

Superconductors

Vorticity in a current of charge carriers moving in an electrical conductor is — like vorticity in the phonon flow in heat conductors — evidently not easily accomplished, though it may occur in ionized gases on a gigantic scale, as appears for instance from photographs

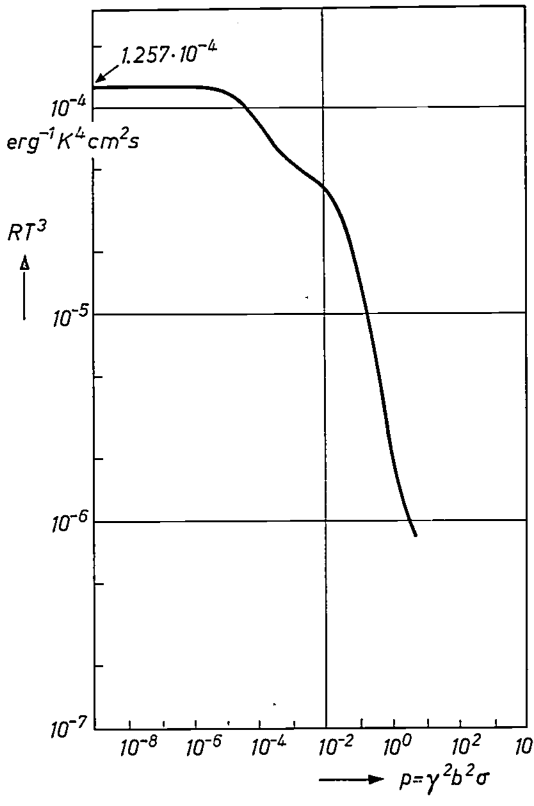


Fig. 7. The Kapitza (heat) resistance R, as calculated [10] for the case in which evanescent lattice waves are absorbed in a thin layer of dislocations at the surface of a copper foil. The product of R and the third power of the absolute temperature T is plotted against the quantity p which is proportional to the concentration σ of the dislocations (γ is a constant which is about 2 and b is the Burgers vector).

of the sun's surface. Plasma effects in well-conducting metals can hardly bring us to the concept of vortices as we discuss it in this survey. Such vortices, however, are found in optima forma in superconductors of the second kind. We shall first summarize [11] the main features of superconductivity.

The gas of conduction electrons in a superconductor has attained a certain degree of ordering. Through the interaction with the metal lattice, electrons form pairs of which the generalized or dynamical momentum

$$p = 2mv + \frac{2e}{c}A \tag{29}$$

vanishes, or has only a very small (everywhere the same) value in the case that a transport current flows in the specimen; v is the velocity of the centre of gravity of the electron pair. In the wave-mechanical treatment a superconductor is described by a wave function Ψ which is phase-coherent over macroscopic distances [*]. The situation is comparable to the case of an electromagnetic field or wave that is constructed from numerous photons and can nevertheless be coherent over macroscopic distances.

The wave function can be written as $\Psi = |\Psi| \exp i\phi$ and p is proportional to the wave vector in the Ψ-field:

$$p = h \operatorname{grad} \phi.$$

Since the curl of the gradient of any variable is identically zero, the momentum field is always irrotational in a superconductor:

$$\operatorname{curl} p \equiv 0. \tag{30}$$

This is the postulate that F. and H. London had already given as a basis for the electrodynamics of superconductors. Outside the region of superconductive ordering we may have $\operatorname{curl} p \neq 0$.

Equation (30) immediately leads to the first equation of London:

$$\operatorname{curl} j = -\frac{ne^2}{mc} H, \tag{31}$$

j being the density of the current carried by the superelectrons of concentration n, while H is the local field derived from a vector potential A. From (31) together with Maxwell's equations the Meissner effect is obtained, i.e. the phenomenon that H inside a superconductor vanishes, except in a thin skin, at the interface between the superconductor and the adjacent non-superconducting medium. The penetration is to a first approximation exponential, the characteristic penetration depth being

$$\lambda = \sqrt{\frac{mc^2}{4\pi ne^2}}\,. \tag{32}$$

In *fig. 8* we show by way of example the situation inside a superconducting rod in which a current flows, determining the same p-value for the whole volume. Note that (31) implies a rotational distribution of the current in the penetration layer. However, only in the situation discussed below we do speak of vortices or vortex lines in the superconductors.

Suppose that somewhere inside the superconductor a small non-superconducting region exists where $\operatorname{curl} p \neq 0$. Then that region where $\operatorname{curl} p \neq 0$ must extend along a line or tube which—as already remarked

[*] The maximum distance over which coherence in a superconducting wire can be established, has been discussed by H. B. G. Casimir [12] and is found to depend on the diameter of the wire. One could venture to suppose that crystallographic coherence in very thin wires over very long distances is also limited.

[6] H. G. van Bueren, Philips tech. Rev. 15, 246 and 286, 1953/54, or H. G. van Bueren, Imperfections in crystals, North-Holland Publ. Co., Amsterdam 1961.

[7] B. Okkerse and P. Penning, Philips tech. Rev. 29, 114, 1968.

[8] J. J. Gilman, Mechanical behaviour of materials at elevated temperatures, McGraw-Hill, New York 1961, p. 17.

[9] J. K. Wigmore, Physics Letters 37A, 293, 1971.

[10] H. Haug and K. Weiss, Proc. ICEC 4, Eindhoven 1972.

[11] J. Volger, Philips tech. Rev. 29, 1, 1968.

[12] H. B. G. Casimir, Proc. Kon. Ned. Akad. Wetensch. B 69, 223, 1966.

— can only end at the surface or end on itself. The value of p at any contour around this vortex line is obviously finite. The interpretation of p as wave vector leads us to a striking point in the theory; we see that as a consequence of the phase coherence,

$$\oint p \, \mathrm{d}l = nh \tag{33}$$

for a closed contour about the vortex line; h is Planck's constant and n is an integer. The relation (33) was already known in the atomic theory of Bohr-Sommerfeld and reappears in the treatment of the macroscopic quantum phenomenon that superconductivity is. With the non-superconductive tube a certain magnetic flux must be present. The accompanying ring current associated with this flux penetrates the adjacent superconducting body over the penetration depth only (*fig. 9*). The flux is quantized, as is apparent from (33):

$$\iint H \mathrm{d}S = \phi = n\phi_0, \qquad \phi_0 = \frac{hc}{2e} = 2.10^{-7} \text{ gauss cm}^2.$$

ϕ_0 is the elementary quantum of flux.

The generation of such flux lines is a complicated affair. A trivial case would be — of course — that we drill a fine cylindrical hole in the specimen and catch a certain magnetic flux in it. The more interesting case, however, is the spontaneous generation of flux lines at the surface and their penetration deep into the material during the magnetization process — such as is found with a large class of superconductors, called superconductors of the second kind. This phenomenon, treated in the theory of Ginzburg-Landau-Abrikosov, can be made plausible by referring to the theory of Gorter and Casimir [13]. In this theory the thermodynamics of the magnetization process was considered and for the first time the superconducting situation was interpreted as a different phase whose free energy F_{so} was smaller than the free energy of the normal phase by a certain amount. In a magnetic field, however, a term $H^2/8\pi$ per unit of volume must be added to F_{so}, due to the perfect diamagnetism of the superconductor, and so a critical field H_{c} can be derived from

$$H_{\mathrm{c}}^2/8\pi = F_{\mathrm{n}} - F_{\mathrm{so}}. \tag{34}$$

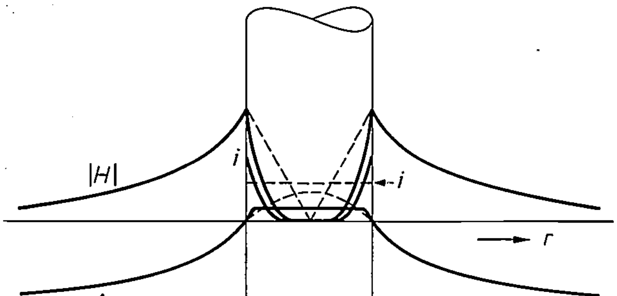


Fig. 8. The current density i, the vector potential A and the magnetic field H in the case of a long straight superconducting wire. Outside the wire H and A do not depend on the wire being superconductive or not, but inside they do. The dotted lines refer to the normal state. At the surface A has been chosen equal to zero. The penetration depth is exaggerated.

In fact, (34) denotes the energy necessary to depair the electrons.

When considering the equilibrium between two phases one may not forget the free energy of the boundary. The imperfect exclusion of the field at the skin of the superconductor diminishes the free energy. Therefore a fine mish-mash of normal and superconductive regions with a large interface must be formed provided that the positive contribution to the surface free energy, which is due to the decay of the ordering energy at the surface, is small enough. We encounter this situation with alloys in particular, where the penetration depth of the disordering — the so-called coherence [**] length ξ — equals the mean free path of the electrons and is therefore very small indeed. Magnetic flux penetrates as millions of vortices, the so-called mixed state, each vortex being in the lowest quantum state $n = 1$. The core of each vortex is a cylindrical region with a diameter of the order ξ, which is not superconducting.

Note that — as in the case of dislocations — the stiffness of the ordering is reduced in an energetically-economic way by the invasion of vortices.

Vortex motion

From a thermodynamical analysis of the mixed state follows the critical value H_{c1} of the magnetic field — and consequently also the critical value of the surface current — at which vortices begin to be emitted from the surface current. It appears that

$$\frac{H_{\mathrm{c1}}}{H_{\mathrm{c}}} = [c]\,\frac{\xi}{\lambda}, \tag{35}$$

in which $[c]$ is of the order unity. If we want to consider the corresponding critical value of the average drift velocity of the superelectrons, we find:

$$\frac{v_{\mathrm{cr}}}{\delta v} = [c]\,\frac{\xi}{\lambda}, \tag{36}$$

in which δv is the increase in velocity one has to give to an electron in order to overcome the pairing energy. Again we observe that the intrinsic limit of the "stiffness" of the ordering is reduced by the generation of a vortex structure. The reduction factor is the ratio between two characteristic lengths: in the denominator the thickness of the current layer, in the numerator a mean free path, *cf.* (18) and (26).

With superconductors the magnetic plasticity — that is the accomplishment of the mixed state — is a matter with two aspects — the one favourable, the other unfavourable. On the one hand the vortex structure allows us to retain superconductivity — as the mixed state — up to remarkably high and therefore technically interesting values of the magnetic field strength. On the other hand we must deal with the motion of flux lines when subject to interaction with a transport current through the superconductor. The current exerts a force upon the flux lines which makes them move perpendicular to the current. Fig. 4 shows a situation that also may occur in a superconductor of the second kind [14]. There will be a steady generation of vortex rings at the surface of the wire, which will then contract and eventually vanish again due to the current. This phenomenon, which appears under various aspects and is called flux flow, brings about a measurable potential difference along the wire. In fact it gives rise to a resistance, even in a superconductor! The value of the resistivity, i.e. the magnitude of the potential difference, may be derived by considering the induction phenomenon brought about by the moving internal flux pattern [15]. More interesting from the technical point of view are the so-called hard superconductors, which are superconductors of the second kind characterized by substantial irregularities in the build-up of the crystal e.g. fluctuations in the composition of alloys, fine precipitates of second phase, stress fields of lattice defects. These defects may act as pinning points for the flux lines, immobilizing them even if heavy currents pass the specimen. We touch here upon a technological problem akin to the hardening of metals and to Bloch-wall pinning in ferromagnetics. This problem has largely been solved, bringing the application of superconductors in heavy-current systems like those for power transmission nearer. However, large-scale application will depend upon reliability and efficiency of refrigeration systems, i.e. the progress in cryotechnology.

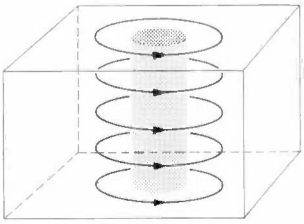


Fig. 9. A vortex in a superconductor. It consists of a threadlike normal region (shaded) at whose surface a circular current is flowing which penetrates over a short distance in the superconductive material.

Let us consider once more the phenomenon of flux flow. The statistical nature of the flux movement leads to noise, as has been found by various investigators [16]. An interesting case is found when the region where flux flow occurs, is only small. The total number of flux lines passing per second may then be found as a frequency, f, determined by

$$V = f\frac{hc}{2e}, \tag{37}$$

in which V is the voltage across the region, depending upon the current applied. With the help of a suitable detector one can indeed observe r.f. signals at frequency f [17]; see *fig. 10*.

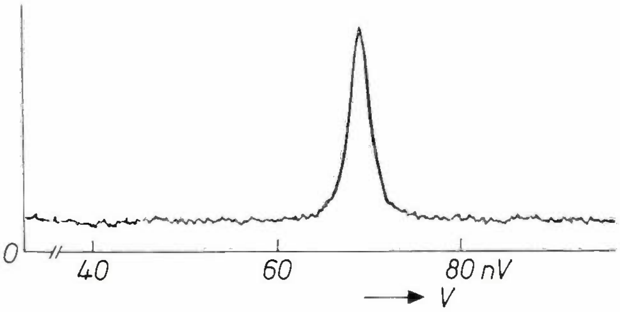


Fig. 10. R.f. signals detected at 34 MHz when a voltage V of about 70 nV is applied over a "weak link" in a superconductor of the second kind [17].

Particularly interesting and experimentally feasible is the case that the weak spot is so small that only one flux quantum at the same time can be present in it. Now a high frequency modulation upon the current occurs — the dominant frequency being again given by (37). This is none other than the famous dynamic Josephson effect [***]. A Josephson contact or weak-coupling device may be constructed as a small bridge or constriction in a thin metallic film, or as a solid-state diode with an extremely thin insulating layer (30 Å thickness is typical), through which by way of tunneling the superwave function can retain its coherence. Experiments on the emission of microwaves have been performed but the influence of absorbed microwaves upon the I-V characteristic has also been observed and studied, revealing anomalies at just that voltage that is given

[**] A rather confusing term, since coherence properly speaking — that is phase coherence — is preserved at very long distance.

[***] Harmonics and sub-harmonics are also present as a consequence of non-linearities.

[13] C. J. Gorter and H. B. G. Casimir, Physica 1, 306, 1934.

[14] W. F. Druyvesteyn and J. Volger, Philips Res. Repts. 19, 359, 1964.

[15] A. G. van Vijfeijken, Thesis, Amsterdam 1967 (also published as Philips Res. Repts. Suppl. 1968, No. 8).
H. B. G. Casimir, Physics Letters 17, 177, 1965.

[16] G. J. van Gurp, Thesis, Eindhoven 1969 (also published as Philips Res. Repts. Suppl. 1969, No. 5).

[17] R. K. Kirschman, H. A. Notarys and J. E. Mercereau, Physics Letters 34A, 209, 1971.

by (37). Recently T. D. Clark [18] has observed the synchronous emission (and absorption) of microwaves by a great number of Josephson contacts placed in an array.

The experimental skill necessary for successful work with the fragile weak-link devices has greatly advanced in the last few years. As a final remark we might mention that delicate current measurements in superconducting circuits are practicable with the aid of Josephson contacts. It need not be explained here, but it deserves mention in connection with the question of the physical meaning of a vector potential, put in the first section of this article. Consider a superconducting ring which encircles a long solenoid. Inside the solenoid an arbitrary magnetic flux can be adjusted. At the position of the ring the magnetic field as caused by the solenoid is zero. Nevertheless it appears from the measurements that a certain ring current is set up, corresponding to an additional flux through the ring which, together with the flux in the solenoid, makes the total flux just a multiple of $hc/2e$. It looks as if the vector potential can indeed be sensed even in field-free space. A careful discussion of this problem requires consideration of all energy terms that are involved in the system superconducting ring + solenoid + current supply [19].

Superfluid helium

Liquid helium (normal isotope ^{4}He) is below 2.1 K in an ordered state (He II) which is in many respects comparable to the superconducting state of metals. The atoms are partially condensed in a superfluid phase, and consequently the liquid shows no viscosity, at least for low velocities. Here too the physical situation is characterized by one wave function — or order parameter — which is phase-coherent over the whole volume of the liquid. This "stiffness" leads to the hydrodynamical property that everywhere curl $\boldsymbol{v}_s = 0$, except at the core of vortices which may be present in the superfluid. Again we have, as in the case of superconductors, the Bohr-Sommerfeld relation:

$$\oint p_s \, \mathrm{d}l = nh. \qquad (38)$$

The normal situation for vortices in superfluid helium is $n = 1$ and therefore the circulation K of the velocity field of a vortex is

$$K = \oint \boldsymbol{v}_s \, \mathrm{d}l = \frac{h}{m}. \qquad (39)$$

Vortices are extremely easily generated in He II and it would be perfectly justified to call it a superfluid of the second kind. In a tube or capillary with diameter 0.1 cm the critical velocity for vortex formation is only about 0.1 cm/s. The vortices are presumably vortex rings born at the inner surface of the tube. The hydrodynamics of He II is not yet mature, due to the complicated interrelation with the statistical and wave-mechanical aspects of this wonderful liquid, but some phenomena are already clear.

Vortices represent a certain amount of kinetic energy E of the flowing liquid (density ϱ_s). The calculation of the kinetic energy is similar to the calculation of the magnetic field energy near a wire. The diameter, $2a$, of the core of a vortex in He II is probably only a few Ångstrom. To a good approximation we have for vortex rings of radius r_0, assuming that $r_0 \gg a$:

$$E = \tfrac{1}{2} \varrho_s K^2 r_0 \left(\ln \frac{r_0}{a} + 0.33 \right). \qquad (40)$$

In the magnetic case [20]

$$E = \tfrac{1}{2} J^2 r_0 \left(\ln \frac{r_0}{a} + 0.33 \right). \qquad (41)$$

The constant term may have a value different from 0.33 if an inhomogeneous distribution of vorticity (current) within the core (wire) is assumed.

Vortex rings move spontaneously. They interact with the wall of the vessel in which the liquid is confined [21]. Even in an infinite fluid they would move because a vortex ring finds itself so to speak in its own velocity field. It is still a matter of further study to understand what precisely happens at the wall when a vortex ring is formed and gets its identity [22]. The critical velocity v_{cr} in a tube of diameter d seems to be, to a good approximation,

$$v_{cr} \cong \frac{K}{2\pi d}. \qquad (42)$$

If we assume the intrinsic maximum fluid velocity is v_0 at the core, (42) may be written as

$$\frac{v_{cr}}{v_0} = [c] \frac{a}{d}, \qquad (43)$$

where $[c]$ is of order unity. Again the ratio of a typical microscopic and a typical macroscopic length appears. Experimental investigations support (42) rather well but the phenomenon of vortex formation is very complicated and especially with regard to small vortex rings much work remains to be done.

Nobody, arriving at this point, will be surprised to learn that in superfluid helium, phenomena completely analogous to the dynamical Josephson effect in superconductors have been observed. A pressure difference i.e. a difference in liquid head H between two compartments filled with superfluid helium and connected through a very small hole (weak link), means that near the hole a steady stream of vortex rings must occur. The laterally transported vorticity is equal to a frequency, f, times the quantum of circulation, and equation (22) becomes therefore:

$$mgH = hf, \qquad (44)$$

in which g is the gravitational constant. The experimental situation is sketched in *fig. 11* [23].

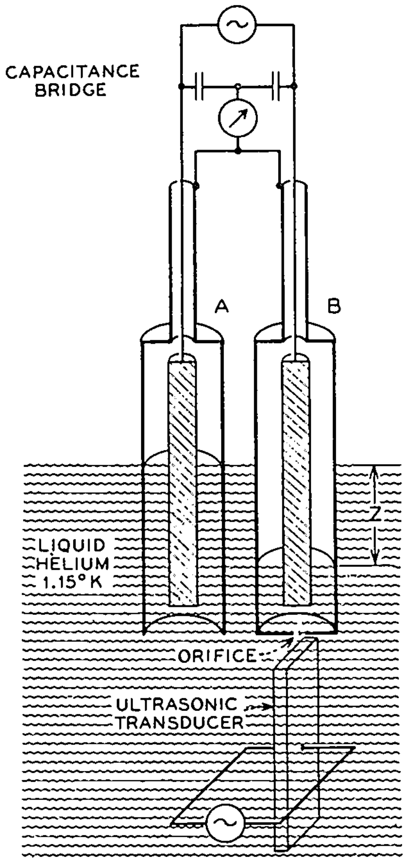


Fig. 11. Sketch of an experiment [23] with superfluid helium, the pressure head being determined from capacity measurements. The flow of helium through the hole is modulated by an ultrasonic source.

Equation (22) not only requires some finesse with respect to the vorticity, since this is quantized, but it is also necessary to use a more generalized form of the left-hand side. In the generalized formalism the thermodynamic potential μ comes instead of the hydrostatic pressure p. The driving force on the fluid particles is grad μ. This generalization is important, because we now can include temperature gradients in our considerations and in conclusion to this review we shall indeed do so.

The effects of vorticity in He II can be suppressed by bringing the fluid in a tube filled with fine porous material. Such a porous plug is called a superleak in the jargon of superfluids. Because the right-hand side of (22) now vanishes, we have a constant thermodynamic potential throughout the superleak:

$$\mathrm{grad}\,\mu = \frac{1}{\varrho}\,\mathrm{grad}\,p - S\,\mathrm{grad}\,T = 0\,, \qquad (45)$$

S being the entropy per unit of mass. Given a temperature difference over the superleak a pressure difference arises according to (45), as London already realized. This pressure difference may lead to a fountain effect. The forces that check the vortices in the porous structure are sometimes compared with the forces pinning the flux lines at the irregularities in hard superconductors.

To a good approximation one can take it that flow through a superleak transports only that part of the fluid which is condensed in the ordered or ground state, and this implies that *no entropy flow* is associated with it. On the other hand, according to (22) grad $\mu \neq 0$ in a capillary with *mobile* vorticity and this implies that also the non-condensed part of the fluid, i.e. the energy- or entropy-carrying part will be displaced. This state of things causes interesting thermal effects when a superleak and a capillary in which the critical velocity is surpassed, are connected in series. The situation resembles a Peltier element, i.e. a series connection of two conductors with a difference in relative position of the averaged energy levels of the charge carriers.

The fluid emerging from a superleak may be considered as being effectively at absolute zero. This property may be put to use. In Mendelssohn's experiment [24] helium purified of entropy in this way, is collected in a recipient and the content is indeed found to be colder (mechano-caloric effect).

However, if one wishes to make a continuously operating device, one must provide for a continuous removal of heat. For that purpose the vorticity in the drain — which mixes as it were the ground state and the excitations — is an essential feature.

An embodiment of the idea has recently been published by F. A. Staas and A. P. Severijns [25]. In *fig. 12* the essential part is shown. We have seen already that in the superleak

$$\Delta p = \varrho S \Delta T, \qquad (46)$$

[18] T. D. Clark, Physica **55** (Superconductivity, Proc. int. Conf. Stanford 1969), 432, 1971.

[19] H. B. G. Casimir, in: A. De-Shalit, H. Freshback and L. van Hove (editors), Preludes in theoretical physics, North-Holland Publ. Co., Amsterdam 1966.

[20] Lord Rayleigh, Proc. Roy. Soc. A **86**, 562, 1912.

[21] A. Walraven, Phys. Rev. A **1**, 145, 1970.

[22] A. G. van Vijfeijken, A. Walraven and F. A. Staas, Physica **44**, 415, 1969.

[23] P. L. Richards and P. W. Anderson, Phys. Rev. Letters **14**, 540, 1965.

[24] J. G. Daunt and K. Mendelssohn, Nature **143**, 719, 1939.

[25] F. A. Staas and A. P. Severijns, Cryogenics 9, 422, 1969, and Proc. ICEC 3, Berlin 1970, page 320.

in other words: dp and dT are proportional to one another, but in the drain capillary grad $p \cong 0$ and the fluid flows against a temperature gradient there. The apparatus is driven with the help of a fountain pressure — comparable to the use of the thermoelectric or Seebeck effect in electrical circuits. With this "vortex cooler", a device without moving parts, an interesting cooling power is attainable in the temperature region from, say, 2.0 K to 0.7 K.

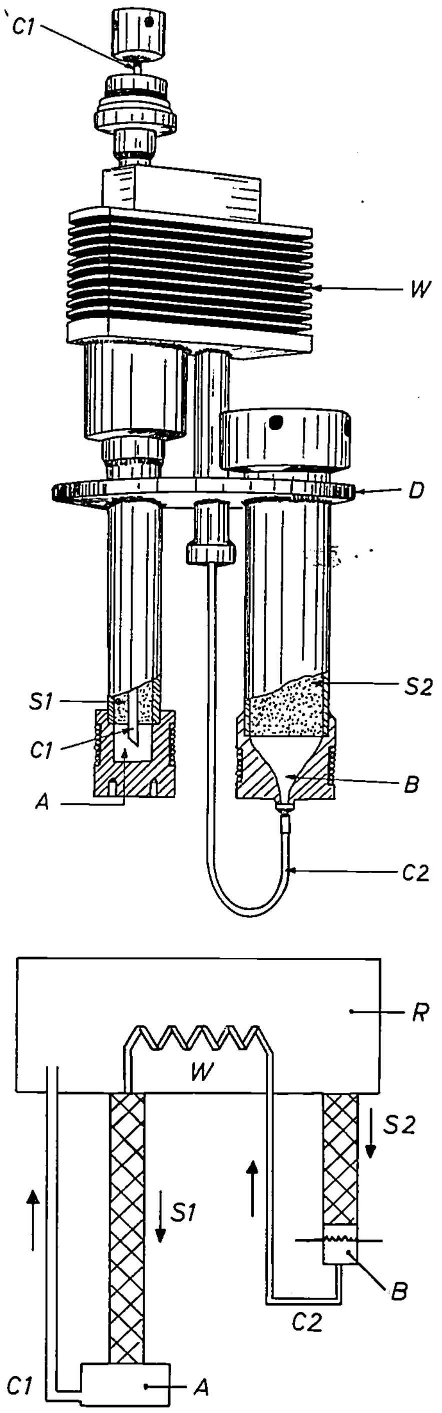


Fig. 12. The vortex refrigerator of Staas and Severijns [25]. The apparatus consists of two chambers A and B which are connected to a heat exchanger W and the helium reservoir R of a cryostat by the superleaks S_1 and S_2 and the capillaries C_1 and C_2. The size of the vortex refrigerator is 5 × 15 cm. It is mounted on either side of the bottom plate of the helium reservoir. The liquid helium in R is cooled down by reducing the vapour pressure to a temperature below the λ-point, e.g. to 1.5 K.

The chamber B can be heated electrically to a temperature of e.g. 1.7 K, resulting in a flow of superfluid helium from R to B caused by the fountain effect. Fluid that has become normal in B gradually flows to W via capillary C_2. It is converted there into superfluid again which can flow through superleak S_1, chamber A and capillary C_1 back to R. The velocity of the superfluid in C_1 is so high that vortices are created in the superfluid, setting up a pressure gradient that pushes the normal, entropy-carrying fluid away from A (i.e. heat flows from A to R). Since only superfluid helium (devoid of entropy) is supplied to A, the result is a cooling of chamber A. The temperature of A can be set between 0.7 K and the temperature of the fluid in R very accurately by varying the amount of heat supplied to B.

The Head Office of N.V. Philips' Gloeilampenfabrieken at Eindhoven, the seat of the Board of Management.

Philips Laboratories at Briarcliff Manor, N.Y., U.S.A.

Philips tech. Rev. 32, 261-280, 1971, No. 6/7/8

The physics of radiative centres in GaP

R. N. Bhargava

I. Introduction

An important byproduct of recent solid-state technology which has evolved in recent years is the "visible" solid-state lamp. Today, the most efficient visible solid-state lamps are fabricated from the III-V semiconductor GaP. In this article we shall review the physics and chemistry of those luminescent centres in GaP which are responsible for this efficient solid-state light emission. The physics of the luminescence consists in discovering and describing the mechanism of hole-electron recombination which leads to the generation of light. The chemistry consists in discovering which impurities in the crystals are important in the physical processes and determining how to add or remove them to produce luminescence of a particular colour. Before we discuss the physics and chemistry of luminescent centres in GaP, we shall briefly review the present status of various other means of generating light in solids.

Solids can generate visible light in several ways. For example, in incandescent lamps, electrical energy is converted to radiation by utilizing the intermediate step of heat generation. Emission of light in these incandescent lamps depends upon the temperature and size of the source (e.g. a tungsten filament). However, in electroluminescence devices, electronic energy can be converted directly into light without the added complexity of high temperature and vacuum sealing. The emission characteristics depend on the chemical composition of the host crystal and the incorporation of small amounts of special impurities. These impurities can give rise to centres which generate either useful light or non-useful light and heat; henceforth, these centres will be referred to as radiative or nonradiative centres respectively.

Electroluminescence was first observed by O. W. Lossew in 1923 [1] [*] in SiC, who observed that light originated within the crystal near a contacting electrode and later established that this luminescence was due to radiative recombination of charged carriers injected across built-in *P-N* junctions. Another type of electroluminescence was observed by G. Destriau [2] in which a ZnS phosphor was suspended in a liquid and an a.c. field was applied between the two electrodes immersed in the liquid. This type of electroluminescence is believed to result from the excitation of radiative recombination centres by inelastic collisions with thermally generated charge carriers that have been accelerated to high energies under the electric field. Other methods of injecting minority carriers in semiconductors to produce luminescence have been demonstrated which do not require a *P-N* junction [3]. All these devices have had little success in the practical world because of either too low light output or too short life when operating at the required brightness.

In 1962, the discovery of GaAs semiconductor lasers [4] in the infrared region and demonstration of highly efficient red-emitting GaP *P-N* junction devices [5] [6] generated intense interest in these light-emitting devices. Since then considerable progress has been made towards the improvement of the light output and in generating various colours in these devices. These *P-N* junction solid-state lamps are commonly referred to as Light-Emitting Diodes or LEDs. They enjoy the same advantages over the incandescent lamps that the transistor had over the conventional vacuum tubes, in that they are extremely reliable (half-life $\gtrsim 10^6$ hours), efficient (as high as 12% efficiency in the red), and require low power (milliwatts) compatible with solid-state circuitry. As such, they are being widely incorporated in new equipment having solid-state circuitry to perform many of the display functions such as indicator lights and alphanumerics-readout, which have heretofore been carried out with miniature incandescent lamps, numerical indicator tubes, and neon glow lamps. Several review articles have appeared in recent years concerned with junction luminescence [7-16]. This article will therefore deal chiefly with radiative and nonradiative centres in GaP, which play a significant role in these devices operating at room temperature. Stress will be on the recent developments pertaining to radiative centres which generate high luminescent devices.

In LEDs electroluminescence is produced by forward biasing a *P-N* junction of the semiconductor. In principle this is a very straightforward, simple process in which holes and electrons are driven together from *P* and *N* regions (see *fig. 1*). Recombination of these holes

Dr. R. N. Bhargava is with Philips Laboratories, Briarcliff Manor, N.Y., U.S.A.

[*] *The references are listed at the end of the article.*

and electrons can then result in the emission of photons with energy equal to or less than that of the energy-band gap of the semiconductor. For visible light the band gap should be greater than 1.8 eV (wavelength about 7000 Å). Recombination of the hole and electron leads to production of useful visible light or invisible light and heat. The internal quantum efficiency is the ratio of number of minority carriers (holes and electrons) that recombine radiatively (useful light) to the total number that recombine. After generation of a photon by radiative recombination, the photon may be lost by self-absorption and not escape from the solid. Thus, the most meaningful quantity in terms of efficiency, the external quantum efficiency, is the ratio of the number of *externally* emitted photons to the number of electrons that flow in the external bias circuit. Beside the internal and external quantum efficiency there is a concept which involves the product of the quantum efficiency and the response of the human eye which is referred to as luminous efficiency. For applications of LEDs, luminous efficiency is probably the most important property. Because of the human-eye response, diodes of external quantum efficiencies of 3 % in red and 0.1 % in green may look equally bright in a dark room because the eye is approximately 30 times more sensitive to green than to deep red.

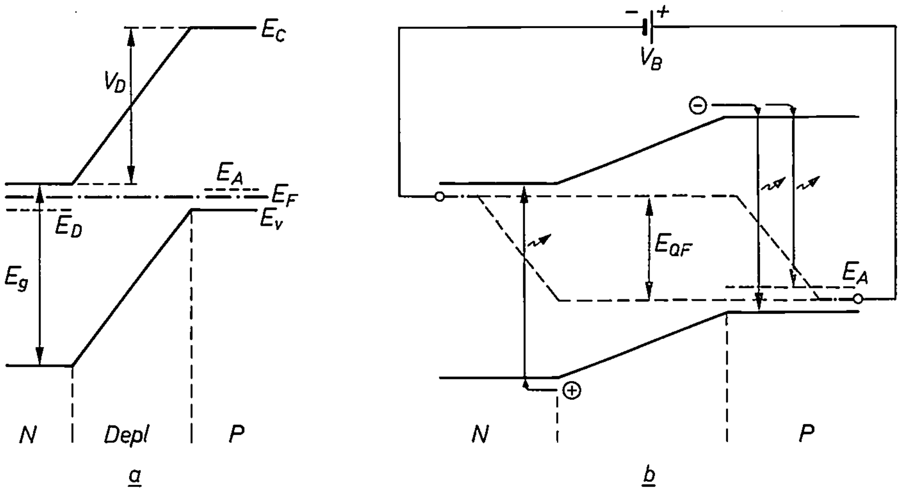


Fig. 1. Schematic representation of a semiconductor *P-N* junction in thermal equilibrium under zero bias (*a*) and under forward bias (*b*). V_D is the built-in potential; V_B is the applied bias; E_F is the Fermi level; E_A and E_D are the acceptor and donor levels and E_{QF} are the quasi-Fermi levels, whose energy difference is approximately equal to eV_B. Minority carrier electrons (holes) are injected to the *P* side (*N* side) of the junction and there recombine radiatively or nonradiatively with the available hole (electron). (From P. J. Dean [12].)

Luminescent properties of the bulk material are normally studied by photoluminescence, i.e. study of the luminescence from crystals excited by light with a photon energy greater than the energy band gap. For a good LED one needs both good photoluminescent and injection efficiencies but in this review article we shall only be concerned with one of the demands, namely the photoluminescent properties of the semiconductor. An alternative excitation process is by irradiative electrons, giving rise to cathodoluminescence.

Currently *P-N* junction devices which emit infrared and visible light are being made from semiconductors like GaAs, GaP, GaAsP, GaAlAs, GaInP and SiC. Infrared, red, yellow and green LEDs are available today in the market. During the past ten years considerable progress has been made towards the understanding of the physics and chemistry of luminescence in these semiconductors. The semiconductors GaAs and GaP have received particular attention primarily because of the high quantum efficiencies that have been achieved. The highest external quantum efficiencies in the GaAs (infrared) and GaP (red) diodes have been reported to be 32 % [17] and 12 % [18] respectively. The essential difference in the internal quantum efficiencies in GaAs and GaP is simply related to the radiative transition probabilities in a direct and indirect band-gap semiconductor. The discussion on this aspect will be postponed until the next section.

Electroluminescence in GaAs has found its importance in the development of a continuous solid-state laser operating at room temperature [19]. Furthermore, a GaAs diode can also be used in conjunction with rare-earth-doped upconverting phosphors to generate multicolor, visible solid-state lamps [20]. In these visible lamps, infrared emission from GaAs diodes is coupled into special phosphors which are capable of converting the infrared radiation into visible light. These phosphors are known as infrared-stimulated phosphors, and examples are fluorides and oxides of rare-earth elements (e.g. YF_3, YOF). The two promising features these infrared-visible converting diodes offer today are firstly that diodes have been made for the blue region of the visible spectrum as well as for the green and red regions, and secondly they provide some colour tunability in the visible region. However, these diodes have low efficiencies. The efficiency in these infrared-stimulated phosphors is low firstly because the infrared-to-visible-light conversion is brought about by absorption of two or three infrared photons to generate one visible photon, a process which has low transition probability, and secondly because the phosphors absorb about 10 percent or less of the emitted infrared radiation from the GaAs diode. Thus, to get enough brightness in these infrared-stimulated phosphors GaAs LEDs have to be driven rather hard (100 mA or more) and a rapid drop off in efficiency is observed as the excitation intensities are lowered.

Table I. Comparison of quantum efficiencies for various visible LEDs.

Material	Radiative transition: direct (d) indirect (i)	Peak emission		Quantum efficiency	Ref.	Luminous efficiency (lumens/watt)	
		H (eV)	λ (nm)			for 100% quantum efficiency	best measured values
$GaAs_{0.6}P_{0.4}$ (red)	d	1.91	649	2×10^{-3}	I	75	0.15
$GaAs_{0.18}P_{0.82}$ (yellow)	i	2.10	590	3×10^{-4}	II	~500	~0.15
$Ga_{0.65}Al_{0.35}As$ (red)	d	1.84	675	1.3×10^{-2}	III	~ 35	~0.45
$In_{0.4}Ga_{0.6}P$ (green)	d	2.17	571	2×10^{-4}	IV	610	0.12
GaP (red)	i	1.77	699	12×10^{-2}	V	20	2.4
GaP (green)	i	2.22	558	6×10^{-3}	VI	600	3.6
GaP (yellow)	i	2.05	605	1×10^{-3}	VII	~400	0.4
SiC (yellow)	i	2.1	590	3×10^{-5}	VIII	330	0.01
GaN (green)	d	2.4	515	1×10^{-2}	IX	~100	0.01

I. A. H. Herzog, W. O. Groves and M. G. Craford, J. appl. Phys. **40**, 1830, 1969.
II. Ref. [68].
III. J. M. Blum and K. K. Shih, Proc. IEEE **59**, 1498, 1971.
IV. B. W. Hakki, J. Electrochem. Soc. **118**, 1469, 1971, and private communication.
V. Ref. [18].
VI. Ref. [27].
VII. Refs. [65] and [52].
VIII. R. M. Potter, J. M. Blank and A. Addamiano, J. appl. Phys. **40**, 2253, 1969.
IX. Ref. [77]; please note that in this case a power efficiency of 10^{-4} is used to compute luminous efficiency.

Gallium phosphide on the other hand, can directly generate efficient visible light in the green, yellow and red regions of the spectrum. The luminous efficiencies of various available visible diodes are listed in *Table I.* The two best diodes are GaP-red and GaP-green, which respectively possess 2.4 lm/W and 3.6 lm/W luminous efficiencies. The discussion in the rest of the article will pertain to how these high luminescent efficiencies have been successfully achieved in GaP. In conclusion, we shall speculate on the quantum efficiencies achievable in the future.

II. Band structure

Gallium phosphide possesses satisfactory electrical and crystal-growth properties as well as a large enough forbidden energy gap (2.26 eV at 300 K) for generation of visible light. However, it is an indirect band-gap semi-conductor. In a direct energy-gap material (e.g. GaAs), the minimum energy of the conduction band and maximum energy of the valence band lie at the same crystal-momentum wave vector k. For an indirect-gap semiconductor (e.g. GaP), the minimum of the conduction band does not lie at the same value of k as the maximum in the valence band. For comparison, the band structures of GaAs and GaP are presented in *fig. 2a* and *2b*. Thus, for an electron in the conduction band to recombine with a hole in the valence band in an indirect semiconductor, a momentum equal to $k = k_c - k_v$ has to be accounted for since momentum is to be conserved: k_c and k_v are the electron and hole momenta at the conduction band and valence band respectively. Because of the restriction on the crystal momentum, the intrinsic recombination probability of the electrons and holes is low. However, the recombination of electrons and holes can be greatly enhanced by addition of impurities which interact strongly with free carriers, enabling crystal momentum to be conserved through an impurity-carrier interaction. We can thus think of the impurity inducing an efficient luminescence. It is just such an impurity-induced emission which produces efficient light generation in GaP.

Of the many possible impurity-induced transitions in

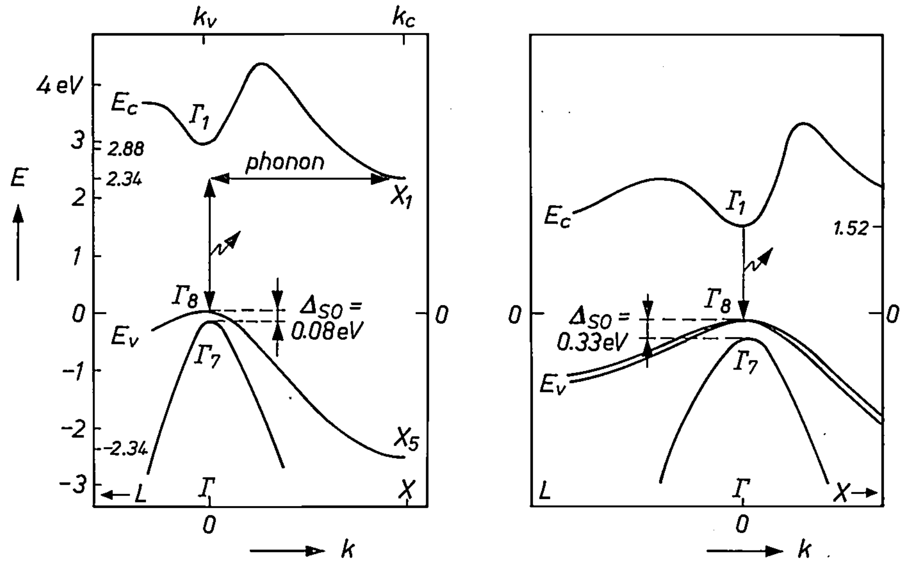


Fig. 2. The band structure of the III-V compound semiconductors GaP (*left*; indirect band gap X_1-Γ_8) and GaAs (*right*; direct band gap Γ_1-Γ_8). In the case of GaAs, the electrons in the conduction band recombine with the holes in the valence band to generate a photon, and momentum is automatically conserved. In GaP, the electrons recombine with the holes to generate a photon through a phonon-assisted transition so that the crystal momentum can be conserved. The energy difference Δ_{so} is the spin-orbit splitting in the valence band.

GaP, some simple examples are: (a) recombination of an electron trapped on a donor, and a hole trapped on an acceptor, (b) a trapped electron at a deep donor recombining with a free hole, (c) excitonic (an electron and hole pair bound together as a pair) recombination at "isoelectronic" and other neutral traps. An isoelectronic trap refers to a bound state produced by an isoelectronic substituent (e.g. N replacing P in GaP). These isoelectronic traps, which are electrically neutral, could either be a point defect or a molecule in the GaP lattice. Such a trap can capture an exciton resulting in efficient radiative recombination. The best example of a point defect is a nitrogen isoelectronic centre which replaces phosphorus in GaP (green diodes). A well-known molecule-type isoelectronic trap is the $\mathrm{Zn_{Ga}}^-$-$\mathrm{O_P}^+$ nearest-neighbour complex which is responsible for the visible-red emission in GaP. Such isoelectronic traps provide the greatest impurity-induced luminescence efficiencies at room temperature in GaP, and will be discussed in detail in sections IV and V respectively. In section VI we shall discuss the problems associated with the formation of the neutral nearest-neighbour complexes.

In addition to the radiative recombination at these desirable impurity centres, it is very instructive to study competing nonradiative recombination mechanisms, which dissipate the injected electron energy and hence reduce the light output. For instance, at 4 K, the efficiency of the donor-acceptor transition in GaP is close to 100%! That is, for each injected electron, we obtain close to one photon out on the average. However, at room temperature (300 K) the quantum efficiency is less than 0.01%. This drastic decrease in efficiency is related to the fact that trapped electrons and holes have an appreciable probability of being thermally released from acceptors and donors at higher temperatures combined with the relatively low transition probabilities of D-A-pairs [78]. Once back in the conduction band, they may find other nonradiative paths for recombination. Most of the energy released in exciton recombination at a neutral donor or acceptor goes into nonradiative processes. However, since the excitons at these centres are only very weakly bound, they are not expected to behave as nonradiative centres at high temperatures where thermalization is important. We shall postpone the important discussion of these nonradiative centres until section VII.

III. Donor-acceptor-pair emission

As mentioned earlier, the recombination mechanism involving simple donors and acceptors can be very efficient at low temperatures. Though they do not produce efficient light emission at room temperature, the understanding of the radiative recombination at these donors and acceptors has provided a great deal of information about the chemical nature of the donors and acceptors. We shall now discuss these recombination processes in more detail.

III-A. Shallow donor-acceptor-pair emission

Fig. 3 shows a photoluminescence spectrum taken at 1.6 K for S- and C-doped GaP. The emission is in the green region of the spectrum. The spectrum consists of a large number of sharp lines that culminate in a broad peak on the low-energy side. The understanding of the existence of all these lines from a simple system was provided by J. J. Hopfield, *et al.* [21]. As depicted in *fig. 4*, free electrons can be captured by positively charged (ionized) donors. Similarly, free holes can be captured by negative (ionized) acceptors. And, since the temperature is low, the electrons and holes will

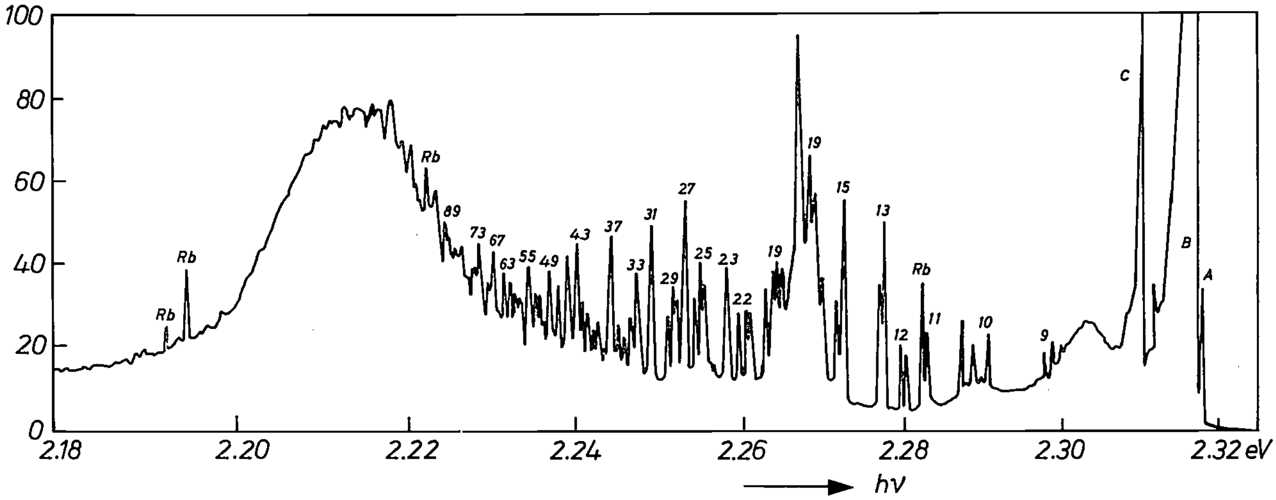


Fig. 3. Photoluminescence spectrum (intensity vs. photon energy $h\nu$) from nominally pure GaP at 1.6 K, depicting the donor-acceptor-pair spectrum. An electron bound to a *sulphur* donor recombines radiatively with a hole bound to a *carbon* acceptor. Different discrete separations between donor and acceptor lead to different lines. The numbers refer to the lattice shell number and are simply related to actual separations. The lines *Rb* are rubidium lamp-calibration lines; the line *C* arises from an exciton bound to a neutral sulphur donor; the lines *A* and *B* are due to a bound excitonic transition at the nitrogen isoelectronic trap. (From D. G. Thomas [8].)

stay put because there is not enough thermal energy to ionize them into the conduction or valence bands. However, this is only partially true, since electrons and holes have wave functions that spread out in the crystals away from the impurity atoms. Consequently, even though the donors and acceptors are separated from one another, there is a non-negligible overlap of wave functions which results in a finite probability of recombination of the electron and hole with the generation of a photon of visible light (see fig. 4). The emitted energy is equal to the band-gap energy minus the acceptor and donor binding energies plus a Coulomb term which, in its simplest form, is $e^2/\varepsilon r$:

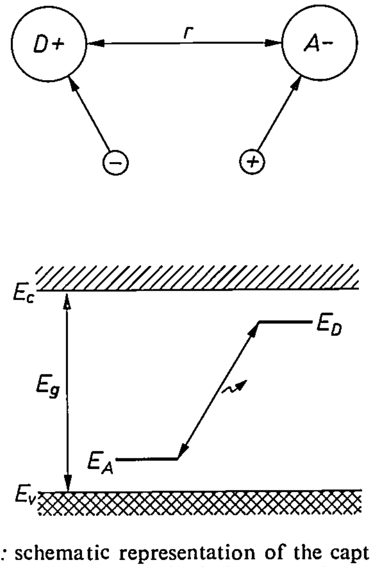


Fig. 4. *Above:* schematic representation of the capture of a free electron-and-hole pair at ionized donor and acceptor atoms respectively. *Below:* subsequent radiative recombination of the trapped particles results in the donor-acceptor-pair emission. The emitted energy $h\nu$ decreases with increasing r due to coulombic interaction. (From D. G. Thomas [8].)

$$E(r) = E_g - (E_A + E_D) + e^2/\varepsilon r. \qquad (1)$$

Here E_g is the forbidden-gap energy; E_A and E_D are the acceptor and donor binding energies; ε is the dielectric constant and r is the actual separation between acceptor and donor. This Coulomb term can be considered as a correction to the donor ionization energy since the electron is not removed to infinity.

A most important aspect of the above equation is that since the acceptors and donors occupy discrete lattice sites, the observed spectrum consists of many sharp lines corresponding to discrete various values of r. Thus, each line in the spectrum corresponds to recombination at donor-acceptor pairs with a particular separation. The intensities of these "pair" lines would be proportional to the number of pairs that can exist with a particular separation. Assuming that the donors and the acceptors are arranged randomly, the number of possible pairs does not increase regularly as the separation increases. The pattern of "pair" lines then can be used in establishing the shell numbers around the impurity, since one can obtain the exact value of r for a particular line. Once the value of r is known, one can plot the observed energy against the inter-impurity separator r and extrapolate to $r = \infty$, obtaining from eq. (1) $E_g - (E_A + E_D)$. Since E_g is known, if either E_A or E_D is known, the other can be accurately determined. This technique of obtaining E_A and E_D has been the most powerful way of determining binding energies. Another piece of information which can be obtained from the intensity pattern and the plot of energy against r is whether the impurities (donors and acceptors) are on different lattice sites or on identical lattice sites. This additional information, besides giving the arrangement of the impurities on particular lattice sites, gives the specific atoms they replace (e.g. Ga or P). The binding energies for various impurities so obtained are listed in Table V of the reference of note [14].

Thus, the analysis of the pair spectra, combined with knowledge of the incorporated impurities, gives details of the physics of the radiative recombination mechanism as well as the chemical nature of the donors and acceptors.

Since at 300 K most of the zinc acceptors are ionized, we expect that any infrared emission due to the O-Zn donor-acceptor pair should be absent. There are however a considerable number of free holes approximately equal to the number of substitutional zinc atoms, in the valence band at these temperatures, which have a wave function like a plane wave and, hence, can interact with the deep bound electrons at oxygen donors. This interaction results in a *bound* electron to a *free* hole recombination process [23]. The detailed analysis of how the spectrum changes in going from a donor-acceptor recombination to a bound-free recombination, is depicted in *fig. 6*. The peak emission at 1.35, as the

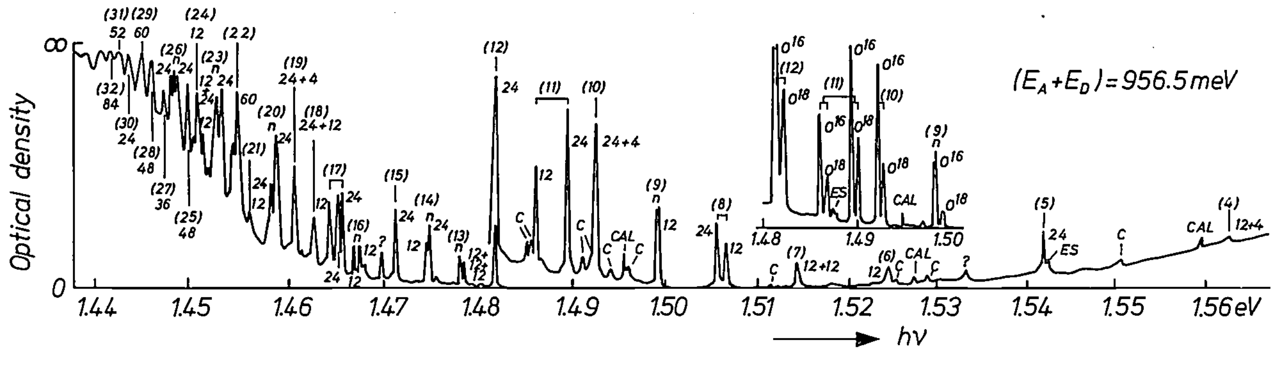


Fig. 5. Infrared luminescence spectrum in (Zn,O)-doped GaP at 1.6 K due to pair emission involving the deep oxygen donor and the shallow Zn acceptor. The bracketed integers are the shell numbers and other numbers denote the pair degeneracy, accounting for the inequivalent pair sites within a given shell. The inset shows the isotope shift when some of the O^{16} are replaced by O^{18}. (From P. J. Dean, C. H. Henry and C. J. Frosch [22].)

III-B. Deep-donor oxygen

Whereas the green luminescence due to shallow donor and acceptor pair emission is 100% efficient at 1.6 K, it quenches as we approach room temperature since the carriers ionize and recombine through non-radiative "killer" centres. However, if the donor or acceptor is a deep trap, the pair luminescence may persist with reasonably good efficiency at room temperature. An example of this type of luminescence is that which is due to the deep-donor oxygen. The oxygen level is approximately 0.90 eV below the conduction band at 1.6 K. At room temperature the donor energy $(E_D)_0 \approx 0.83$ eV, which is approximately 32 kT at $T = 300$ K. Consequently, the electron trapped at the oxygen donors has a very low probability of thermalizing at 300 K.

The deeply trapped electron at an oxygen donor can recombine with a hole at an acceptor at low temperature. The pair spectra with Zn as the acceptor are depicted in *fig. 5* [22]. The pair emission occurs in the near infrared region. From the analysis of O-Zn donor-acceptor-pair spectrum $(E_D)_0 = 0.895 \pm 1$ meV was derived [22].

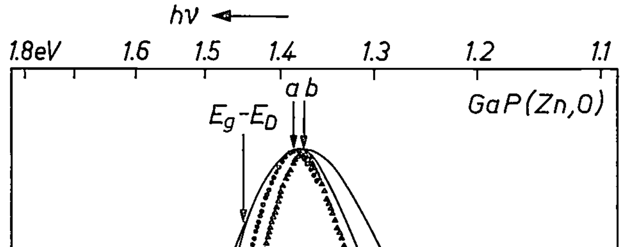


Fig. 6. Temperature variations of spectral characteristics of the infrared luminescence spectra in (Zn,O)-doped GaP, depicting the shift from donor-acceptor-pair emission to bound-electron to free-hole emission. The two arrows on the top of the peak signify a shift of peak opposite to that of the band gap when the temperature decreases from 160 K (dots; peak at *a*) to 77 K (triangles; peak at *b*) suggesting the change in the nature of the optical transition.

temperature is raised, shifts to the higher energy. This can be simply explained by expressing the peak emission for the donor-acceptor (*DA*) and bound-electron to free-hole (*BF*) transitions as

$$E_{h\nu}^{\mathrm{DA}} = E_{\mathrm{g}} - (E_{\mathrm{A}} + E_{\mathrm{D}}) + e^2/\varepsilon r - E_{\mathrm{phonon}} \qquad (2)$$

and

$$E_{h\nu}^{\mathrm{BF}} = E_{\mathrm{g}} - E_{\mathrm{D}} - E_{\mathrm{phonon}}, \qquad (3)$$

where $E_{h\nu}^{\mathrm{DA}}$ and $E_{h\nu}^{\mathrm{BF}}$ are the observed peak energies of the donor-acceptor and bound-to-free transitions; E_{phonon} represents the energy of the contributing phonons which is the same in both cases, and is due to the tightly bound electron at the deep donor. In the O-Zn system $E_{\mathrm{A}} = 64$ meV and $e^2/\varepsilon r \approx 40$ meV, which suggests that $E_{h\nu}^{\mathrm{BF}}$ should be shifted approximately 24 meV towards higher energy. The observed shift in fig. 6 has a similar value. Thus, as the temperature is raised from 77 K, the transition has changed from dominant donor-acceptor to dominant bound-to-free. This is also consistent with the lifetime measurements performed on the deeply trapped electron as explained below.

In a simplified picture the decay time of the luminescence due to the radiative recombination of an electron with a hole follows a simple exponential law, i.e. $I(t) = I(0) \exp(-t/\tau_{\mathrm{D}})$, where $I(t)$ is the intensity of the luminescence at a given time t and τ_{D} is the decay time. When $t = \tau_{\mathrm{D}}$, $I(t)/I(0) = 1/\mathrm{e}$. Hence the measured value of τ_{D} is referred to as the 1/e value. In case of donor-acceptor spectra, the recombination kinetics are dependent on the lattice spacing so that decay rates are different for different sites. The observed decay rate is thus a sum of many exponentials, giving rise to a nonexponential decay; for convenience however it is described by a τ_{D} of the 1/e variety. In the case of recombination of the bound electrons with holes, the time decay should follow a true exponential law. Thus, by examining the time decay of the luminescence, one can determine whether the recombination is due to a donor-acceptor pair or due to a bound-to-free transition. This information was also used to confirm that above 150 K the infrared emission at 1.35 eV is due to a bound-electron/free-hole transition [23].

Even though the infrared emission at room temperature is due to a deep electron recombining with a free hole, it is still not very efficient. This can be qualitatively explained by understanding the capture process for the conduction-band electrons into the ground state of the oxygen donor and is also due to the low concentration of these donors. Since the oxygen donor is deep, if an electron from the conduction band has to be captured into it, it has to lose energy equivalent to 17 optical phonons; the energy of one optical phonon is 50.1 meV. This suggests that the probability of electron capture into this deep trap is extremely small. However, the electron can be captured into an excited state of the donor and then be transferred to its ground state through radiative energy transfer. In fact, this has been observed for oxygen by P. J. Dean and C. H. Henry [24], and shown in *fig. 7*. The transition is referred to as an internal-capture luminescence in the oxygen donor. The lifetime of this internal-capture luminescence is measured to be approximately 20 μs [13] at low temperatures. If we assume that this capture time remains constant with increasing temperature, one anticipates the oxygen donors to readily saturate as a function of increased injection of electrons in the conduction band. In the regime of saturation the effective capture time would be at least 20 μs. If this is so, one could make an estimate of the quantum efficiency of the infrared emission by making a comparison with the quantum efficiency of the red zinc-oxygen complex emission. A capture time for the zinc-oxygen complex is approximately 20 ns.

We can conclude by comparison of the two capture times that the infrared emission should be approximately 10^3 times weaker than the red emission, which puts the infrared quantum efficiencies in the order of 10^{-4}. A further reason for the low infrared efficiencies is the low number of substitutional oxygen donors. Recent photo-capacitance data due to H. Kukimoto *et al.* [25] suggests that the number of isolated oxygen donors in *P*-type GaP is approximately 2×10^{16} cm^{-3}. Though the oxygen donor does not play a major role as a radiative centre, we shall see later that it has produced very interesting radiative and nonradiative centres in association with either impurities or crystal defects (vacancies). Consequently, the chemistry of the incorporation of the oxygen donor becomes quite important towards the understanding of the formation of radiative centres.

IV. Isoelectronic traps; nitrogen

The discussion in the last section described the luminescence properties of an oxygen donor, which is a deep trap for electrons in GaP. In this section we shall discuss luminescence due to another class of impurities which exhibit some properties of a deep trap without actually possessing a large binding energy and are referred to as "isoelectronic" traps. These isoelectronic traps were discovered in GaP in 1965 [26] and are crucial since they are capable of enhancing the recombination rate in an indirect-gap semiconductor. We shall discuss, mainly, why isoelectronic traps can generate efficient luminescence. The discussion will be related to nitrogen-doped GaP; nitrogen gives the simplest isoelectronic trap and is also responsible for the efficient luminescence in green GaP LEDs [27].

An isoelectronic *trap* is an isoelectronic sukstituent (impurity atoms which are in the same column of the periodic table as the host atom replaced) which introduces a *bound* state for either an electron or a hole. Once one carrier is bound, the centre is charged and the carrier of opposite sign is readily trapped through Coulomb attraction to form a bound exciton. For example, nitrogen occupies phosphorus sites in Gap and produces a bound state for electrons. Qualitatively, this can be understood since one expects nitrogen to be more attractive to electrons than is phosphorus, because of the more exposed nuclear charge of nitrogen. Thus, the electron is somewhat localized at the central nitrogen atom and hole attracted into a coulombic orbit. There is therefore a resemblance to an acceptor and the state is referred to as an "isoelectronic" acceptor [28]. A heavier-element isoelectronic with phosphorus might be expected to produce a centre attractive to holes; an electron would then be bound by coulombic attraction and an isoelectronic donor would result. This is the situation when bismuth replaces phosphorus. Arsenic, on the other hand, does not produce a bound state in the forbidden gap of GaP [13].

Whenever a particle becomes bound to an impurity centre, it becomes localized in real space. Since real r and momentum space k are conjugate quantities, the wave function of the trapped electron is then much more extended in k-space. This is schematically illustrated as a large effect for the electron in *fig. 8*. The extent to which the wave function spreads out depends critically on the type of binding. In case of a normal donor or acceptor where the binding is coulombic, the wave function spreads out very little [13], as depicted for the hole in fig. 8. The reason is that this is a long-range force and produces no abrupt changes in the wave function in real space. However, there are short-range forces which are not coulombic in origin. This is the case of N in GaP, where the bound state for an electron is believed to occur primarily from the large difference in electronegativity between the N atom and the P it replaces. The bare pseudo-potential difference suggests that the binding energy of an electron should be of the order of 1 eV. Hydrostatic deformation of the GaP lattice around the bare N atom will tend to reduce this binding energy. The result is that nitrogen will bind an electron with about 10 meV of energy [28-30]. The binding effect is felt only over one or two lattice spacings. Because of the highly localized distortions produced in that part of the electron wave function which is close to nitrogen atom, the electron wave function will extend throughout the k-space. Furthermore, due to the presence of a subsidiary minimum of

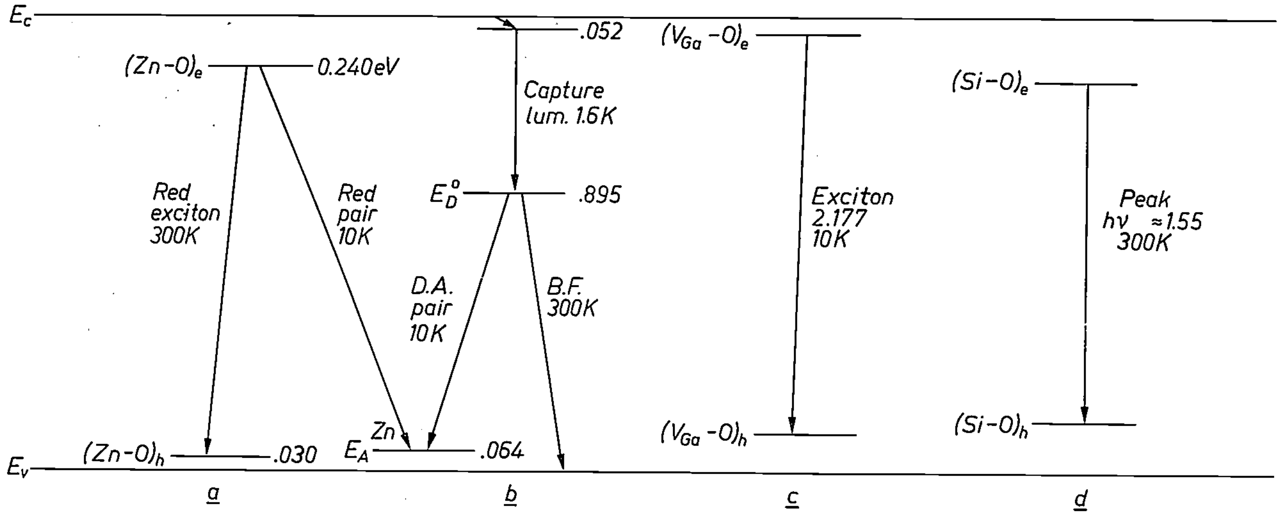


Fig. 7. Schematic representation of various radiative transitions associated with either an isolated oxygen donor or its associated complexes. *a*) The bound excitonic transition at a Zn-O nearest-neighbour complex. $(\text{Zn-O})_e$ and $(\text{Zn-O})_h$ are respectively the electron and hole binding energies. At low temperatures the electron can recombine with the hole at a zine acceptor (E_A^{Zn}) giving rise to a red "pair"-like spectrum. *b*) The capture luminescence in the infrared region of 0.8 eV due to an internal transition from an excited state to the ground state E_D^O of the oxygen donor. The electron from the ground state can either recombine with a hole at the zinc acceptor giving rise to a donor-acceptor (D.A.) pair emission or with a hole in the valence band to give rise to bound-to-free (B.F.) transition. *c*) The bound excitonic transition at a gallium-vacancy oxygen-donor complex (V_{Ga}-O_{P}) in the orange region of the spectrum. *d*) Infrared emission due to a nearest neighbour $\text{Si}_{\text{Ga}}^{+}$-$\text{O}_{\text{P}}^{+}$ complex, which predominately acts as a nonradiative centre.

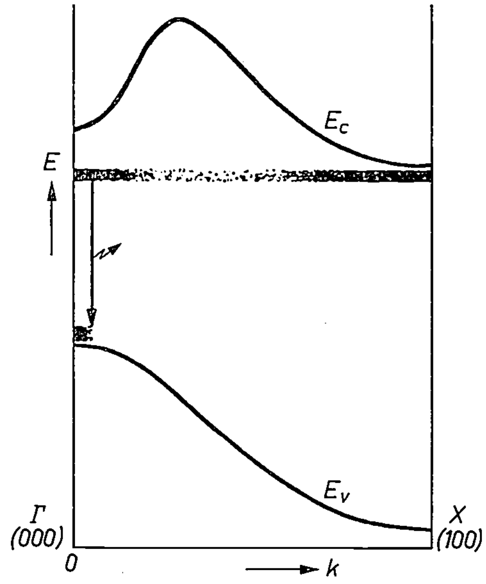


Fig. 8. A representation of the wave-vector (k) dependence of the probability densities (shaded area) of an electron bound to an isoelectronic trap (N in GaP) near the conduction band (energy at lower edge E_c) and of a hole bound to a shallow acceptor near the valence band (E_V at top). The spread of the wave function of the electron in the k-space shown, is responsible for the efficient luminescence in N-doped GaP. (From D. G. Thomas [15].)

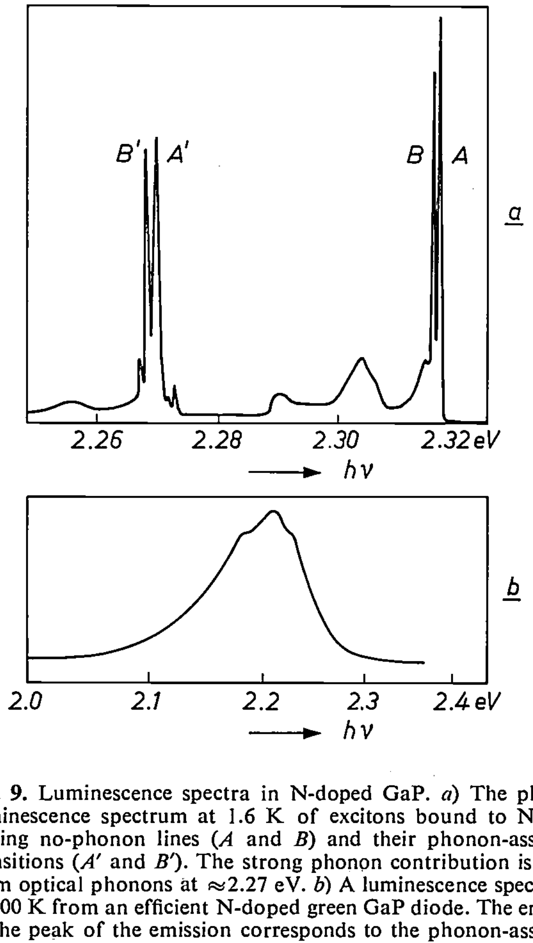


Fig. 9. Luminescence spectra in N-doped GaP. *a*) The photoluminescence spectrum at 1.6 K of excitons bound to N, depicting no-phonon lines (A and B) and their phonon-assisted transitions (A' and B'). The strong phonon contribution is seen from optical phonons at ≈2.27 eV. *b*) A luminescence spectrum at 300 K from an efficient N-doped green GaP diode. The energy at the peak of the emission corresponds to the phonon-assisted transitions of the line A. (From R. A. Logan, H. G. White and W. Wiegmann [27].)

the conduction band at $k = 0$ (zone centre), see fig. 2, this spreading is largely concentrated at the zone centre $k = 0$. This corresponds to the k value of the hole and so recombination can occur easily as illustrated. This explains why the isoelectronic trap nitrogen is effective in promoting radiative recombination near the band edge of GaP in the green region of the spectrum, see *fig. 9*. The spectrum consists of a sharp no-phonon line (referred to as the A line) at 2.3171 eV at 1.6 K and strong optical phonon contribution is depicted in fig. 9*a*. Since the binding energy of the isoelectronic trap is particularly sensitive to deformations of the central cell, the optical branch of the lattice vibrations contributes strongly. The spectrum in fig. 9*b* is from a green-emitting diode at 300 K [27]. The peak of the emission corresponds to an energy of the A line minus one optical phonon (LO). The maximum external quantum efficiency for green diodes achieved to date is 0.6% [27], which corresponds to a luminous efficiency of 3.6 lm/W.

One interesting fact to note is that the deeper bismuth isoelectronic trap, to which the hole is bound by 0.04 eV, does not produce efficient luminescence at 300 K. This is because the transition probability is much lower for excitons bound to bismuth than for those bound to nitrogen. This is due to the fact that it is the hole which is bound by short-range forces and is spread out in the Brillouin zone. The band structure of GaP has no maximum in the valence band at the point corresponding to the conduction-band minimum. However, at this point, the valence-band edge is several volts removed from the $k = 0$ (zone centre) valence-band maximum. As a result, there is much less opportunity for direct electron and hole recombination for bismuth than nitrogen traps. One should also note that for neither isoelectronic trap do the phonon wings emphasize the momentum-conserving phonons that are often thought to be the mechanism by which optical transitions acquire strength in indirect semiconductors.

V. Nearest-neighbour complexes

The generation of red light in (Zn,O)-doped GaP uses both of the required properties for the efficient generation of light, i.e. (a) the trap should be deep so that no or little thermalization of the bound particles occurs, and (b) the localization of the electron and hole occurs through formation of the bound exciton at an "isoelectronic" trap. How this is brought about is very interesting. As we have seen in section III, the oxygen donor is approximately 0.83 eV deep at 300 K, and generates useless emission in the near infrared. It is deep, not because of coulombic attraction, but because of the same type of short-range forces which produce the attraction with nitrogen, i.e. a large proportion of the bound-electron wave function is in the vicinity of the central cell. Consequently there is again a favourable spread of the electron wave function in k-space. To bring the binding energy into a useful range (from infrared to visible), it may be reduced by introducing a nearby acceptor with an opposite charge. In fact, this is what we find when the Zn atom is substituted on a nearest-neighbour Ga site. The negatively charged core of the Zn_{Ga} acceptor reduces the binding energy of the electron to 0.24 eV and generates an isoelectronic trap (complex) which is a molecule [31] [32]. This originally neutral Zn_{Ga}^{-}-O_{P}^{+} centre, after having captured an electron, will bind a hole in a coulombic orbit. The recombination of electron and hole results in a very efficient luminescence. External photoluminescent quantum efficiencies of approximately 85% at 77 K and approximately 21% at 300 K [33-34] have been reported. The best external quantum efficiency of red LEDs reported is 12% [18]. In the future, with better control and understanding of the P-N junction formation, we expect that the achieved electroluminescent efficiencies in LEDs will be approximately the same as the photo-luminescent efficiencies.

V-A. Identification of radiative centres

The identification of the red luminescence due to a bound exciton annihilation at a nearest-neighbour Zn-O complex has an interesting history. J. W. Allen and coworkers [5] showed in 1963 that on Zn- and O-doped GaP solution-grown platelets, alloyed P-N junctions gave diodes with efficiencies of about 0.01 to 0.1%. H. G. Grimmeiss and H. Scholz of Philips Forschungslaboratorium Aachen [6] in 1964 reported efficiencies as high as 1.5% but reproducibility was poor. The most promising method of junction formation in GaP was provided by M. R. Lorenz and M. Pilkuhn [35] in 1966, who extended the liquid-phase epitaxy method (referred to as LPE) first developed by H. Nelson [36] in 1963 to grow junctions in GaAs.

The formation of P-N junctions by liquid-phase epitaxy consists in tipping a solution of GaP, Ga and suitable dopants (say Zn and Ga_2O_3 for a P-type layer) on a GaP-substrate crystal of conductivity opposite to that of the above (i.e. N-type substrate using Te as dopant). Using a slow cooling process ($\approx$ 1-5 °C/min) from 1100 °C, one can grow thin layers ($\approx$ 50-100 μm) of the opposite conductivity in several hours (P layer on N substrate in the above case).

To date, the best results have been obtained for P-on-N tipping. In recent years, because of the availability of large single crystals of GaP grown by the Czochralski method [37], red diodes have been grown by a double liquid-phase epitaxial process, in which a Te-doped N layer ($N \approx 7\times10^{17}$ cm^{-3}) is grown on an N^+ substrate. Subsequently, a Zn- and O-doped P layer ($P \approx 3\times10^{17}$ cm^{-3}) is grown by LPE. Such a device has given the external quantum efficiency of 12% mentioned earlier [18].

Initial studies on the physics of the radiative recombination in (Zn,O)-doped GaP have suggested that the red luminescence is due to the distant donor-acceptor pair emission involving the oxygen donor and a zinc acceptor [38]. In this picture the oxygen donor was believed to be $\approx$ 0.47 eV deep and the high efficiency of the red emission was attributed to the deep bound electron. In 1967 this picture was proved incorrect [31] [32]. When chemically similar Cd was used in place of Zn, an entirely new spectrum was obtained, see *fig. 10a.* The spectrum consisted of a sharp A line at 1.907 eV at 20 K and more fine structure (phonon replicas) was observed as compared to the spectrum containing (Zn,O)-doped GaP shown in fig. 10*b*. The sharp no-phonon line allows one to perform precise experiments.

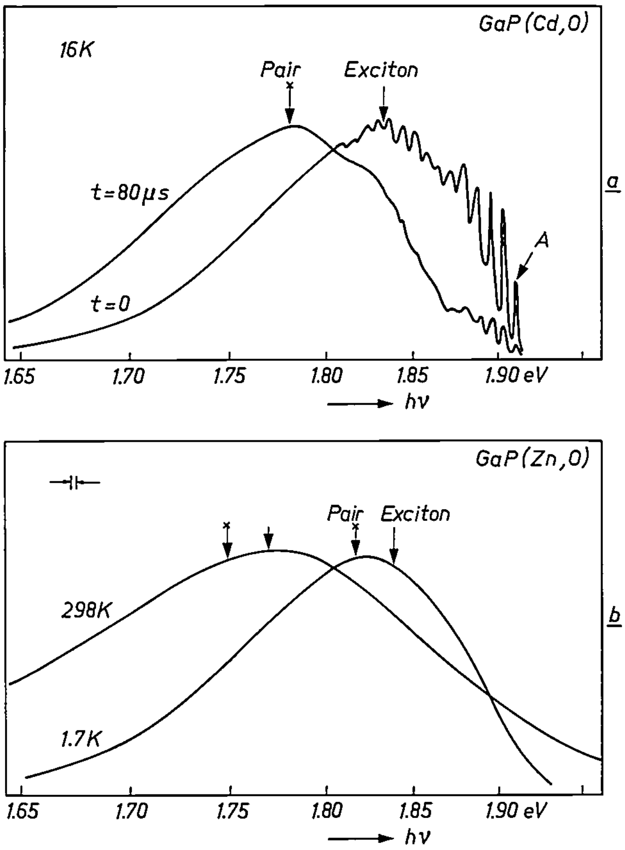


Fig. 10. Luminescence spectra from (Cd,O)- and (Zn,O)-doped GaP, due to bound excitonic and pair-like transitions from Cd-O and Zn-O nearest-neighbour complexes. *a*) The time-resolved spectra due to the Cd-O complex. At time $t = 0$, the structured luminescence due to the exciton recombination at Cd-O complex dominates. This emission consists of a no-phonon line A as well as local and lattice phonon-assisted transitions. At $t = 80$ μs a completely new spectrum appears which corresponds to a recombination of an electron at Cd-O complex with a hole at the Cd acceptor. *b*) The red spectrum in (Zn,O)-doped GaP at two temperatures. Though no fine structure can be seen, Zn-O emission possesses identical properties to Cd-O emission. The slit width is shown at the top left-hand corner of the figure. (See J. D. Cuthbert, C. H. Henry and P. J. Dean [39].)

From these experiments, which are described below, it has been shown that the red emission in (Cd,O)-doped GaP is due to a recombination of a bound exciton at a nearest-neighbour Cd_{Ga}^{-}-O_{P}^{+} complex.

1. An extremely powerful way of identifying the chemical impurities responsible for the emission is to perform an "isotope" experiment in which an isotope of the element in question differing from the natural isotope is added. When isotope O^{18} was added instead of normal O^{16} isotope, the sharp A line shifted by 0.7 meV to higher energies [31]. Similarly, when isotope Cd^{114} was added for Cd^{110}, the sharp A line did not shift significantly, but its phonon replicas did [32]. Without answering the difficult question as to precisely why these isotope shifts occurred, involvement of Cd and O was proven beyond doubt.

2. The relative orientation of Cd and O atoms could be determined from magneto-optical experiments [32]. At 1.6 K the hole and electron trapped at the centre freeze out from an energy level which corresponds to the A line in which the hole and electron spins are antiparallel, to an energy level 2.3 meV lower in energy in which the hole and electron spins are parallel. The transition corresponding to the lower energy is referred to as the B line. The Zeeman splitting of the B line is highly anisotropic. The data so obtained is consistent with a model in which Cd and O atoms are aligned along [111] directions, of which there are *four* in the crystal. If the magnetic field is along the [100] direction, all four become equivalent and only a simple pattern is observed. Since Ga and P atoms fall along [111] directions in GaP lattice, it is safe to say that Cd occupies a Ga site and O occupies an adjacent P site. The discrete-pair lines described in section III involving isolated O donors and Zn or Cd acceptors confirm the interpretation.

3. The nature of the recombination mechanism producing the emission could be verified further by studying the decay time of the luminescence. As discussed earlier, if the luminescence is due to a bound excitonic transition only, the decay is exponential. In case of Cd-O spectra, the decay was found to be nonexponential [39]. Furthermore, when time-resolved spectroscopy was performed (i.e. taking the spectra at different times after the exciting pulse had been switched off), one obtained two distinct spectra as shown in fig. 10*a*. This led to the interpretation that the deep bound electron at the Cd-O donor level can recombine radiatively either with a hole bound at the Cd-O complex or with a hole at the isolated Cd acceptors. (Schematically this is depicted for the Zn-O system in fig. 7.) Though the latter emission involved separated Cd acceptors, the emission is still different from the conventional donor-acceptor-pair spectra. The reason is that the Cd-O donor is neutral after recombination, and hence there is no Coulomb term in the equation for the energy of the emitted photon in eq. (1). This is also confirmed through time-resolved spectroscopy. Since in the bound excitonic transition the hole is localized, this emission should be dominant over the pair-like luminescence at room temperature.

4. Since the emission is due to a nearest-neighbour complex, thermal treatments can readily associate or dissociate the complex. The molecular bond between Cd and O or Zn and O in the GaP lattice can be broken around 900 °C and hence a fast thermal quenching from this temperature should decrease the red emission considerably. Conversely thermal treatments several hours long at lower temperatures (600 °C) should increase the number of complexes. In the case of Zn-O this has been confirmed [40]. We have used it to study extensively the kinetics of recombination leading to the red emission [41]. This type of study has enabled us to determine the relative strengths of the various competitive processes (both radiative and nonradiative). By definition a fully nonradiative process cannot be studied optically. However, the combination of thermal treatments with the recombination kinetic studies has in recent years yielded substantial information about them. We shall discuss this in some detail in the next section.

Just above we have discussed in detail the experiments performed on the Cd-O system. However, analogous properties are observed for the useful visible-red-emitting Zn-O complex, except for the sharp structure which is absent. In fact, the initial confusion in interpreting the red emission in (Zn,O)-doped GaP as due to conventional donor-acceptor-pair spectra, was due to two unexpected features of the spectra: firstly no sharp structure was seen, and secondly the pair-like spectrum coexists with the bound excitonic spectrum at a Zn-O complex.

V-B. Kinetics of (Zn,O)-doped GaP

The kinetics of radiative recombination of excitons bound to the nearest-neighbour Zn-O complex has been extensively studied in recent years [41-44]. We shall review some of the salient features of these studies and describe the principal results one obtains from it.

In the simplest model [41], recombination of electrons (the minority carriers) with holes (the majority carriers) in *P*-type (Zn,O)-doped GaP is depicted in *fig. 11*. The excited electrons in the conduction band can either be trapped at the Zn-O complex or disappear through other undesirable paths (henceforth referred to as nonradiative paths). The capture times into the Zn-O complex and thermal emission time out of the complex are given by τ_{nt} and τ_{tn}, respectively. The capture time

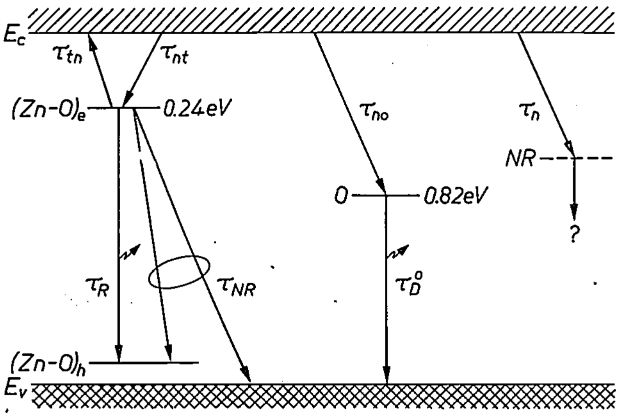


Fig. 11. Schematic representation of recombination of minority electrons in *P*-type GaP(Zn,O) at 300 K. The capture time into the Zn-O complex is given by $\tau_{\rm nt}$; the thermal emission time out of the complex is given by $\tau_{\rm tn}$. The capture time of nonradiative centres is represented by $\tau_{\rm n}$. The decay time of an electron out of the complex and into the valence band is given by $\tau_{\rm R}$ for radiative processes and $\tau_{\rm NR}$ for nonradiative processes. $\tau_{\rm no}$ and $\tau_{\rm D}^{\rm o}$ represent the capture time into and decay time of an electron from the oxygen donor O, respectively.

for nonradiative centres is represented by $\tau_{\rm n}$. The capture time for a particular impurity is inversely proportional to its concentration. The decay time of an electron out of the complex and into the valence band is given by $\tau_{\rm R}$ for radiative processes and $\tau_{\rm NR}$ for nonradiative processes. The measured decay time $\tau_{\rm D}$ of the excitonic emission depends on the number of Zn-O complexes and the number of nonradiative centres through thermalization of the electron bound to a Zn-O trap ($E_{\rm trap} \approx 0.24$ eV at 300 K) and can be expressed as:

$$\frac{1}{\tau_{\rm D}} = \frac{1}{\tau_{\rm R}} + \frac{1}{\tau_{\rm NR}} + \frac{1}{\tau_{\rm tn}(1 + \tau_{\rm n}/\tau_{\rm nt})}. \qquad (4)$$

The capture rates $1/\tau_{\rm nt}$ and $1/\tau_{\rm n}$ are respectively proportional to the number of radiative and nonradiative centres. The ratio of the concentrations of these two centres controls the quantum efficiency η, which is expressed as:

$$\eta = \left(\frac{\tau_{\rm D}}{\tau_{\rm R}}\right)\left(\frac{\tau_{\rm n}/\tau_{\rm nt}}{1 + \tau_{\rm n}/\tau_{\rm nt}}\right), \qquad (5)$$

where $\tau_{\rm R}$ and $\tau_{\rm NR}$ depend on the free-hole concentration p. By plotting $\tau_{\rm D}$ and η as a function of p, one can estimate the value of $\tau_{\rm n}/\tau_{\rm nt}$ as given in *fig. 12* and *fig. 13*. Two types of experiments have been done to get information on the ratio $\tau_{\rm n}/\tau_{\rm nt}$.

1. Crystals have been examined before and after the thermal treatments defined earlier, i.e. quenching and annealing. On annealing at 600 °C, one of the things that happen is that more Zn-O complexes are formed, resulting in an increased $\tau_{\rm n}/\tau_{\rm nt}$ ratio. The increased value of $\tau_{\rm n}/\tau_{\rm nt}$ results in higher efficiency and longer decay time as shown in figs. 12 and 13. On fast thermal quenching from ≈ 900 °C, $\tau_{\rm n}/\tau_{\rm nt}$ decreases, which corresponds to lower quantum efficiency and a low value of decay time. Thus precise values of the parameter $\tau_{\rm n}/\tau_{\rm nt}$ can be obtained.
2. If other foreign impurities are added resulting in degradation of the lifetime, one can conclude that a particular impurity has changed $\tau_{\rm n}/\tau_{\rm nt}$, provided that $\tau_{\rm R}$ and $\tau_{\rm NR}$ remain constant, i.e. free-hole concentration does not change. Such an experiment has been performed with chlorine dopants and has been shown that it does in fact degrade efficiency and gives a lower decay time [45] as shown in figs. 12 and 13. A further discussion of nonradiative centres will be taken up in detail in section VII.

From the preceeding data on decay time and efficiency, only $\tau_{\rm n}/\tau_{\rm nt}$ can be obtained. Several efforts [46] [47] have been made in measuring the total minority-carrier lifetime $\tau_{\rm L}$ which is defined through the equation:

$$\frac{1}{\tau_{\rm L}} = \frac{1}{\tau_{\rm n}} + \frac{1}{\tau_{\rm nt}}. \qquad (6)$$

When the thermalization time $\tau_{\rm tn}$ is much shorter than the electron-hole recombination time $(1/\tau_{\rm R} + 1/\tau_{\rm NR})$,

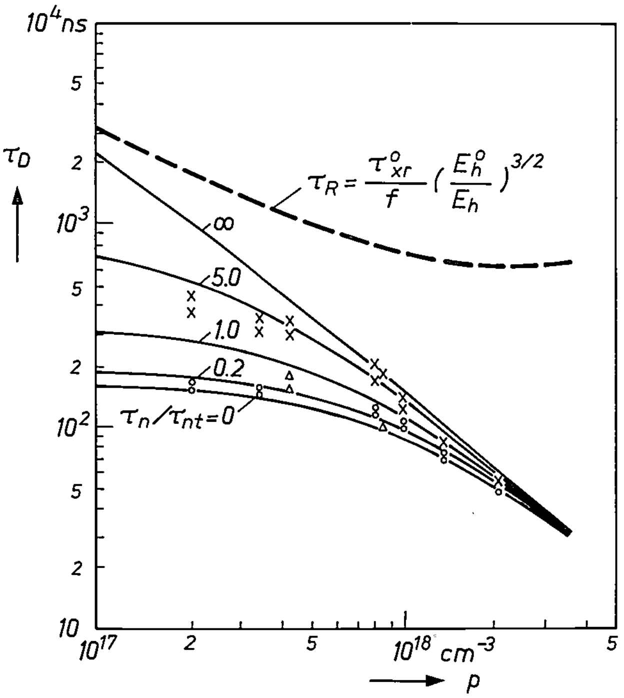


Fig. 12. The decay time $\tau_{\rm D}$ as a function of free-hole concentration p in (Zn,O)-doped GaP measured at 300 K. Predicted values of $\tau_{\rm D}$ are given by the solid lines using a value of the thermalization time $\tau_{\rm tn} = 161$ ns and various values of the parameter $\tau_{\rm n}/\tau_{\rm nt}$. (See J. S. Jayson, R. N. Bhargava and R. W. Dixon [41].) Heat treatment of a sample with gallium chloride brings down the value of $\tau_{\rm n}/\tau_{\rm nt}$ suggesting that nonradiative contribution in the sample has increased [45]. The crosses refer to a sample annealed at 600 °C, the circles to a sample quenched from 1000 °C.

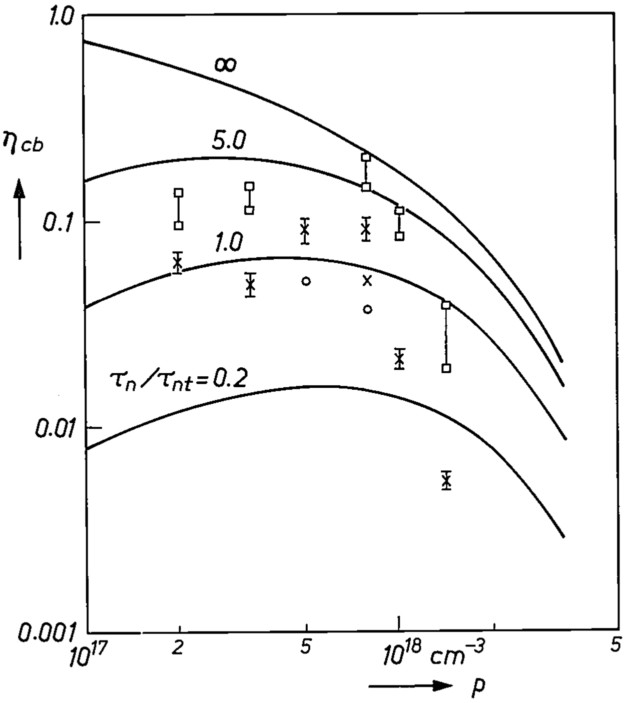


Fig. 13. The quantum efficiency η of the conduction band electrons as a function of free-hole concentration p measured at 300 K. Predicted values of quantum efficiency are given by solid lines using a value of the thermalization time τ_{tn} = 161 ns and for different values of τ_n/τ_{nt}. (See Jayson, Bhargava and Dixon [41].) Heat treatment of a sample with gallium chloride brings down the value of τ_n/τ_{nt} suggesting that the nonradiative path in the sample has increased. The squares represent the predicted internal efficiencies, the crosses and circles the measured external efficiencies for samples whose heat treatment took place whit gallium chloride absent (crosses) or present (circles).

the decay time τ_D equals τ_n. This occurs at reasonably high temperatures (≈ 500 °C). Under these conditions, τ_n has been measured recently [47]. (Such a direct measurement of τ_n yields the number of nonradiative centres which otherwise are too difficult to obtain through luminescence studies.)

It is clear that higher efficiencies could be achieved by either increasing the number of Zn-O complexes (radiative centres) or decreasing the number of nonradiative centres. In the next section we shall discuss how these nearest-neighbour Zn-O complexes are formed and how one can maximize their number.

VI. The gallium-vacancy/oxygen complex

It has been known for some time that the commercially viable vapour-growth process did not produce efficient luminescence in GaP. Recent spectroscopic studies [48-51] on the vapour-grown material have yielded some valuable information about this fact and have further answered the question of how these radiative nearest-neighbour complexes like Cd-O or Zn-O complexes are formed. In addition the formation of new radiative nearest-neighbour complexes like Mg-O [52] and the presence of certain nonradiative centres have been explained [49] [53].

VI-A. Identification

The discovery involved was the identification of a spectrum due to a bound excitonic transition at a nearest-neighbour complex of a gallium vacancy (V_{Ga}) and a deep oxygen donor at the adjacent phosphorus site [49] [50] (henceforth this complex is referred to as V_{Ga}-O_P). The spectrum from this complex occurs in the orange region and consists of a sharp no-phonon line at 2.117 eV at 1.6 K, designated as the line *A* in *fig. 14*. Additional structure is attributed to the lattice phonons and some weak vibrational modes. The intensity of the spectrum quenches rapidly with increasing temperature. The observed intensity of this characteristic orange luminescence (henceforth referred to as COL) is reasonably weak. The measured luminescent decay time for this emission is 110 μs at 4 K. The steps, which led to the identification of this COL luminescence as a bound excitonic transition at a V_{Ga}-O_P complex, are described below. The experiments were carried out on a water-vapour-grown GaP crystal which showed the strongest characteristic orange luminescence.

1. Annealing such a crystal at 400 °C for 15 hours in vacuum quenches the COL completely as shown in *fig. 15a*.

2. Indirect evidence for the existence of a V_{Ga}-O_P complex was obtained recently by Dean [48] in his study of the spectrum due to a complex Li_{Int}-Li_{Ga}-O_P; Li_{Int} is the lithium atom at an interstitial location in the lattice. When Li was diffused in these crystals at 400 °C for 15 hours, emission due to Li_{Int}-Li_{Ga}-O_P was observed. In this case the COL is completely quenched as

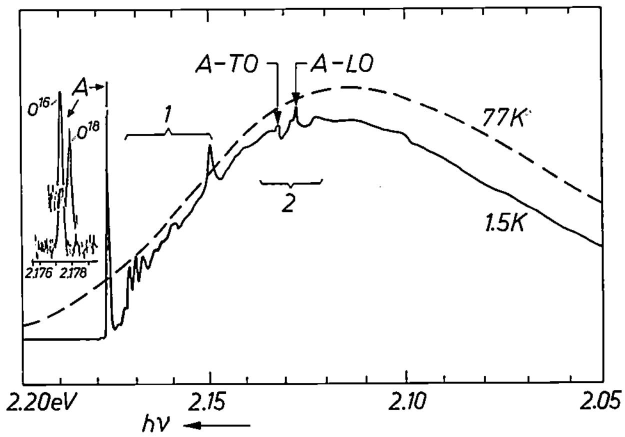


Fig. 14. Characteristic orange luminescence (COL) observed in non-intentionally doped water-vapour-grown crystals at 1.5 K and 77 K. The emission at 1.5 K consists of a sharp no-phonon line *A* and additional lines due to localized vibrational modes (*1*) and to lattice modes (*2*). The inset shows the isotope splitting of the line *A* when O^{18} is used instead of O^{16}.

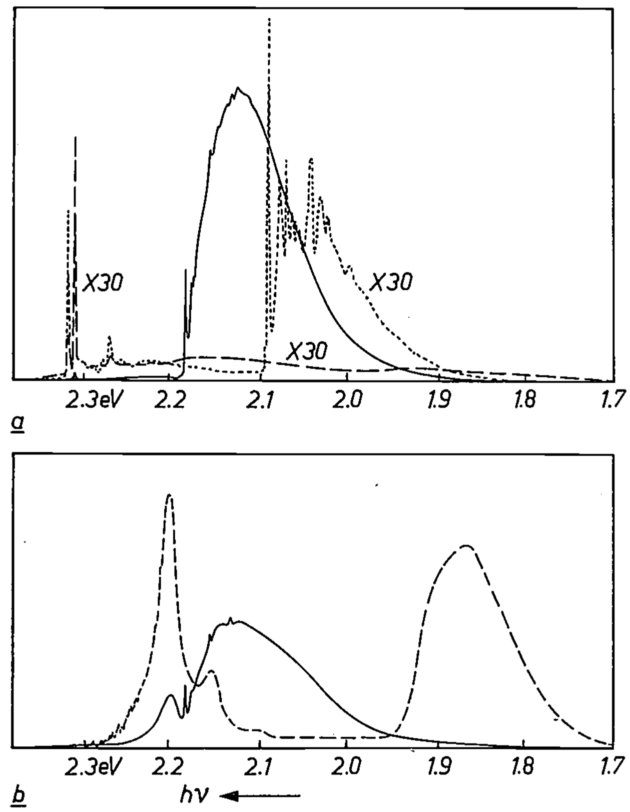


Fig. 15. The results of annealing on the photoluminescent spectrum of samples that show strong characteristic orange luminescence. *a*) An undoped GaP crystal grown in water vapour and annealed for 15 hours at 400 °C, in vacuum (dashed line) and in the presence of Li (dotted line). In vacuum the intensity of the COL increases significantly, whereas in the presence of Li the characteristic emission of Li_{Int}-Li_{Ga}-O_P is observed to be 30 times weaker than the COL. The solid line shows the spectrum of the sample before annealing. *b*) A Zn-doped GaP crystal grown in water vapour and annealed for 50 hours at 500 °C. The results show quenching of COL emission and simultaneous increase of red emission due to the Zn-O complex (dashed line). The solid line again shows the spectrum of the sample before annealing.

predicted by Dean's model [48] of Li replacing a V_{Ga} in a V_{Ga}-O_P complex. However, Li diffusion in already annealed crystals, in which the COL was completely quenched, gave at least 30 times less red luminescence due to Li_{Int}-Li_{Ga}-O_P complex. This was strong evidence that the COL is associated with the V_{Ga}-O_P complex.

3. Strongest confirmation on the role of oxygen was again provided by the isotope experiments [50] where O^{18} was used in place of O^{16}. The data in the insert of fig. 14 shows the isotope shift which gives irrefutable evidence that oxygen is positively involved in centres giving rise to the COL.

4. Magneto-optical data suggests [50] that the transition is due to a bound exciton and that the final state is a neutral double acceptor. In the neutral acceptor the spins of the two holes are paired off. Furthermore, it is determined that the tight binding of the exciton (by 160 meV) to V_{Ga}-O_P contains an appreciable contribution from the bound hole, unlike the Cd-O or Li_{Int}-Li_{Ga}-O_P complex, where the hole is weakly bound to the electron by a long-range coulombic attraction.

VI-B. Role of the V_{Ga}-O_P complex

Having confirmed that COL is due to an electronic transition at a V_{Ga}-O_P complex, we discuss below some of the aspects of the existence of such a complex which we believe to be important in the formation of some of the radiative and non-radiative complexes in GaP.

1. The elements of the Group II column of the periodic table such as Cd, Zn, Mg, and Be can fill the gallium vacancy of the V_{Ga}-O_P and produce nearest-neighbour radiative complexes (e.g. Zn_{Ga}^--O_P^+). This mechanism of formation is believed to occur because of the following reasoning. The charged state of the V_{Ga}-O_P is a double acceptor, i.e. $(V_{Ga}\text{-}O_P)^{--}$, as determined by the magneto-optical data on the sharp A line in COL [50]. The Zn atom is assumed to diffuse in GaP by an interstitial-substitutional mechanism. Furthermore an interstitial Zn atom is positively charged, i.e. donor-like. The coulombic attraction between Zn_{Int}^+ and $(V_{Ga}\text{-}O_P)^{--}$ can thus lead to the formation of a neutral complex Zn_{Ga}^--O_P^+; this process is described by the equation:

$$Zn_{Int}^+ + (V_{Ga}\text{-}O_P)^{--} \rightarrow Zn_{Ga}^-\text{-}O_P^+ + e. \qquad (7)$$

Thus the V_{Ga}-O_P complexes play the crucial role of precursors to the formation of radiative Zn_{Ga}-O_P complexes, and it becomes necessary to control and optimize the concentration of V_{Ga}-O_P complexes in order to have a control over the Zn_{Ga}-O_P complex concentration [51] [54]. This will be discussed later in section VI-C.

2. Another important aspect of the V_{Ga}-O_P complexes is their behaviour in Zn-doped water-vapour-grown crystals. The spectra before and after annealing are shown in fig. 15*b*. After annealing, strong quenching of COL occurs with simultaneous increase in the red Zn-O complex emission at 1.86 eV, indicating an increase in the number of radiative centres. However, the intensity of the green Zn-S donor-acceptor-pair spectrum also increases, suggesting that the V_{Ga}-O_P complex may be acting as nonradiative centres at low temperature.

3. Lastly, the quenching of the COL on annealing at temperatures of ≈ 400-500 °C can be attributed to the formation of centres like C_{Ga}^+-O_P^+ or Si_{Ga}^+-O_P^+ which act as nonradiative centres [45] [53]. How these centres act as nonradiative centres is discussed in section VII. However, the formation in these centres is guided by an equation (say for Si), similar to eq. (7):

$$Si_{Int}^+ + (V_{Ga}\text{-}O_P)^{--} \rightarrow Si_{Ga}^+\text{-}O_P^+ + 3e. \qquad (8)$$

Thus in all GaP crystals, both radiative centres like Zn_{Ga}-O_P and nonradiative centres like Si_{Ga}^+-O_P^+ are formed during the annealing process. Understanding the equilibrium constant for the formation of these various complexes from the V_{Ga}-O_P complex is therefore necessary to optimize the quantum efficiency of the material. Considerable work is in progress and we shall briefly discuss some of the important results obtained to date in section VI-C.

An important discovery which has resulted from the study of V_{Ga}-O_P complexes is that the element Mg can be incorporated in place of the gallium vacancy in the V_{Ga}-O_P complex, giving rise to new efficient yellow-orange emission, attributable to a bound excitonic transition at a nearest-neighbour Mg-O complex [52] [55]. The details of the optical properties of this emission are discussed in section VII.

VI-C. Formation of V_{Ga}-O_P complexes

The maximization of the number of V_{Ga}-O_P complexes is rather important for the formation of more Zn-O radiative complexes, as expressed in equation (7) in the last section. Since the characteristic orange luminescence is attributed to the bound excitonic transition at a V_{Ga}-O_P complex, its relative intensity should give the relative number of V_{Ga}-O_P complexes. We describe briefly the experiments which have been performed recently on water-vapour-grown GaP crystals [51] to maximize the number of V_{Ga}-O_P complexes, utilizing the spectral studies on COL.

Several pieces of the same water-vapour-grown GaP crystals were annealed at different temperatures ranging from 500 °C to 1000 °C and subsequently quenched. Annealing time varied from 2 hours at lower temperatures to 15 minutes at higher temperatures. The resultant photoluminescent intensity of the COL obtained in these crystals is plotted as a function of the quenching temperature in *fig. 16*. The data goes through a sharp maximum with activation energies of 3.3 eV and 1.25 eV respectively on the high- and low-temperature sides of the curve. It is interesting to note that the emission intensity decreases sharply on both sides of the maximum for two entirely different reasons: the low-temperature region is consistent with an impurity incorporation as given by eq. (8), and the decrease at high temperatures represents the dissociation of the gallium-vacancy/oxygen-donor complex. Understanding of both of these processes is rather crucial for the maximization of the quantum efficiency in (Zn,O)-doped GaP since one produces more nonradiative centres and the other destroys the associated complexes responsible for generating radiative centres. These concepts are being put into practice presently to increase the quantum efficiency in (Zn,O)-doped GaP.

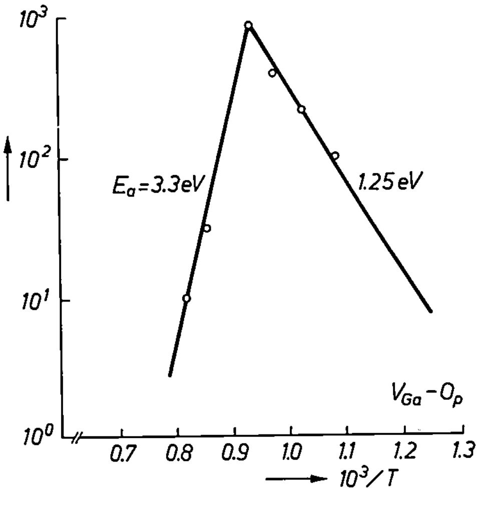


Fig. 16. Relative intensity of the V_{Ga}-O_P spectrum as a function of annealing temperature T. The activation energies E_a for high- and low-temperature reactions are 3.3 and 1.25 eV respectively.

VII. Mg-O nearest-neighbour complexes

Since an Mg acceptor is smaller in size and less electronegative than Zn, one anticipates that the bound exciton emission from a Mg_{Ga}-O_P complex would be shifted towards a higher energy than that from a Zn_{Ga}-O_P complex [52] [55]. Recently an efficient emission in the yellow-orange region (peak at 6030 Å at 300 K) has been observed in (Mg,O)-doped GaP [52] [55]. This emission is attributed to a bound excitonic emission at a nearest-neighbour complex [52]. Confirmation of the role of oxygen is again obtained from an O^{18}-isotope experiment [55].

This yellow-orange luminescence in Mg-diffused samples, at low temperatures, is shown in *fig. 17a*. The luminescence consists of two separate bands: a structured luminescence designated as spectrum A, and a structureless spectrum designated as spectrum B. Time-decay experiments (fig. 17*c*) and luminescence-excitation experiments have been used to derive the energy-level diagram in fig. 17*b*, and the two different spectra A and B are attributed to the radiative recombination of electrons at an Mg-O complex with either a hole at the Mg-O excitonic level E_h to produce the exciton spectrum A, or a hole at an isolated Mg acceptor level E_A to produce the pair spectrum B. This is similar to the dual "exciton" and "pair" spectra discussed in section V for Cd-O and Zn-O complexes. The ratio of Mg-O, A and B spectra depends critically on the ratio (Mg-O)/(Mg). However, as expected, the room-temperature emission is mainly due to the exciton transition.

It is important to note that on thermal quenching from 950 °C, the Mg-O complexes can be destroyed with simultaneous appearance of the COL originating at the V_{Ga}-O_P complex. This strongly suggests that formation of Mg-O is preceded by the formation of a V_{Ga}-O complex. As noted earlier, optimization of the V_{Ga}-O_P complexes is therefore an important step in the improvement of the efficiency of GaP light-emitting diodes which utilize isoelectronic nearest-neighbour complexes as radiative centres.

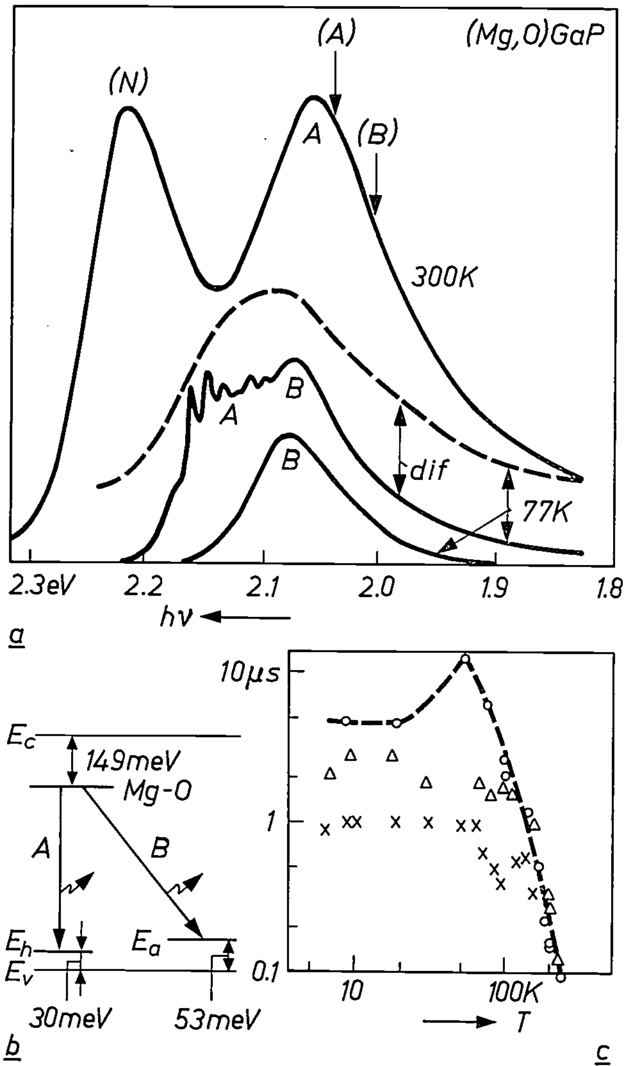


Fig. 17. The recombination properties of the electrons bound at the Mg-O complex. *a*) Photoluminescent (solid lines) and electroluminescent (dashed line) spectra for Mg-diffused (curves *dif*) and solution-grown crystals. For the diffused samples, both exciton spectra (*A*) and pair spectra (*B*) are observed, whereas in the case of the solution-grown crystals the pair spectrum *B* dominates. At 300 K the emission is due to exciton recombination (*A*). The vertical arrows depict the expected peak position of the two types of emission assuming that the temperature dependence of the Mg-O electronic level is the same as that of the band gap. The emission peak at 2.22 eV is an A-line transition due to nitrogen. *b*) Schematic representation of the energy-level diagram showing the "exciton" and "pair" transitions (*A* and *B*) respectively; E_h and E_A are the binding energies of holes at the exciton and acceptor levels respectively. *c*) The 1/e decay time is plotted against the temperature *T* for a solution-grown platelet (circles), for an LPE layer (triangles) and a diffused sample (crosses). The results are reminiscent of (Zn,O)-doped GaP under low-intensity excitation [43].

The external quantum efficiency of photoluminescence in (Mg,O)-doped GaP has been measured to be greater than 50% at 100 K. Diodes have been fabricated by growing a Te-doped *N* layer on solution-grown platelets. An external quantum efficiency of ≈ 0.1% has been measured at 77 K. Optimization of the room-temperature quantum efficiencies requires a further increase of Mg-O complexes near the junction and/or a significant reduction in the nonradiative centres which, in addition to the normal capture process, capture electrons that thermalize from the relatively shallow (Mg-O)-complex level.

VIII. Nonradiative centres

Nonradiative recombination mechanisms — those that release energy as heat (e.g. lattice vibrations) but do not produce light — are difficult to study since direct observation of this process is almost impossible. However, they play a very important role in the light emitting properties of indirect semiconductors. The quantum efficiency in GaP at room temperature is indeed limited by electron-hole recombination at these nonradiative centres, as discussed in section V-B. Consequently, a discussion of the nature of these centres is important, since their concentration must be reduced in order to achieve significantly higher quantum efficiencies in LEDs.

The most relevant nonradiative process in these semiconductors is the Auger process. In the Auger process, energy released due to electron-hole recombination, instead of being carried away by a photon, is transferred as kinetic energy to a nearby third particle (electron or hole). This third particle could be bound or free. Several cases have been studied and we discuss here three of these relevant cases which are separately depicted in *fig. 18*.

The first case relates to S-doped GaP in which S acts as a shallow donor. At low temperatures neutral sulphur can bind an exciton giving rise to an excitonic transition at 2.309 eV at 4.2 K commonly referred to as the C line [57]. The measured lifetime of the C-line luminescence is 21 ns. The radiative lifetime estimated from absorption and electrical measurements for the above transition is ≈ 11 μs. This amounts to a discrepancy of a factor 500 between observed and the estimated values for the lifetime. This discrepancy can be explained if the state not only decays radiatively, producing the C-line luminescence, but also decays nonradiatively by an Auger process. In this Auger process, the bound electron at the sulphur atom takes the energy released in the exciton annihilation (fig. 18*a*), this nonradiative process being 500 times more rapid than the radiative decay.

It is known that all neutral acceptors [58] [59] and donors [60] bind excitons at low temperatures and that the resulting emission is primarily nonradiative. However, these centres do not act as nonradiative centres at room temperature, because the excitons are already dissociated at this temperature.

The second type of Auger process is related to neutral centres such as the Zn-O complex (fig. 18*b*). In efficient *P*-type (Zn,O)-doped material, free holes are in abundance at room temperature ($\approx 5 \times 10^{17}$ cm^{-3}). Whenever the bound hole of the exciton recombines with the deep bound electron, energy could be transferred as kinetic energy to one of the free holes, resulting in an Auger process [61] [62]. To get a large number of Zn-O complexes, the concentration of Zn doping is required to be high [63]. However, at these high concentrations, the nonradiative process due to Auger recombination dominates [41] [42]. Hence, an optimum Zn concentration has been found. For the best diodes [18] it amounts to approximately 3×10^{17} cm^{-3}.

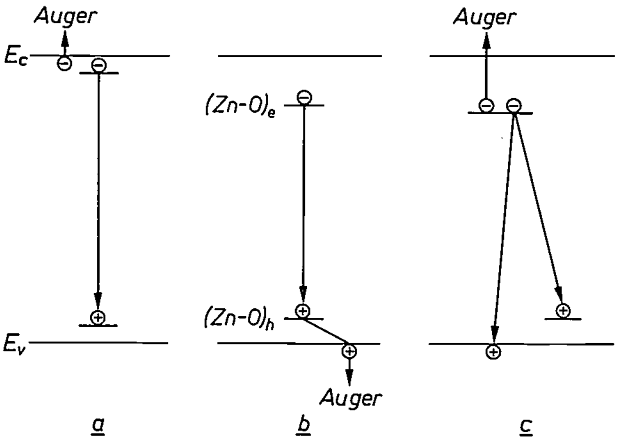


Fig. 18. Schematic representation of three types of Auger processes. In these processes energy released due to an electron-hole recombination is transferred as kinetic energy to a nearly third particle (electron or hole). *a*) Auger process associated with the excitonic transition at a neutral sulphur donor. *b*) Auger process associated with the bound excitonic transition at Zn-O complex, utilizing a free hole in the valence band as the third particle. An alternative process is when electron at Zn-O complex combines with the free hole and energy is transferred to the hole at Zn-O complex. *c*) Auger recombination process at a doubly ionized donor. The energy released in electron-hole recombination is transferred to the second bound electron at the donor.

An alternative approach to achieve higher efficiencies is to achieve a large number of Zn-O complexes, without doping with a large number of Zn atoms. Experiments in this direction are in progress [51].

One point concerning the two types of isoelectronic traps (i.e. N and Zn-O complex) is worth making. As we have seen, high concentrations of Zn result in predominantly nonradiative Auger processes. It is however fortunate that isoelectronic traps like N by definition do not add charge carriers to the crystals and hence do not promote nonradiative Auger recombinations — e.g. no impurity-band Auger effects [64]. The most efficient green diodes, in fact, contain a nitrogen doping greater than 10^{19} cm^{-3} and yellow nitrogen-doped GaP diodes contain approximately 10^{22} cm^{-3} N atoms [65].

Lastly, we discuss a group of nonradiative centres which are deep centres and could act as nonradiative centres at room temperature. These centres are doubly ionized centres, which are capable of binding two particles of the same type. Let us take for an example the case of Cl doping in GaP. Since Cl is a Group VII element it will replace P to generate deep doubly ionized donors [45] [66]. Normally, Cl^{++} atoms can capture two electrons. If there are free holes available, hole and electron recombination will take place. However, this energy does not result in photon generation, but instead can be transferred to the second bound electron at the chlorine atom (fig. 18*c*). Similarly, when inadvertent impurities like Si or C fill the V_{Ga} of the V_{Ga}-O_P complex, doubly ionized complexes like Si_{Ga}^+-O_P^+ and C_{Ga}^+-O_P^+ result. These centres are again doubly charged and will promote strong Auger recombination [53] [67]. Certain properties of these doubly ionized centres are summarized below.

(a) Since these centres are doubly ionized, the capture cross-section of such centres is expected to be large [67], i.e. they can control the recombination kinetics in GaP.

(b) Impurities like Cl and Si are inadvertently present in most GaP crystals, and in fact may be major factors currently limiting the observed quantum efficiency. Their removal should result in higher observed quantum efficiencies, both in diodes and in bulk crystals.

An interesting comment on the various types of growth processes used for making GaP diodes should be made here. Recently, diodes have been made by vapour-phase epitaxy [68] and liquid-phase epitaxy [69] in which green luminescence is attributed to the radiative recombination of the free exciton and its phonon-assisted transitions. In these diodes it seems that the low quantum efficiencies as well as the luminescence decay time are limited by the presence of the nonradiative centres. In both types of diodes a luminescence decay time of 400 ns has been observed, which possibly suggests that the concentration of nonradiative impurities in vapour-phase and liquid-phase epitaxially grown GaP is approximately the same.

IX. Concluding remarks and future developments

As can be seen from the preceding discussion of radiative centres, the recent improvements in quantum efficiency and range of colour give a promising future to GaP LEDs. However, one question remains to be answered: "What is the theoretical upper limit on the external quantum efficiency in this indirect semiconductor?" Stated in practical terms, "Can we ever hope to produce a solid-state lamp capable of generating lumens of optical power?"

Thus far no theoretical predictions have been made based on the physics of the radiative recombination at localized centres in GaP. Because of the importance of this question, we shall attempt a semiquantitative answer [70] in the case of radiative recombination at the Zn-O complex, which is the primary centre discussed in this article.

To achieve the higher quantum efficiencies in (Zn,O)-doped GaP, one requires that the ratio of the radiative centres (in this case the number of Zn-O complexes) to nonradiative centres be as high as possible. The reduction of the nonradiative centres is achieved (a) by reducing Zn concentration to avoid Auger processes due to free holes on the *P*-side, and (b) by reducing nonradiative impurities like C^+-O^+ or Si^+-O^+, etc.

Presently, the number of Zn-O complexes, $N_{\text{Zn-O}}$, achieved is approximately 3×10^{16} cm^{-3} for the Zn concentration, N_{Zn}, to be $\approx 3\times10^{17}$ cm^{-3}, i.e. $N_{\text{Zn-O}}/N_{\text{Zn}} = 0.1$. By understanding the thermodynamics of the formation of these complexes, one can hope to achieve $N_{\text{Zn-O}}/N_{\text{Zn}} \approx 1.0$. Initial calculations in this direction [71] suggest that to maximize $N_{\text{Zn-O}}/N_{\text{Zn}}$, one should incorporate Zn after the number of V_{Ga}-O_{P} complexes has been maximized. Furthermore, nonradiative centres like Si^+-O_P^+ or Cl^{++} can be avoided by growing crystals without these unwanted impurities. Alternatively, impurities like Si^+-O^+ can be minimized by appropriate heat treatment [56].

Assume that one can obtain $N_{\text{Zn-O}}/N_{\text{Zn}} = 1.0$ and a considerable reduction in the nonradiative impurities by appropriate growth conditions and thermal treatments. Then, in our model as discussed in section V-B, this would correspond to $\tau_{\text{n}}/\tau_{\text{nt}} \gg 1$, and from the estimated variation of τ_{D}, τ_{R} and τ_{NR} as a function of free-hole concentration (fig. 12), we predict $\tau_{\text{R}} \approx 3000$ ns and $\tau_{\text{D}} \approx 1500$ ns for $N_{\text{Zn}} \sim 10^{17}$ cm^{-3}. From these numbers eq. (5) gives $\eta \approx 50\%$. Such a high value of quantum efficiency may be possible in the near future.

Using quantum efficiencies of 50% in red diodes, one can think of large-area indicator light sources as one of the possibilities for solid-state lighting. A comparison is made with an incandescent lamp of 15 W which generates approximately 10% of its power in the visible and 1% in the red (i.e. 1.5 W of visible-light output and 0.15 W of red-light output). The luminous efficiency of such a lamp is about 10 lm/W corresponding to about 1 lm/W in the red region of the visible spectrum. A red diode of quantum efficiency 50% at 2 V d.c. and 50 mA would generate 0.05 W of optical power in the red. An array of three diodes would correspond to 0.15 W of optical power which is the same as generated by the incandescent lamp in the red region mentioned above. Furthermore, the luminous efficiency of these red diodes would correspond to 10 lm/W because of the higher quantum efficiency.

In the case of the radiative centres responsible for green and yellow emission in GaP, much more research is required before similar projections can be made. However, because of the greatly increased sensitivity of the eye in these regions, efficiencies of 1 to 10% would already give 5-50 lm/W.

It has been suggested recently that the high efficiencies in nitrogen and (Zn,O)-doped GaP might lead to stimulated emission of coherent radiation, i.e. laser action and experiments aimed at this goal have been reported [72] [73]. However, a characteristic inhomogeneous broadening is observed [72] instead of the expected line narrowing and further work is needed to clarify whether laser action in the visible is occurring in GaP. Beside GaP and the III-V ternaries, there are some other semiconductors which could be potential candidates for the future study of radiative centres. An example is SiC, which can give blue light; injection luminescence can be produced in both *N*- and *P*-type as well as several polytypes. The two most interesting forms are the cubic (β-SiC) and rhombohedral (α-6H), which possess indirect energy-band gaps of 2.38 eV and 3.08 eV at 0 K respectively. Since the melting point of SiC is ≈ 2600 °C, this semiconductor has been plagued by a serious growth problem. Furthermore the physics and the chemistry of these nonradiative recombination centres in SiC has not been studied in that great detail, primarily because no proper controlled doping of material has been done. Epitaxial layers have been grown to achieve yellow or blue diodes with relatively poor efficiency at desirable injection level [74] (Table I).

GaN, on the other hand has a band gap of 3.50 V and is a direct band-gap semiconductor. High photoluminescent efficiencies (12% at 300 K) [75] and laser action by optical pumping [76] have been recently achieved in this semiconductor. The fundamental problem with GaN besides availability of a single-crystal growth process is that only *N*-type GaN has been produced to date. Making it *P*-type may not be feasible as noticed in the case of other direct wide-band semiconductors like the II-VIs. To date, GaN diodes of the insulating-to-*N*-type have been made by Zn diffusion exhibiting green and blue d.c. electroluminescence at room temperature [77]. The external power efficiency is 10^{-4} and an external quantum efficiency of 10^{-2} has been achieved.

A review of progress on the radiative and nonradiative centres in GaP has been presented, which clearly demonstrates that mutually beneficial interaction can be derived between the physics of the subject on the one hand and the chemical aspects of the subject on the other. How far the LEDs will advance towards true solid-state lamps depends at this stage largely on our ability to translate this newly acquired understanding into diodes of much higher efficiency using commercially viable processes. At present it seems that these solid-state lamps will enjoy a bright future.

The author wishes to acknowledge the valuable comments and the encouragement from Dr. D. D. King and Dr. S. K. Kurtz. Useful comments have also been made by Ir. A. T. Vink.

Bibliography

[1] O. W. Lossew, Telegrafia i Telefonia **18**, 61, 1923, and C.R. Acad. Sci. U.R.S.S. **29**, 360 and 363, 1940.

[2] G. Destriau, J. Chimie phys. **33**, 587, 1936.

[3] A. G. Fischer, Electroluminescence in II-VI compounds, in: P. Goldberg (editor), Luminescence of inorganic solids, Academic Press, New York 1966, pp. 541-602.

[4] M. I. Nathan, W. P. Dumke, G. Burns, F. H. Dill, Jr., and G. Lasher, Appl. Phys. Letters **1**, 62, 1962.
R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys and R. O. Carlson, Phys. Rev. Letters **9**, 366, 1962.
T. M. Quist, R. H. Rediker, R. J. Keyes, W. E. Krag, B. Lax, A. L. McWhorter and H. J. Zeiger, Appl. Phys. Letters **1**, 91, 1962.

[5] J. W. Allen, M. E. Moncaster and J. Starkiewicz, Solid-State Electronics **6**, 95, 1963.

[6] H. G. Grimmeiss and H. Scholz, Physics Letters **8**, 233, 1964.

[7] M. Gershenzon, Electroluminescence from p-n junctions in semiconductors, in: P. Goldberg (editor), Luminescence of inorganic solids, Academic Press, New York 1966, pp. 603-684.
M. Gershenzon, Radiative recombination in the III-V compounds, in: R. K. Willardson and A. C. Beer (editors), Semiconductors and semimetals Vol. 2, Academic Press, New York 1966, pp. 289-369.

[8] D. G. Thomas, Physics Today **21**, No. 2, 43, Feb. 1968.

[9] M. R. Lorenz, Trans. Met. Soc. AIME **245**, 539, 1969.

[10] D. G. Thomas, Brit. J. appl. Phys./J. Phys. D **2**, 637, 1969.

[11] Y. P. Varshni, Phys. Stat. sol. **19**, 459, 1967, and **20**, 9, 1967.

[12] P. J. Dean, Junction electroluminescence, in: R. Wolfe and C. J. Kriessman (editors), Applied solid state science Vol. 1, Academic Press, New York 1969, pp. 1-151.

[13] P. J. Dean, J. Luminescence **1/2**, 398, 1970.

[14] H. C. Casey, Jr., and F. A. Trumbore, Mat. Sci. Engng. **6**, 69, 1970.

[15] D. G. Thomas, IEEE Trans. ED-**18**, 621, 1971.

[16] A. A. Bergh and P. J. Dean, Proc. IEEE **60**, 156, 1972.

[17] I. Ladany, J. appl. Phys. **42**, 654, 1971.

[18] R. Solomon *et al.* have recently achieved red GaP diodes of external quantum efficiencies of 12% utilizing a dipping process (see R. Solomon and D. DeFevere, J. Electronics Mat. **1**, 26, 1972). The technique utilizes the double epitaxial growth process reported by R. H. Saul, J. Armstrong and W. H. Hackett, Jr., Appl. Phys. Letters **15**, 229, 1969, where external quantum efficiencies of 7.2% were achieved.

[19] I. Hayashi, M. B. Panish, P. W. Foy and S. Sumski, Appl. Phys. Letters **17**, 109, 1970.

[20] F. W. Ostermayer, Jr., Metall. Trans. **2**, 747, 1971.
S. V. Galginaitis, Metall. Trans. **2**, 757, 1971.

[21] J. J. Hopfield, D. G. Thomas and M. Gershenzon, Phys. Rev. Letters **10**, 162, 1963.

[22] P. J. Dean, C. H. Henry and C. J. Frosch, Phys. Rev. **168**, 812, 1968.

[23] R. N. Bhargava, Phys. Rev. B **2**, 387, 1970; see also J. M. Dishman, Phys. Rev. B **3**, 2588, 1971.

[24] P. J. Dean and C. H. Henry, Phys. Rev. **176**, 928, 1968.

[25] H. Kukimoto, C. H. Henry and G. L. Miller, Bull. Amer. Phys. Soc. **17**, 233, 1972.

[26] D. G. Thomas, J. J. Hopfield and C. J. Frosch, Phys. Rev. Letters **15**, 857, 1965.

[27] R. A. Logan, H. G. White and W. Wiegmann, Solid-State Electronics **14**, 55, 1971.

[28] J. J. Hopfield, D. G. Thomas and R. T. Lynch, Phys. Rev. Letters **17**, 312, 1966.

[29] R. A. Faulkner, Phys. Rev. **175**, 991, 1968.

[30] J. C. Phillips, Phys. Rev. Letters **22**, 285, 1969.

[31] T. N. Morgan, B. Welber and R. N. Bhargava, Phys. Rev. **166**, 751, 1968.

[32] C. H. Henry, P. J. Dean and J. D. Cuthbert, Phys. Rev. **166**, 754, 1968.

[33] M. Gershenzon and R. M. Mikulyak, Appl. Phys. Letters **8**, 245, 1966.

[34] K. Maeda, M. Naito and A. Kasami, Jap. J. appl. Phys. **8**, 817, 1969.

[35] M. R. Lorenz and M. Pilkuhn, J. appl. Phys. **37**, 4094, 1966.

[36] H. Nelson, RCA Rev. **24**, 603, 1963.

[37] S. J. Bass and P. E. Oliver, J. Crystal Growth **3/4**, 286, 1968.

[38] M. Gershenzon, F. A. Trumbore, R. M. Mikulyak and M. Kowalchik, J. appl. Phys. **36**, 1528, 1965.

[39] J. D. Cuthbert, C. H. Henry and P. J. Dean, Phys. Rev. **170**, 739, 1968.

[40] A. Onton and M. R. Lorenz, Appl. Phys. Letters **12**, 115, 1968.

[41] R. N. Bhargava, J. appl. Phys. **41**, 3698, 1970.
J. S. Jayson, R. N. Bhargava and R. W. Dixon, J. appl. Phys. **41**, 4972, 1970.

[42] J. M. Dishman and M. DiDomenico, Jr., Phys. Rev. B **1**, 3381, 1970.
J. M. Dishman, M. DiDomenico, Jr., and R. Caruso, Phys. Rev. B **2**, 1988, 1970.

[43] J. S. Jayson and R. W. Dixon, J. appl. Phys. **42**, 774, 1971.
J. S. Jayson and R. Z. Bachrach, Phys. Rev. B **4**, 477, 1971.

[44] A. Kasami, Jap. J. appl. Phys. **9**, 946, 1970.

[45] R. N. Bhargava, Proc. Tenth Int. Conf. Phys. Semicond., Cambridge, Mass., 1970, p. 640.

[46] M. A. Afromowitz and M. DiDomenico, Jr., J. appl. Phys. **42**, 3205, 1971.

[47] J. A. W. van der Does de Bye and A. T. Vink, to be published in J. Luminescence.

[48] P. J. Dean, Phys. Rev. B **4**, 2596, 1971.

[49] R. N. Bhargava, S. K. Kurtz, A. T. Vink and R. C. Peters, Phys. Rev. Letters **27**, 183, 1971.

[50] P. J. Dean, Solid State Comm. **9**, 2211, 1971.

[51] G. M. Blom and R. N. Bhargava, to be published.

[52] R. N. Bhargava, C. Michel, W. L. Lupatkin, R. L. Bronnes and S. K. Kurtz, Appl. Phys. Letters **20**, 227, 1972.

[53] R. Z. Bachrach, O. G. Lorimor, L. R. Dawson and K. B. Wolfstirn, Bull. Amer. Phys. Soc. **16**, 436, 1971.

[54] M. Toyama and A. Kasami, to be published in Jap. J. appl. Phys.

[55] P. J. Dean and M. Ilegems, J. Luminescence **4**, 201, 1971.

[56] R. N. Bhargava, unpublished results.

[57] D. G. Thomas, M. Gershenzon and J. J. Hopfield, Phys. Rev. **131**, 2397, 1963.
D. F. Nelson, J. D. Cuthbert, P. J. Dean and D. G. Thomas, Phys. Rev. Letters **17**, 1262, 1966.

[58] P. J. Dean, R. A. Faulkner, S. Kimura and M. Ilegems, Phys. Rev. B **4**, 1926, 1971.

[59] A. T. Vink and R. C. Peters, J. Luminescence **3**, 209, 1970.

[60] P. J. Dean, Phys. Rev. **157**, 655, 1967.
P. J. Dean, R. A. Faulkner and S. Kimura, Phys. Rev. B **2**, 4062, 1970.
A. T. Vink, A. J. Bosman, J. A. W. van der Does de Bye and R. C. Peters, J. Luminescence **5**, 57, 1972.

[61] B. Welber and T. N. Morgan, Phys. Rev. **170**, 767, 1968.

[62] K. P. Sinha and M. DiDomenico, Jr., Phys. Rev. B **1**, 2623, 1970.

[63] J. C. Tsang, P. J. Dean and P. T. Landsberg, Phys. Rev. **173**, 814, 1968.

[64] J. D. Wiley, J. Phys. Chem. Solids **32**, 2053, 1971.

[65] R. Nicklin, C. D. Mobsby, G. Lidgard and P. B. Hart, J. Physics C **4**, L 344, 1971.

[66] R. N. Bhargava, Bull. Amer. Phys. Soc. **16**, 409, 1971.

[67] G. F. Neumark and D. Polder, Bull. Amer. Phys. Soc. **17**, 238, 1972, and private communication.

[68] W. O. Groves, A. H. Herzog and M. G. Craford, Appl. Phys. Letters **19**, 184, 1971.
M. G. Craford, private communication.

[69] R. Z. Bachrach and O. G. Lorimor, J. appl. Phys. **43**, 500, 1972.

[70] Recently a qualitative estimate was made (J. M. Dishman, Int. Electron Devices Meeting 1970, Washington, D.C., p. 154, Abs. 24.2), which suggests that removal of nonradiative centres by a factor of 10 can only increase of a factor of 2 in the quantum efficiency of diodes. However, such calculations are based on several estimated parameters (concentration of radiative centres, capture cross-sections, etc.) and would change the results as these parameters are determined more accurately.

[71] G. M. Blom, unpublished work.

[72] N. Holonyak, Jr., D. R. Scifres, H. M. Macksey, R. D. Dupuis, Y. S. Moroz, C. B. Duke, G. G. Kleiman and F. V. Williams, Phys. Rev. Letters **28**, 230, 1972.

[73] R. D. Burnham, N. Holonyak, Jr., D. L. Keune and D. R. Scifres, Appl. Phys. Letters **18**, 160, 1971.
R. E. Nahory, K. L. Shaklee, R. F. Leheny and R. A. Logan, Phys. Rev. Letters **27**, 1647, 1971.

[74] E. E. Violin, A. A. Kal'nin, V. V. Pasynkov, Yu. M. Tairov and D. A. Yas'kov, in: Silicon Carbide 1968, Proc. int. Conf., University Park, Penn. (special issue of Mat. Res. Bull. **4**, 1969), p. S 231.
R. W. Brander and R. P. Sutton, Brit. J. appl. Phys./J. Phys. D **2**, 309, 1969.
G. Kamath, private communication.

[75] M. Ilegems, R. Dingle and R. A. Logan, Bull. Amer. Phys. Soc. **17**, 233, 1972.

[76] R. Dingle, K. L. Shaklee, R. F. Leheny and R. B. Zetterstrom, Appl. Phys. Letters **19**, 5, 1971.

[77] J. I. Pankove, E. A. Miller and J. E. Berkeyheiser, RCA Rev. **32**, 383, 1971.

[78] D. G. Thomas, J. J. Hopfield and W. M. Augustyniak, Phys. Rev. **140**, A 202, 1965.
J. A. W. van der Does de Bye, A. T. Vink, A. J. Bosman and R. C. Peters, J. Luminescence **3**, 185, 1970.

Volume 32, 1971, No. 6/7/8 pages 145-280 Published 16th June, 1972

PHILIPS TECHNICAL REVIEW

VOLUME 32, 1971, No. 9/10/11/12

Microwave electronics

Recent years have seen so many changes taking place in microwave technology that the whole field is taking on a completely new look. The greatest changes are to be seen, naturally enough, in the laboratory, but the impact on the manufacturers and the user is already considerable. In the past microwave technology was concerned with relatively large and expensive components such as magnetrons, klystrons and waveguides. Today, however, lower-power microwave signals can be generated and processed in solid-state elements, and these are small and much less expensive. With these new components systems can now be made that earlier would have been inconceivable, and microwave equipment is no longer excluded from some likely applications because it is too large or too expensive. At last even the non-professional applications have become a practical possibility.

This issue of Philips Technical Review includes twelve articles that describe microwave studies at various Philips laboratories and give an impression of the nature of these new developments. They are introduced by an article that gives a general account of present trends and also includes a look at the likely pattern of future developments.

Solid-state microwave electronics

N. E. Goddard

Introduction

The history of modern technology includes many examples of unrecognized inventions, revived classical theories, premature technical developments and some frustrated industrial applications. These are the consequences of vigorous pioneering effort and the recent technology of microwave electronics is no exception. Nevertheless steady progress in technique and exploitation is maintained. Microwave technology is of increasing importance in fields such as communications, broadcasting and domestic electronics, as well as in the location, control, and navigation of land, sea, air and space vehicles.

This issue of Philips Technical Review is devoted to an account of some topics of recent research on microwave devices and associated circuits carried out in research laboratories associated with the Philips group of companies. The present time is opportune because radical changes in microwave technique are occurring simultaneously with the pressures of overcrowding in frequency, space and time in the application and user areas. There are pressures to expand the overcrowded communications and broadcasting frequency spectrum further into the microwave bands. There are strong demands for systems to increase the density and control of traffic flow in the overcrowded space available for ships, aircraft and road vehicles. There are demands for ever-increasing data processing and transmission speeds involving gigabit-per-second logic functions and gigahertz network bandwidths. Extensive research and development in microwaves has been undertaken in the past, particularly for demanding and sophisticated military applications. With recent developments in low-cost devices and circuits a challenging and expanding field of application can be foreseen with great scope in civil markets.

The early stages of electronic development in the microwave spectrum, broadly at frequencies from 1 GHz to 300 GHz, were beset with problems of radiation losses from conductors, dissipation in conductors and insulators, and the transit time of electron trajectories in conventional grid-controlled vacuum tubes. To overcome these problems, large, enclosed transmission-line circuits and complex, high-voltage, velocity-modulated generators and amplifiers were developed. A very precise and accurate circuit technique was established, backed up by an academically and practically rewarding body of elegant theoretical analysis. But paradoxically microwave electronics was not synonymous with microelectronics, even at micropowers. Nor were small wavelengths associated with small component and equipment costs. This situation is now undergoing radical change.

As with other branches of electronics, the greatest recent impact on the design of devices and equipments in the microwave spectrum has been made by solid-state physics. The consequent development of novel generator, amplifier and control devices was rather like the opening of Pandora's box. New devices and speculation on their applications proliferated; new investigations spread and multiplied with the aid of what some may consider the excessively immediate and diverse means of scientific communication. However, unlike the mythological analogue, more than Hope remains and many desirable objectives have been achieved without, as yet, creating too much havoc among mankind.

Some notes on the history of these developments and the main lines of present investigations will provide a backcloth for the papers which follow and references to some of the better established techniques which are not reported in this issue.

Solid-state devices

"This specious conjecture has no convincing theoretical argument in its favor." Thus commented H. C. Torrey and C. A. Whitmer [1] [*] in 1946 on the long-accepted assumption that a simple contact rectifier used as a frequency converter must always have a conversion loss greater than unity. After describing H.Q. North's welded-contact germanium crystals and showing that a negative intermediate-frequency conductance requires a barrier capacitance which varies strongly with barrier voltage, they comment further: "No use has yet been made of the amplifying properties of the welded rectifiers used as converters. It is possible that a lightweight receiver with no tubes, yet capable of detecting audio modulation of a weak microwave signal close to noise level, might be constructed by keeping these crystals on

N. E. Goddard, M.A., is head of the Systems Division of Mullard Research Laboratories, Redhill, Surrey, England.

the verge of oscillation by a suitable feedback arrangement. Some work along these lines was started by R. H. Dicke, but inconclusive results had been obtained when the war came to an end."

A further ten years elapsed and the transistor had been invented before the growth of solid-state microwave electronics, foreshadowed in this prophetic comment, gathered pace.

In 1948 A. van der Ziel [2] underlined the possibility of low-noise amplification with semiconductor diodes, and in 1956 J. M. Manley and H. E. Rowe [3] published an analysis of the general properties of non-linear elements and established the energy relationships for multi-frequency converters and amplifiers. The first example of a parametric amplifier was shortly afterwards reported by H. Suhl [4] and M. T. Weiss [5]. This amplifier was a non-linear inductance amplifier based on the non-linear properties of ferrite materials, but its further development was hindered by the requirement for very high microwave pump powers. This restriction did not apply to the non-linear-capacitance semiconductor diode or varactor which was also studied at this time for converter applications and frequency control [6] [7]. The development of degenerate and non-degenerate parametric amplifiers using varactors quickly followed [8] [9]. Negative-resistance, single-port, varactor parametric amplifiers proved capable of extensive development in low-noise and broad-bandwidth applications [10] [11] [12], despite the difficulties of stability and separation of input and output signals. Fortunately the solution to the latter problem had already been found in the prior extensive work on microwave ferrite gyrators, in particular the three- and four-port circulators. An extensive range of applications for varactors was developed, including frequency multipliers, limiters, switches and tuners.

In 1955, C. H. Townes and co-workers [13] reported the maser — a new type of microwave amplifier, frequency standard and oscillator. Coherent microwave oscillations were obtained by stimulating transitions between two quantum states of an ammonia system. N. Bloembergen [14] then proposed the three-level solid-state maser and this was demonstrated with gadolinium ethyl sulphate by H. E. D. Scovil *et al.* [15] in 1957. With an internal noise temperature of only a few degrees Kelvin when cooled to liquid-helium temperatures, and with the development of travelling-wave amplifier structures of 1% bandwidth, the solid-state maser was a vital component in the first Earth-station receivers for satellite communications [16]. It has now been largely superseded in this application by parametric amplifiers which have a much better capability and flexibility in bandwidth and operating bath temperature [12].

A new type of semiconductor diode with negative-resistance characteristics was discovered by L. Esaki [17] in 1957. Based on the phenomenon of electron-tunnelling across a *P-N* junction, the tunnel diode is capable of operating over a very wide frequency range and requires only a simple d.c. power supply and a simple signal circuit. Although at first a very wide range of applications was predicted for this diode, limitations of power-handling, stability, fragility and modest noise factor led to its adoption in only a limited range of circuits in which its simplicity could be exploited.

The severe power limitations of the small-area, point-contact and junction diodes required for microwave-frequency operation had long been recognized and the desirability of a volume mechanism in bulk material, as in the solid-state maser, but capable of supporting high energy densities, was acknowledged. In 1961 B. K. Ridley and T. B. Watkins [18] predicted bulk negative resistance in III-V semiconductor compounds as a result of the transferred-electron effect, and in 1963 J. B. Gunn [19] demonstrated microwave oscillations in gallium arsenide and indium phosphide which were explained by the Ridley and Watkins mechanism. Therefore the effect described by Ridley and Watkins is often referred to as the Gunn effect. The different modes of operation of Gunn diodes and their application for high-power generation and wide-range tuning have been extensively studied and some of the results are reported elsewhere in this issue [20] [21] [22].

Yet another type of semiconductor diode had been suggested by W. T. Read [23] in the 1950s in which negative resistance could result from avalanche breakdown and carrier transit time in a specially designed semiconductor junction. Practical devices were first reported in 1965 [24]. As with the Gunn or transferred-electron device, a variety of modes of operation have been explored [25] [26] [27]. The relative capabilities and advantages of Gunn and avalanche devices in different applications are not yet clearly resolved.

For switching, modulation and limiting functions, a *P-I-N* diode was developed in 1958. A similar structure had been reported earlier by R. N. Hall [28] for power rectification at low frequencies. At microwave frequencies there is no rectification and the central region of intrinsic semiconductor is insulating. When charge carriers are injected into this region by a forward bias voltage and the depth of the region is about equal to the recombination-diffusion length, an incident microwave signal is absorbed with very little reflection. Thus the device is an electronically controlled attenuator [29] [30] [31].

In parallel with all this work on new devices, a steady,

[*] *The references are listed at the end of the article.*

if less spectacular, application of planar semiconductor technology has been devoted to the improvement of the well-established detector and mixer diodes which continue to be used in larger quantity than other diodes. With the metal-semiconductor junction of the Schottky-barrier diode, substantial improvements have been made in mixer noise figure, conversion loss, burnout, dynamic range and mechanical reliability [32].The backward diode, with a highly non-linear *I-V* characteristic due to controlled tunnelling, has also provided a sensitive detector [33].

The remaining active device to be mentioned here is the transistor. Ever since its introduction in 1948, the maximum operating frequency of the transistor has been steadily increased and with new construction technologies experimental devices have been operated at frequencies beyond 10 GHz. In many microwave applications the transistor will be preferred to the two-terminal diode device because it is more readily controlled and has an inherently better isolation between input and output.

This great variety of device developments has depended on a parallel investigation of material physics and semiconductor technology which is too extensive to recount here. Most of the work on germanium and silicon has been undertaken for lower-frequency devices. At microwave frequencies the requirements for high carrier mobility, low permittivity and small dimensions have been met by the development of III-V compound semiconductors such as gallium arsenide, and technologies for epitaxial layers [34] and area metal-semiconductor junctions.

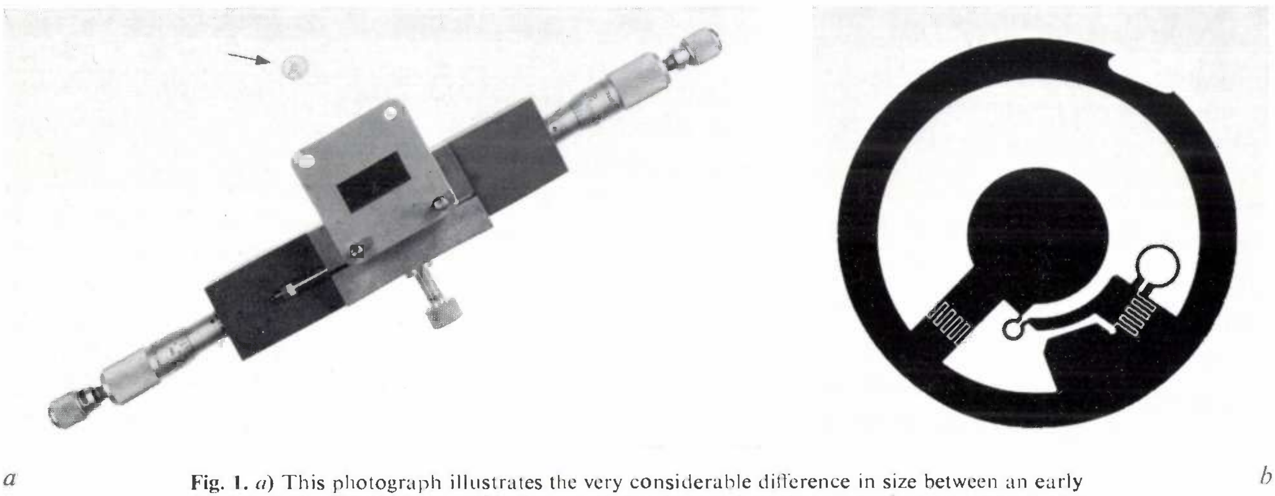


Fig. 1. *a*) This photograph illustrates the very considerable difference in size between an early waveguide tunnel-diode amplifier, tunable in the 10.7 GHz to 11.7 GHz frequency band, and a lumped-circuit amplifier, shown above it. The diameter of the lumped-circuit version is 7 mm. *b*) A diagram of the lumped-circuit tunnel-diode amplifier, now greatly enlarged. The conductors were made by depositing a thin gold-on-nickel chromium film on to a quartz substrate, and the amplifier is coupled to a coaxial circuit by means of concentric rings. At a gain of 10 dB and a noise figure of 7 dB the instantaneous bandwidth is about 3 GHz to 4 GHz.

Circuit techniques

Techniques for transmission, filtering and passive processing of microwave signals have undergone changes as radical as those in active devices. A number of important lines of development can be recognized which include the provision of non-reciprocal circuits with transmission through ferrite materials, the use of transmission modes in stripline techniques and the development of technology for achieving essentially lumped-constant circuits. There has also been a continuing development of the ubiquitous hybrid coupler which is a building brick in so many microwave circuit assemblies.

Stripline transmission circuits in shielded and balanced or in open microstrip form have been known for a very long time [35] [36] but a number of developments had to be made before technically and industrially effective applications could be adopted [37] [38]. The electromagnetic field patterns for which simple closed solutions were not available had to be analysed and translated into mode equations suitable for circuit analysis. Dielectric substrates with low losses and controlled permittivity and dimensions had to be developed. But the main impediment to stripline circuit techniques was the incompatibility of the essentially planar transmission lines and the available coaxial semiconductor detectors or vacuum tubes with coaxial and waveguide oscillatory circuits. Passive planar microwave integrated circuits were well developed: complementary active devices were not.

Solid-state amplifier and oscillator devices such as varactors, avalanche diodes, tunnel diodes and Gunn devices, especially with encapsulations suitable for stripline mounting, have enormously changed the design and application potential of microwave integrated circuits [39].

As a research topic, work on microwave ferrite devices is now limited to one or two special areas such as the propagation of surface waves. Many designs for gyrators and phase-shifters in hollow guide, coaxial line and stripline are available. The problem of compatibility between ferrite devices and integrated circuits has been tackled in several different ways. Thin-film circulators on ferrite-substrate inserts have been successfully designed and are reported elsewhere [37]. Two new approaches are the use of a ferrite substrate for the whole integrated circuit with magnetized regions for the non-reciprocal parts, and the design of circulators with miniature lumped-constant circuit elements on very small chips of ferrite [37] [40].

Simple lumped-constant circuits for microwave frequencies, such as tunable wavemeters, have been known for some thirty years. With the development of precision machining of miniature components, no doubt such circuits could have been more widely employed. However, as with striplines, there were severe compatibility problems with the active devices and such developments would have had little economic or technical value. With semiconductor devices and a planar circuit technique, the compatibility problem is removed and lumped circuits in thin-film form seem certain to become established in many applications [40]. The physical difference, for example, between an early waveguide tunnel-diode amplifier and its lumped-circuit equivalent is dramatic (*fig. 1*).

Two of the most important elements in microwave circuits are the filter and the coupler. For special applications, such as directional filters in a channel-dropping network for millimetric H_{01} transmission systems, it is necessary to use hollow guide for minimum attenuation. But for more general purposes stepped-impedance and stubs-and-lines filters in thin-film technique, suitable for integrated circuits, are widely employed. With a computer-aided design which optimizes performance within specified physical limitations and with a tape-controlled machine, it is practical to program a system for direct production of masks from the basic filter performance data (*fig. 2*, see p. 286). Filters that are tunable have also been developed [37].

At lower frequencies and with lumped elements, the coupler or hybrid junction was known as a symmetrical two-terminal-pair lattice network. Microwave versions were first developed as waveguide magic-T junctions and ring junctions. One of the first broad-band junctions was the phase-reversal coaxial ring which was developed as a 90° proximity coupler incorporated within a 180° ring [38]. A great diversity of couplers and hybrid junctions has been developed and adapted for use in thin-film integrated circuits. Following on the re-entrant and overlay couplers, one of the most promising recent techniques with possible application to other circuits is the use in combination of both the microstrip line and its inverse form known as the slotline (*fig. 3*).

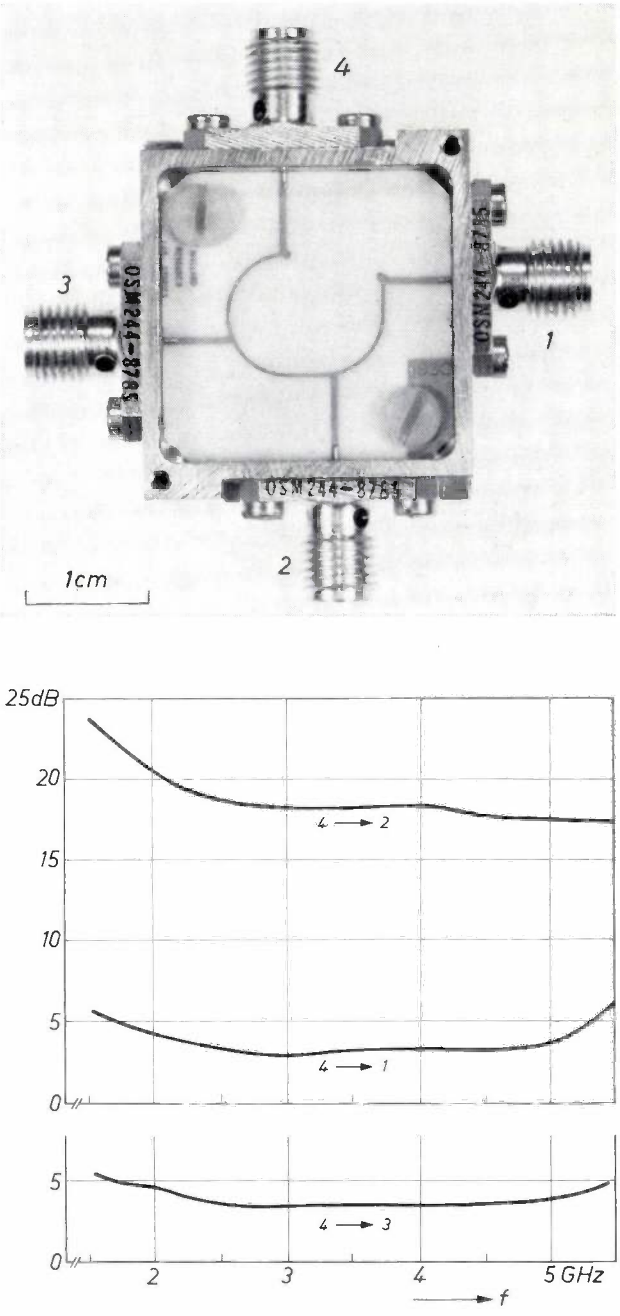


Fig. 3. The "hybrid" coupler in the photograph combines the microstrip and slotline techniques. A slotline is formed by a thin slot cut in a metal plane which is deposited on a dielectric sheet. Three pairs of ports are connected by microstrip; the remaining pair (*1*,*4*) are connected by a slotline. The combination of microstrip and slotline transmission lines gives a flat coupling and high isolation (*4* → *2*) over a wide band of frequencies.

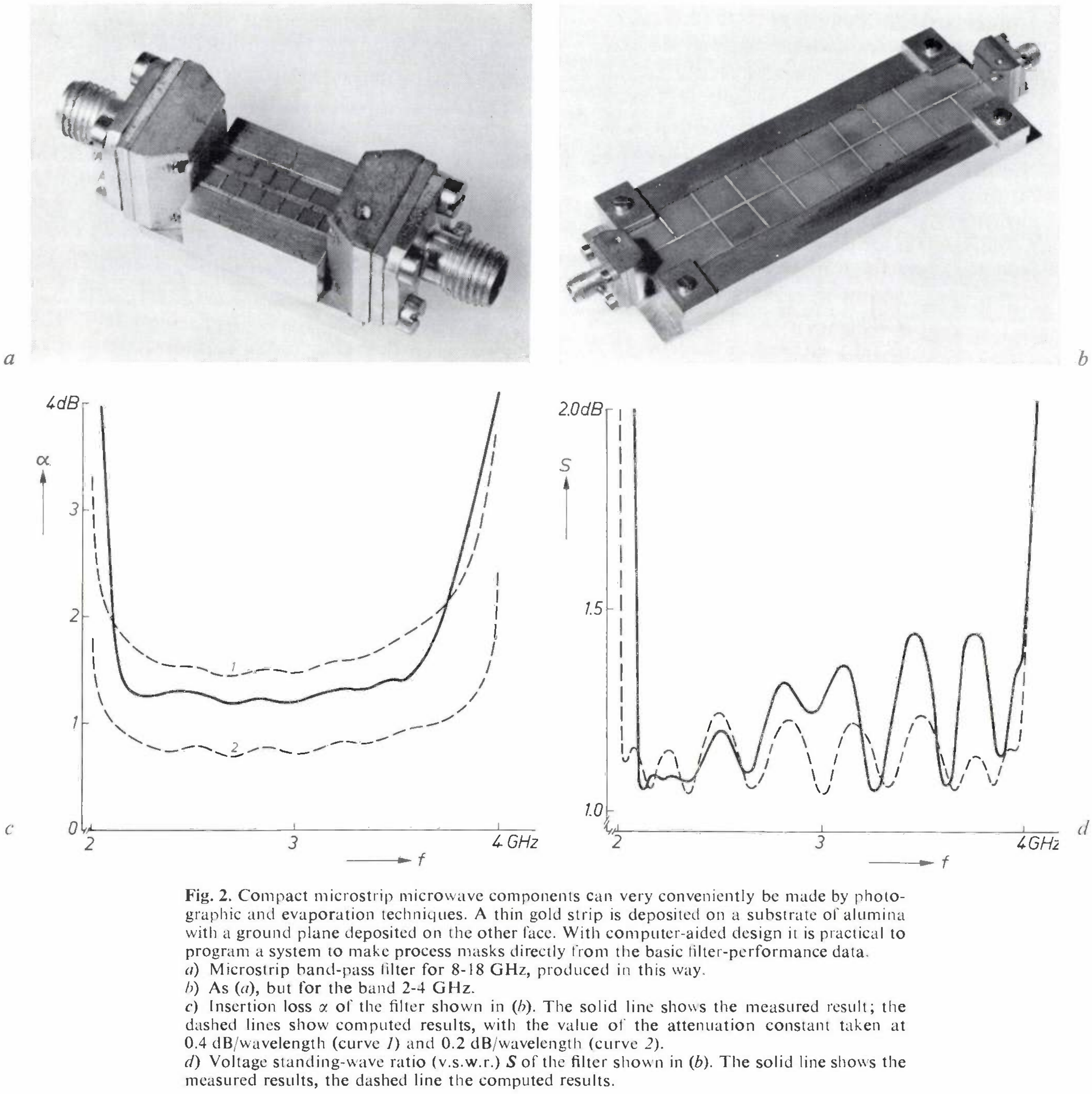


Fig. 2. Compact microstrip microwave components can very conveniently be made by photographic and evaporation techniques. A thin gold strip is deposited on a substrate of alumina with a ground plane deposited on the other face. With computer-aided design it is practical to program a system to make process masks directly from the basic filter-performance data.
a) Microstrip band-pass filter for 8-18 GHz, produced in this way.
b) As (*a*), but for the band 2-4 GHz.
c) Insertion loss α of the filter shown in (*b*). The solid line shows the measured result; the dashed lines show computed results, with the value of the attenuation constant taken at 0.4 dB/wavelength (curve *1*) and 0.2 dB/wavelength (curve *2*).
d) Voltage standing-wave ratio (v.s.w.r.) *S* of the filter shown in (*b*). The solid line shows the measured results, the dashed line the computed results.

Devices for existing systems

Having briefly traced the history and variety of solid-state microwave devices and circuits, it is pertinent to discuss their eligibility for current systems applications and their influence on future requirements for equipment design. The state-of-the-art performance is conveniently summarized in terms of detector and receiver sensitivity (*fig. 4*) and power-generating capability (*fig. 5*). With improving techniques for operation at the highest frequencies such summaries rapidly become out of date and it is also difficult to assess the practical validity of the performance of special laboratory devices.

For communication, broadcast, radar and navigation systems operating within the Earth's atmosphere, the well-known radio window between about 1 GHz and 15 GHz is important. For minimum-cost systems, the transistor, the tunnel diode and the mixer diode provide noise figures in the range 2 dB to 6 dB depending on the frequency and on the type of device. The transistor is expected to find increasing application in this frequency range as new technologies such as ion implantation permit improvement in geometry, and as devices like the field-effect transistor are further developed. The uncooled parametric amplifier has an internal noise temperature less than atmospheric thermal noise at ground level and meets the performance requirements of most other Earth systems. The conventional amplifier requires a pump source at a frequency much higher than the signal frequency but with improved solid-state pumps and thin-film circuits a relatively inexpensive assembly is already possible. For systems with large bandwidth operating between the Earth and space stations, the cooled parametric amplifier provides the best combination of sensitivity and frequency bandwidth. Uncooled parametric amplifiers are logistically more acceptable for mobile earth stations with narrower bandwidth.

Apart from a few special applications, such as 35 GHz radar, the main interest in devices at higher frequencies is for very broad-band trunk communications in low-loss waveguides. Frequencies in the range 35 GHz to 110 GHz are proposed and the development of suitable devices for repeaters is in hand.

For small-signal receiver applications, with few exceptions, the semiconductor device is capable of replacing the vacuum tube. As is apparent from fig. 5, semiconductor devices cannot yet compete with vacuum tubes as generators at the higher power levels but will find many applications as transmitters in low-power systems. The devices are also small enough to be used in matrix arrays and some substantial effective radiated power fluxes may be possible by this means.

The advent of cheap microwave generators and

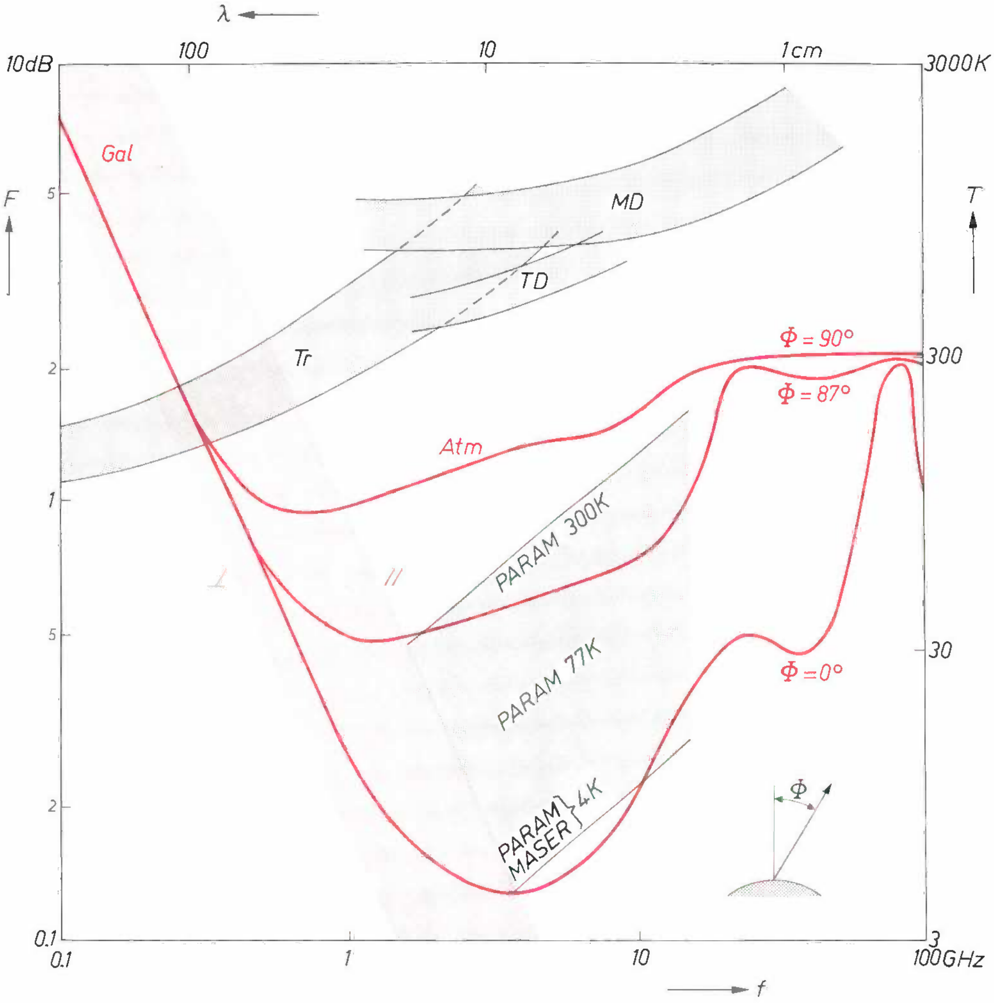


Fig. 4. In this figure the noise performance of various types of microwave receiver (grey regions) is compared with the noise produced at the aerial by atmospheric absorption (*Atm*) and extraterrestrial radiation (galactic noise; *Gal*). The degree of noisiness is plotted in the vertical direction and is shown as aerial noise temperature or excess noise temperature *T* on the right and noise figure *F* on the left; *f* represents frequency and λ wavelength. *Tr* transistors. *MD* mixer diodes. *TD* tunnel diodes. *PARAM* parametric amplifiers. The red curves show the variation with frequency of noise produced at the aerial for $\Phi = 0°$, 87° and 90°, where Φ is the angle between the direction of the aerial and the zenith. At the lower frequencies the aerial noise is mainly galactic noise (shaded region); at the higher frequencies the absorption noise dominates. Between about 1 and 15 GHz there is the "radio window", important for systems that operate within the Earth's atmosphere.

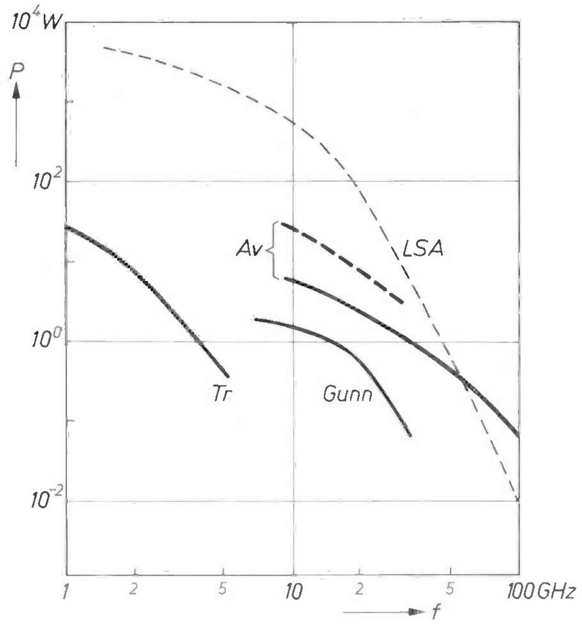


Fig. 5. State-of-the-art power generating capability of various solid-state microwave sources, with power P as a function of frequency f. *Tr* transistor. *Gunn* Gunn oscillator, continuous-wave. *LSA* Gunn oscillator operating at the LSA-mode. *Av* avalanche diode, continuous-wave (drawn curve) and pulsed (dashed curve).

integrated circuits is also expanding the application of radar and communication techniques to new market areas. Simple Doppler radars are already extensively used for protection of commercial and domestic premises from intruders. Experiments have been made with miniature radars as aids to drivers of vehicles operating on airfields in bad visibility. A microminiature radar has been fitted in a road surface in the space normally occupied by a "cat's-eye". It detects vehicles and measures their speeds of vehicles for road-traffic control. Many more applications of this kind may be anticipated.

Quo vadis?

Excepting a dramatic new discovery, the course of development of solid-state microwave devices seems established for the immediate future. For more speculative developments in microwave electronics we must examine the possibilities for novel systems.

The replacement of vacuum tubes by semiconductors throughout the microwave spectrum and in all but the highest-power applications is a challenging task and should result in more compact, efficient and economic systems. This, however, is not the end of the story because an even greater challenge remains to be met. Let us look at the situation at lower frequencies for which semiconductors are at a more advanced stage of development. We note the development of portable radio and television receivers and the rather slower impact on capital equipment such as the telephone network. But the really vigorous and expansive development, made possible by semiconductor technology, is that of computer and data-processing systems. One of the main reasons is that the high component density and low power consumption of semiconductor circuits, particularly monolithic integrated circuits, has made it possible to use a very large matrix of circuit functions, in both parallel and serial modes, which the systems engineer could barely conceive in vacuum-tube techniques. A similar challenge now faces the microwave systems engineer because microwave integrated circuits are potentially very small and economic in power consumption and cost. It is, perhaps, unlikely that a new industry on the scale of the computer industry will arise (though it may be noted that logic speeds are heading into the microwave spectrum) but it seems probable that new and important applications will be found as systems research and development takes account of the new technology.

Let us pursue this analogy between low-frequency and microwave solid-state developments and see where it leads.

Microwave television broadcasting is already being made an economic possibility by the development of the new technology and a practical need by the overcrowding of the lower-frequency bands in densely-populated multilingual areas (*fig. 6*). Microwave-satellite broadcasting is also being seriously studied.

With its large scale of application, broadcasting is clearly an important area for microwave techniques, but, with the solid-state developments at lower frequencies already cited, the major novel systems possibilities

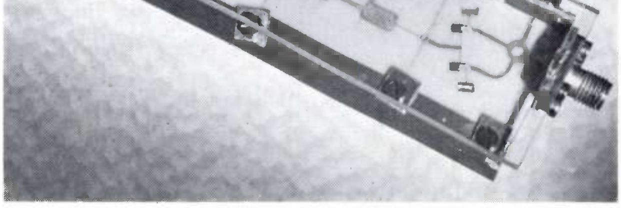


Fig. 6. An experimental converter for microwave television receivers. It converts microwave broadcast signals from Earth or satellite stations to UHF signals suitable for standard receivers. This microstrip mixer and amplifier circuit receives signals in the 12 GHz band and provides output signals in the 600-900 GHz UHF band.

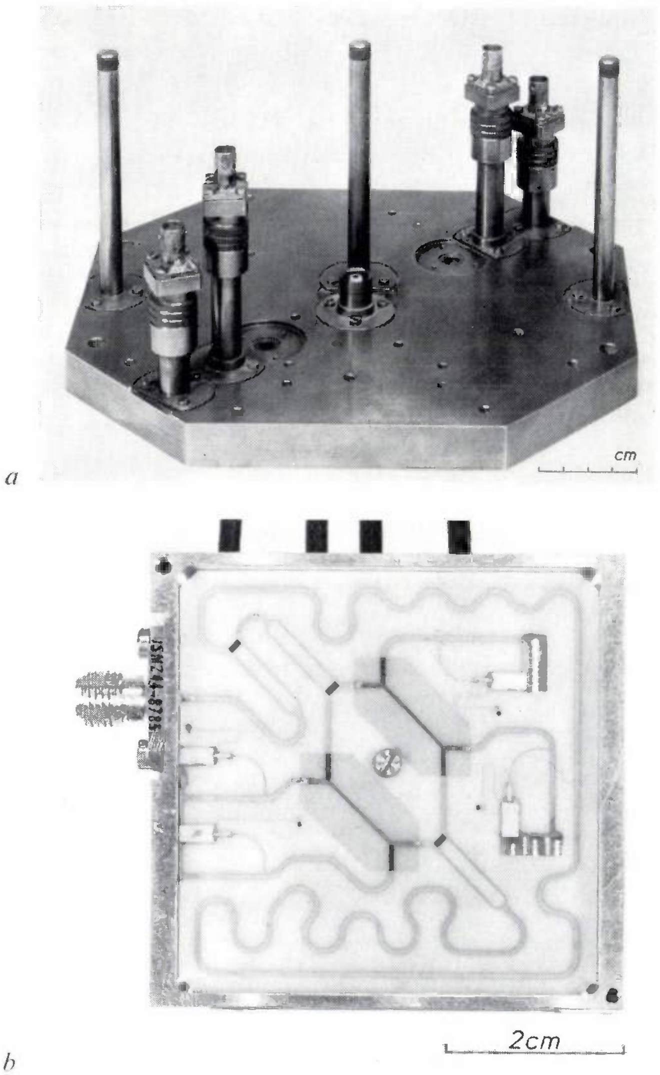


Fig. 7. *a*) Early experimental phase discriminator in coaxial techniques for the 2.0-4.0 GHz band. The circuit includes five hybrid rings, four detector diodes and three matched loads. *b*) A microstrip discriminator for the same band. The phase bridge here is formed from two overlay couplers, and there are also three matched-T power dividers, detectors, loads and a long delay line.

may be found in areas other than broadcasting. Until recently, the microwave systems engineer has had to think in terms of high-voltage klystron and magnetron generators and expensive, precision circuit "plumbing" in waveguide or coaxial technique. In all but the more sophisticated military systems, one transmitter and receiver channel was all he could generally afford or accommodate in a typical installation. It is now economically and technically possible to consider much more complex multichannel systems involving signal processing — much as the computer involves data-processing. A well known example is the phased-array aerial system with a large matrix of microwave radiators or sensors. By controlling the phase [41] and amplitude of the individual element radiations or by processing the signals received by the sensors, a highly controllable and adaptable system may be achieved, with a capability for a large number of parallel or serial functions. Thus such an array in a radar system might be used for simultaneous tracking of a number of targets in combination with a volume surveillance mode.

Another significant example which is perhaps a pointer to the possibility of new systems is the multiple phase discriminator (*fig. 7*). Phase measurement in microwave systems is important, particularly with arrays of sensors [42], because a knowledge of the phases and amplitudes of signals in each sensor entirely defines the information available in the system. In applications concerned with the measurement of the location and the frequency of transmissions, phase measurement is particularly useful.

Phase is a cyclic property of radio signals and by increasing the circuit time delay between two signals of identical frequency or the path difference between two sensors, a phase measurement can be used to determine frequency or direction of arrival of a transmission with a resolution limited only by the time duration of the signal. However, the greater the resolution of the phase-measuring circuit, the greater will be the number of possible ambiguities in frequency or direction, corresponding to the number of signal-cycle periods in the relative time delay or path difference employed. It can be demonstrated that all ambiguities may be resolved, with the maximum tolerance to phase perturbations caused by propagation or instrumental effects, with a multichannel system. In one example of this system there are N time or path delays with values in the ratios $1 : 2 : 4 : 8 : \ldots : 2^N$ and $N + 1$ receivers. With suitable processing of the phase measurements, the accuracy of frequency or direction measurement is determined by the longest time or path delay and the instantaneous frequency or angular coverage by the circuit with shortest delay.

One of the first examples of such a system is the instantaneous frequency-measuring receiver (*fig. 8*). This unique receiver accepts signals in an octave frequency range and instantaneously measures the frequency with an accuracy determined by the number of delay lines and phase discriminators employed and the duration of the signal. A large number of time-interlaced signals may be measured over a wide range of signal amplitudes and without the use of any tuned circuits or receiver controls. The phase tolerance of the system permits the accurate measurement of frequency even in the presence of simultaneous interfering signals within a few decibels of the measured signal. The containment of such a multichannel system within a reasonable equipment size and within practical power supply requirements is made possible by the use of hybrid and monolithic integrated-circuit techniques for all the

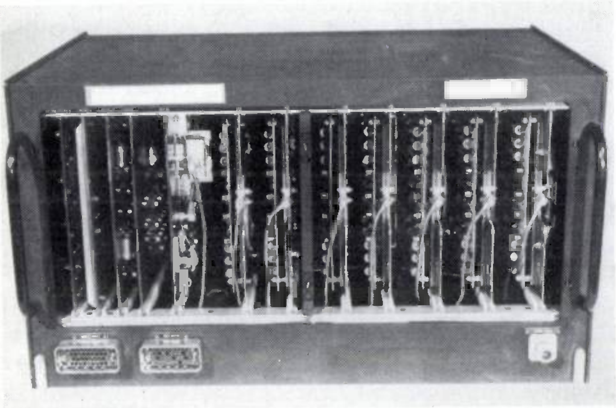


Fig. 8. Instantaneous frequency-measuring receiver. There are seven hybrid-integrated microwave discriminator channels, each with associated integrated-circuit video-frequency amplifiers and digitizing circuits. Simple logic circuits convert the output signals from the discriminators into an accurate, unambiguous set of parallel binary digits representing the signal frequency. The receiver measures the frequencies of signals in an octave frequency range without the use of tuned circuits or other controls.

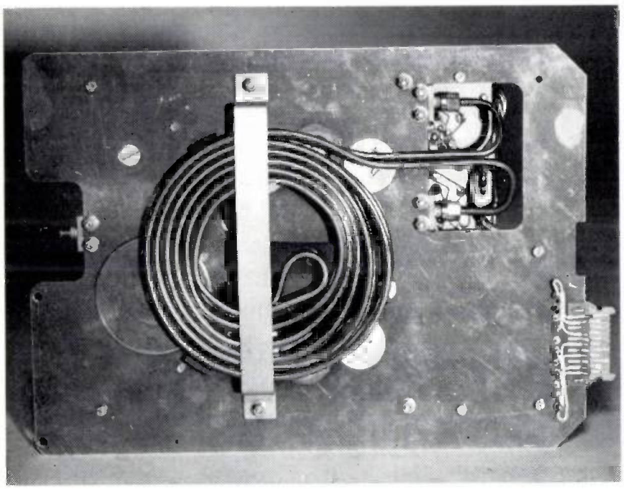


Fig. 9. One of the frequency discriminator units for the frequency-measuring receiver of fig. 8. The unit is made in stripline technique, with a coaxial delay line, and can be plugged in for rapid interchangeability. Detectors, loads and a trombone-type phase adjuster are incorporated within the stripline circuit.

microwave and digital signal processing functions (*fig. 9*).

A similar example of this system technique, again made feasible by microwave hybrid-integrated circuits, is the use of microwave interferometers in a system for the landing guidance of all types of VTOL, STOL and conventional fixed-wing aircraft. In this case a linear array of receivers, with aerial spacings (path delays) chosen to provide optimum sampling over the aperture, is used to measure the angle of arrival of transmissions from an aircraft beacon. Two arrays, one for azimuth and one for elevation (*fig. 10*), permit the instantaneous measurement of the bearing angles of a large number of aircraft over a very wide range of approach directions in both elevation (glide slope) and azimuth (localizer). Very high angular accuracies and high tolerance to phase perturbations caused by the environment can be achieved.

These are but two examples of important systems applications made possible by the new microwave device and circuit technologies in combination with original ideas on multichannel measurement techniques. The use of cheap microwave circuits in signal-processing functions, characterized by the preservation of phase and amplitude from individual sensors in multiple arrays, is perhaps one significant portent for the future of microwave electronics. The physical principles are well known and the new microwave technology is available. With the consequent reorientation of approach to equipment design no doubt many more important new systems will be generated.

Bibliography

[1] H. C. Torrey and C. A. Whitmer, Crystal rectifiers, McGraw-Hill, New York 1948.
[2] A. van der Ziel, J. appl. Phys. **19**, 999, 1948.
[3] J. M. Manley and H. E. Rowe, Proc. IRE **44**, 904, 1956.
[4] H. Suhl, Phys. Rev. **106**, 384, 1957.
[5] M. T. Weiss, Phys. Rev. **107**, 317, 1957.
[6] A. Uhlir, Jr., Proc. IRE **44**, 1183, 1956.
[7] L. J. Giacoletto and J. O'Connell, RCA Rev. **17**, 68, 1956.
[8] M. E. Hines, Amplification in non-linear reactance modulators, 15th Annual Conf. Electron Tube Research, Berkeley, Cal., U.S.A., 1957.
[9] G. F. Herrmann, M. Uenohara and A. Uhlir, Jr., Proc. IRE **46**, 1301, 1958.
[10] C. S. Aitchison, Philips tech. Rev. **28**, 204, 1967.
[11] C. S. Aitchison, R. Davies and P. J. Gibson, Proc. Symp. on Microwave Applications of Semiconductors, London 1965, paper 26, also published in IEEE Trans. **MTT-15**, 22, 1967.
[12] C. S. Aitchison, E. L. Hentley, S. R. Longley and J. C. Williams, Proc. 7th International Conf. on Microwave and Optical Generation and Amplification, Hamburg 1968, page 564.
[13] J. P. Gordon, H. J. Zeiger and C. H. Townes, Phys. Rev. **99**, 1264, 1955.
[14] N. Bloembergen, Phys. Rev. **104**, 324, 1956.
[15] H. E. D. Scovil, G. Feher and H. Seidel, Phys. Rev. **105**, 762, 1957.
[16] J. C. Walling and F. W. Smith, Philips tech. Rev. **25**, 289, 1963/64.
[17] L. Esaki, Phys. Rev. **109**, 603, 1958.
[18] B. K. Ridley and T. B. Watkins, Proc. Phys. Soc. **78**, 293, 1961.
[19] J. B. Gunn, Solid State Comm. **1**, 88, 1963.
[20] G. A. Acket, R. Tijburg and P. J. de Waard, The Gunn diode; this issue, page 370.
[21] J. de Groot and A. Mircea, Computer calculations of the Gunn effect; this issue, page 385.
[22] J. Magarshack, Gunn-effect oscillators and amplifiers; this issue, page 397.
[23] W. T. Read, Jr., Bell Syst. tech. J. **37**, 401, 1958.
[24] R. L. Johnston, B. C. De Loach, Jr., and B. G. Cohen, Bell Syst. tech. J. **44**, 369, 1965.
[25] D. de Nobel and M. T. Vlaardingerbroek, IMPATT diodes; this issue, page 328.
[26] K. Mouthaan, IMPATT-diode oscillators; this issue, page 345.
[27] P. J. de Waard, Anomalous oscillations with an IMPATT diode, this issue, page 361.
[28] R. N. Hall, Proc. IRE **40**, 1512, 1952.
[29] F. C. de Ronde, H. J. G. Meyer and O. W. Memelink, IRE Trans. **MTT-8**, 325, 1960.

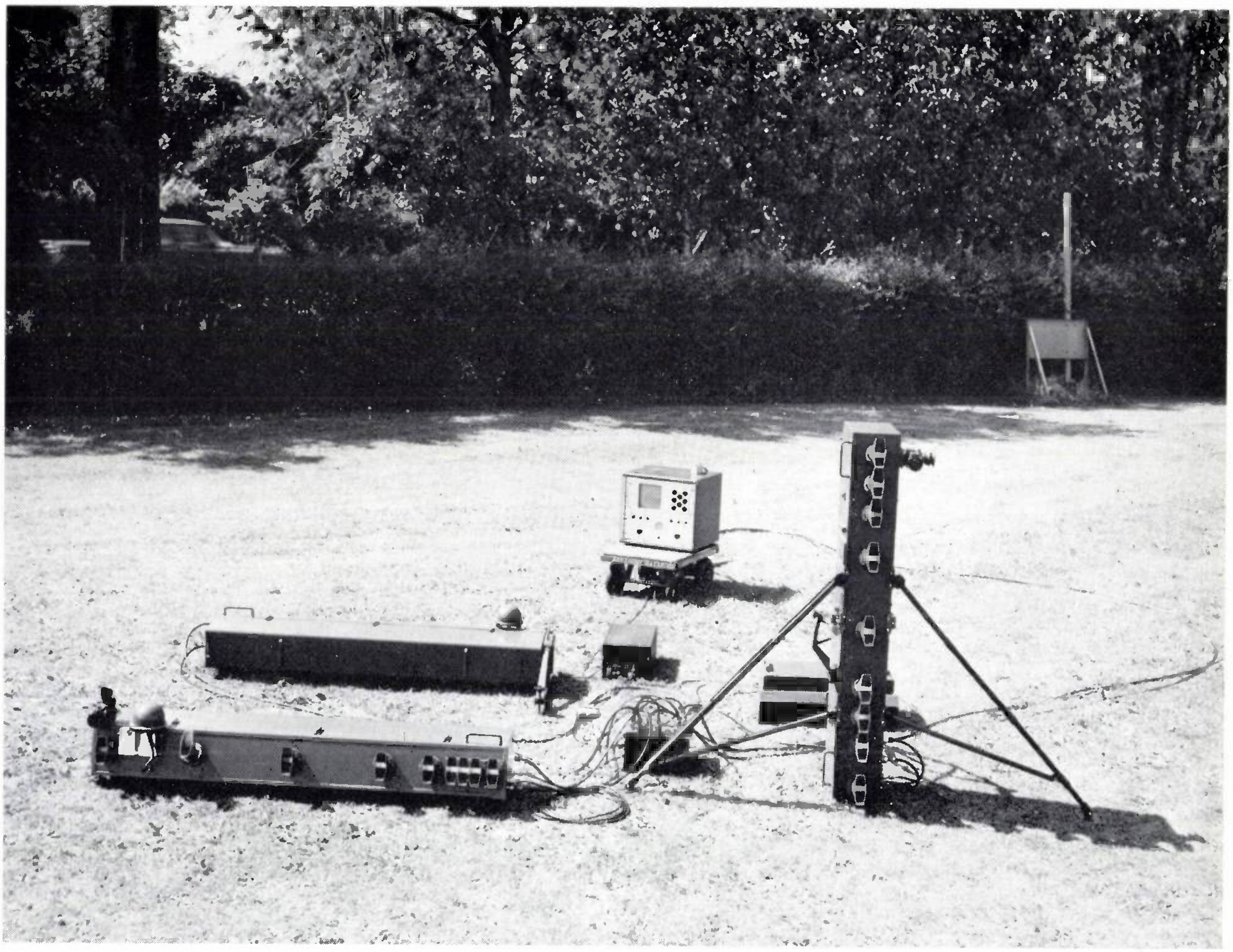


Fig. 10. A system that has become feasible through the use of microwave hybrid-integrated circuits, which are both compact and rugged. Here three microwave interferometers are combined to give a high-performance system for guiding helicopters and fixed-wing aircraft to a landing area. The interferometer units incorporate distance-measuring equipment and a data link. Horizontal horn-aerial arrays (about 2 m long) provide azimuth-approach and overshoot guidance over a wide range of approach angles with an r.m.s. fluctuation of about 0.05°. The vertical array provides simultaneous elevation guidance for approach angles between 2° and 20° with an r.m.s. fluctuation of about 0.1°.

[30] A. Uhlir, Jr., Proc. IRE **46**, 1099, 1958.

[31] T. H. B. Baker, Electronic Technology **38**, 300, 1961.

[32] H. N. Daglish, J. G. Armstrong, J. C. Walling and C. A. P. Foxell, Low-noise microwave amplifiers, Cambridge Univ. Press, 1968.

[33] T. Oxley and F. Hilsden, Radio and Electronic Engr. **31**, 181, 1966.

[34] A. Boucher and B. C. Easton, Epitaxial growth of gallium arsenide; this issue, page 380.

[35] D. D. Grieg and H. F. Engelmann, Proc. IRE **40**, 1644, 1952.

[36] W. E. Fromm and E. G. Fubini, Proc. National Electronics Conf., Chicago 1954, page 58.

[37] M. Lemke and W. Schilz, Microwave integrated circuits on a ferrite substrate; this issue, page 315.
P. Röschmann, YIG filters; this issue, page 322.

[38] S. J. Robinson and P. T. Saaler, Philips tech. Rev. **28**, 211, 1967.

[39] J. H. C. van Heuven and A. G. van Nie, Microwave integrated circuits, this issue, page 292.

[40] C. S. Aitchison, Lumped components for microwave frequencies; this issue, page 305.

[41] J. H. C. van Heuven, *P-I-N* switching diodes in phase-shifters for electronically scanned aerial arrays; this issue, page 405.

[42] R. N. Alcock, Philips tech. Rev. **28**, 226, 1967.

Summary. As an introduction to this issue on microwave solid-state devices the history of device developments and compatible circuit techniques is outlined from the speculations of the 1940s to the present state of the art.

In existing equipment and systems solid-state devices can replace vacuum tubes often with improved performance and smaller cost in all but the higher power applications. Microwave technology is of increasing importance in systems for communications broadcasting domestic electronics and the control and navigation of all forms of transportation.

It is suggested that the full impact of microwave technology on systems innovation has not yet been realized. Some novel multichannel equipments with microwave solid-state circuits integrated into the signal processing functions may be a portent for the future.

Philips tech. Rev. 32, 292-304, 1971, No. 9/10/11/12

Microwave integrated circuits

J. H. C. van Heuven and A. G. van Nie

Introduction

Electrical circuits for low frequencies are usually built up from components and conducting connections between the components. Capacitance, inductance and resistance, for example, are thought of as concentrated in the components, and although the conducting connections also possess these properties to some extent, their contribution compared to that of the components is so small that it is usually neglected. The circuits are therefore referred to as circuits with lumped components.

This idealized representation of an electrical circuit differs increasingly from reality as the frequencies become higher. In the first place stray elements begin to play a more important part; typical strays are the inductance of the conducting connections and the capacitance between them, and also the capacitance between the turns of a coil and the inductance of the plates of a capacitor. The magnitude of these stray elements is usually not known exactly. In the second place, the transit time of the signal in the circuit becomes significant when it constitutes a larger part of the cycle of the r.f. signal — in other words, when the wavelength of the electromagnetic field in and around components and conductors approaches the same order of magnitude as their dimensions. In this case there is generally radiation as well.

These effects have led to the development of an entirely different circuit technique for very high frequencies (above 1 GHz, corresponding to wavelengths shorter than 30 cm), in which the circuits are not built up from lumped components [1]. This *microwave* technique is based on the use of transmission lines in which the electromagnetic field is bounded in the transverse direction, so that no radiation occurs. The electrical behaviour of the transmission-line configurations can be accurately calculated, thus removing uncertainties caused by strays. The most familiar types of transmission line or guide are the coaxial line and the rectangular waveguide (*fig. 1a* and *b*). An electromagnetic wave is able to propagate in the space inside these transmission lines, and a pattern of currents in a thin "skin" on the surface of the conducting wall accompanies the wave. In a coaxial line the electric and magnetic lines of force all lie in planes perpendicular to the longitudinal axis of the conductor; this means that a voltage and a current can be defined for any given cross-section, and that the transmission line can be characterized by a capacitance and inductance distributed along the length. The circuit parameters in microwave technology are thus referred to as distributed. A representation as simple as this is not possible in a hollow waveguide. A picture of what happens in the rectangular waveguide of fig. 1*b* is obtained by imagining that the waves are reflected by the inside walls, giving propagation only in the longitudinal direction.

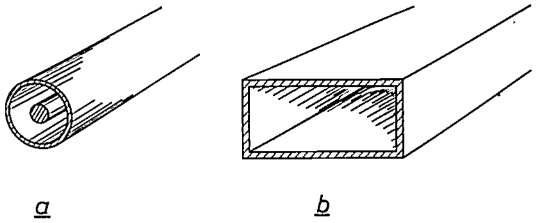


Fig. 1. *a*) Coaxial line. The electromagnetic wave propagates in the space between the inner and outer conductor and is entirely enclosed in the lateral direction. *b*) Rectangular waveguide. Here again wave propagation is only possible in the longitudinal direction of the guide.

When a signal is sent along a length of transmission line that is short-circuited at one end, the signal is reflected at the short-circuit. It returns to the input with a phase angle which differs from that of the incoming signal by an amount that depends on the length of the line. A reflection can also be caused by discontinuities other than a short-circuit. Such a length of transmission line behaves as a reactive element, which is inductive or capacitive depending on the wavelength and on the nature of the termination. Short lengths of line like this (stubs) are often used in microwave circuits, and can be connected in shunt with a continuous transmission line.

Circuits produced in this way with coaxial lines or hollow waveguide are precision-engineering devices, and their high cost rules them out for any but professional applications. Their dimensions were not such a very important consideration when only thermionic valves were available for generating and amplifying microwave signals. Nowadays, however, these functions are being taken over by small semiconductor

Ir. J. H. C. van Heuven and Ir. A. G. van Nie are with Philips Research Laboratories, Eindhoven.

devices [2], and this made it necessary to look for correspondingly smaller kinds of transmission line. The answer was found in the form of planar transmission lines.

Planar transmission lines

Planar transmission lines consist of thin metal strips and metal layers on a non-conducting substrate. Some typical cross-sections are shown in *fig. 2*. Fig. 2*a* shows a cross-section of the symmetrical stripline or sandwich line, which consists of a strip conductor between two metal ground planes. The dielectric occupying the intermediate space carries the thin metal strip but is not essential to the operation of this transmission line. Stripline can be regarded as a flat version of the coaxial line, with side walls omitted. Although the electromagnetic field here is not entirely enclosed by metal, there is nevertheless no radiation. A disadvantage is that the inner conductor is not easily accessible for connecting to other circuit elements, such as semiconductor devices. For this reason microstrip line is more frequently used in practice (fig. 2*b*). The ground plane can be considered as a reflecting plane (fig. 2*c*); the microstrip line is then found to be closely related to the two-wire transmission line, a familiar example of which is the ribbon feeder line.

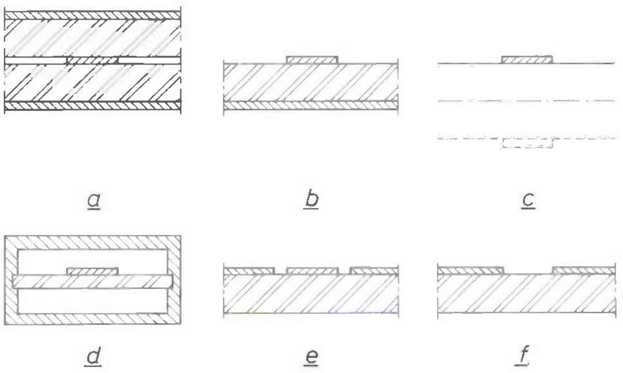


Fig. 2. *a*) Stripline. A strip conductor is sandwiched between two ground planes and is held in place by two layers of dielectric material. The stripline can be regarded as a flat form of the coaxial line (fig. 1*a*). *b*) Microstrip. The strip conductor here is more readily accessible for making connections than symmetrical stripline (*a*). *c*) The microstrip line is the electrical equivalent of a twin-wire transmission line in which the second conductor is situated at the image of the strip conductor in the ground plane. *d*) Suspended stripline. The field of the strip conductor, unlike that in (*a*) and (*b*), is largely in air. *e*) Coplanar waveguide. *f*) Slotline. The r.f. electric field appears across the slot, and slotline is therefore suitable for the connection of parallel impedances.

Other planar transmission lines that have been developed include suspended stripline, the coplanar waveguide and slotline (fig. 2*d*, *e*, *f*). By far the most widely used, however, is microstrip line, and this is the type with which we shall be concerned in the rest of this article, apart from a few remarks on slotline.

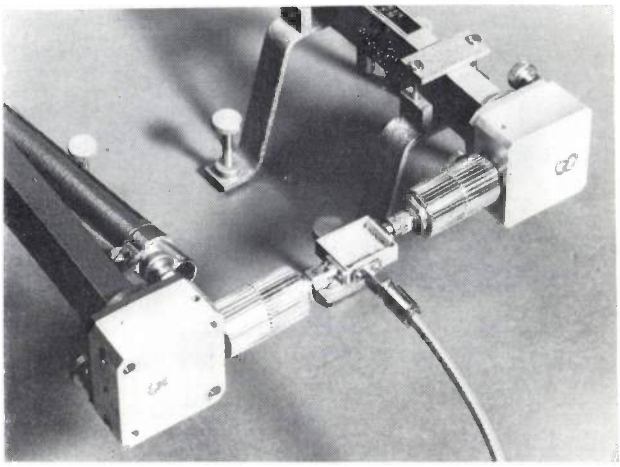


Fig. 3. A microstrip circuit connected by waveguides to a test equipment. The photograph gives an idea of the relative sizes of microstrip line and waveguide for the 10 GHz band.

The conductors are formed on the substrate by vacuum evaporation or electroless deposition; the required pattern of conductors is obtained by means of photo-etching techniques. These make it possible to produce the complete pattern of conductors of a whole circuit at the same time, hence the term "integrated circuit". The semiconductor devices are added to the circuit later, as in the case of hybrid integrated circuits for lower frequencies.

The dimensions of an integrated circuit made in microstrip line are very much smaller than those which the circuit would have if it were made up from lengths of waveguide or coaxial line; this is illustrated by the photograph of part of the test arrangement shown in *fig. 3*. Microstrip line also reduces the number of mechanical connections required in the circuit, and this improves reliability. On the other hand the losses of microstrip are greater than those of coaxial line or waveguide. Owing to these higher losses the characteristics of very narrow-band filters in microstrip are not very good. There are other solutions [3] for these filters, however, and in general it is possible to use microstrip to produce microwave circuits that have good and reproducible electrical characteristics, are small andalso, provided they are produced in sufficiently large quantities, relatively inexpensive. This opens up prospects for applications in phased arrays for

[1] Microwave circuits with lumped components are not fundamentally impossible, but the requirement that components and connections shall be small compared with the wavelength assumes very small dimensions. These can nowadays be obtained by means of planar techniques; see C. S. Aitchison, Lumped components for microwave frequencies, this issue, page 305.

[2] D. de Nobel and M. T. Vlaardingerbroek, IMPATT diodes, this issue, page 328.
G. A. Acket, R. Tijburg and P. J. de Waard, The Gunn diode, this issue, page 370.

[3] P. Röschmann, YIG filters, this issue, page 322.

radar — aerials which are built up from thousands of aerial elements with associated phase-shifters [4], to scan the beam without mechanical movement — and in television transmission at centimetre wavelengths. This brings microwave technology to the consumer market for the first time.

Our circuits, some of which will be described in this article, have all been made on a non-conducting substrate; it would however be feasible to use a semiconducting substrate, e.g. of silicon or gallium arsenide, and to produce the required semiconductor diodes in the substrate by local *P*-type and *N*-type diffusions, giving a monolithic integrated circuit. As yet it has not been possible to make practical circuits by this method; the dielectric losses in the semiconductor material are too high, and the chip area needed for a circuit with distributed elements is greater than is allowed by available single crystals.

In this article we shall therefore confine ourselves to circuits on non-conducting substrate materials such as aluminium oxide and quartz glass. Before dealing with materials and technology, we shall first take a somewhat closer look at the electrical characteristics of microstrip and slotline.

Electrical characteristics of microstrip

Homogeneous dielectric

In a microstrip line the lines of force are partly inside and partly outside the substrate (*fig. 4*). As a rule the substrate does not have the same dielectric constant as the ambient air, and this inhomogeneity of the medium causes considerable complications in the solution of the equations that describe the propagation of waves along the microstrip line. To discuss the effect of the dielectric on the characteristics of a microstrip line we shall therefore begin with the theoretical case in which the entire microstrip line is contained in a homogeneous dielectric with a dielectric constant ε_r and a relative permeability μ_r. In this case the phase velocity c of the wave propagating along the line is given by the expression

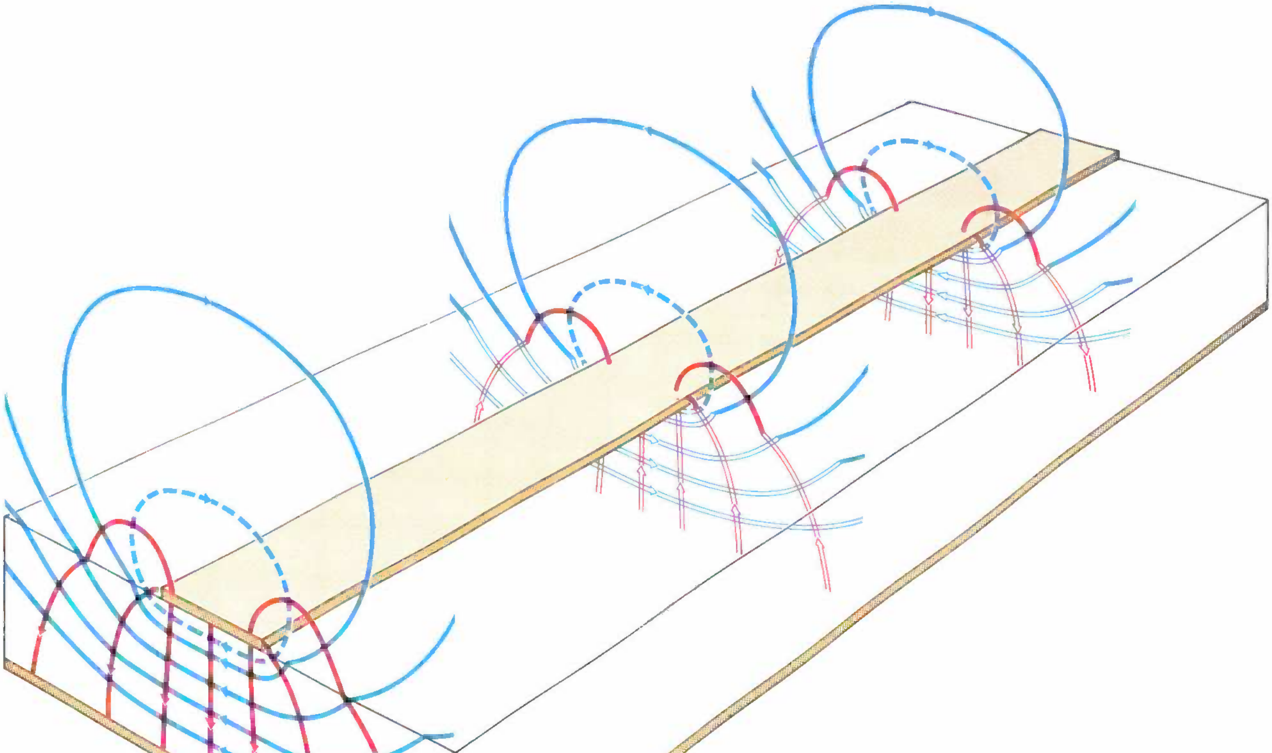


Fig. 4. Pattern of the electromagnetic field of a microstrip line. In this figure both the electric field (red) and the magnetic field (blue) are transverse; this is an approximation which is valid only at the lower microwave frequencies. The lines of force go partly through air and partly through the dielectric. The dashed line shows the magnetic field close to the strip.

$$c = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}}, \qquad (1)$$

where c_0 is the velocity of light in free space. In the materials used ε_r is greater than 1, and μ_r is usually equal to 1; the phase velocity along the line is therefore less than that in free space and indeed, since ε_r may easily be 10 or more, is considerably less. Consequently the wavelength λ along the line is smaller than the wavelength λ_0 in free space, and since there is a fixed ratio between the wavelength and the length of some sections of line in a circuit with distributed elements, the whole circuit is smaller than it would be with no dielectric. For this reason it is preferable to use dielectrics with a high ε_r.

In microstrip with a homogeneous dielectric the electric and magnetic lines of force are perpendicular to the longitudinal axis of the line (transverse electromagnetic or TEM mode). This means that every point on the surface of a cross-section of the strip conductor has the same potential with respect to the ground plane and that current flows only in the longitudinal direction. The ratio of the two is uniquely defined as the local impedance and we can attribute to the line in the usual way a characteristic impedance Z. This depends on the dielectric; if Z_0 is the characteristic impedance of a line with a dielectric constant $\varepsilon_r = 1$ and a relative permeability $\mu_r = 1$, then we can write in general

$$Z = Z_0 \sqrt{\frac{\mu_r}{\varepsilon_r}}. \qquad (2)$$

A dielectric with a high ε_r therefore makes the characteristic impedance lower. Z_0 is determined by the geometry of the line, in particular by the ratio of the width w of the strip to the distance h between strip and ground plane; we call this ratio w/h the form factor.

Inhomogeneous dielectric

If there is no dielectric above the strip, as in fig. 4, we must introduce some approximations in calculating the characteristics of the line. The wave is no longer propagated along the line in the pure TEM mode. At lower frequencies, however, the deviation from the TEM mode is not very great, and as a first approximation it is therefore neglected [5]. A static electric field is assumed to exist between the strip and the ground plane, and an effective value $\varepsilon_{\mathrm{eff}}$ is determined for the relative dielectric constant. Since the lines of force run partly through air, $\varepsilon_{\mathrm{eff}} < \varepsilon_r$. At the edge of the strip the proportion of the field passing tnrough air forms a relatively large proportion of the total. If the strip is made wider without increasing the distance to the ground plane, in other words if the form factor w/h is increased, the share of this edge field in the total field decreases and $\varepsilon_{\mathrm{eff}}$ approximates more to ε_r.

In equation (1) we can now substitute $\varepsilon_{\mathrm{eff}}$ for ε_r (we assume that $\mu_r = 1$). The phase velocity and hence the wavelength are then a function of w/h (*fig. 5a*). To calculate the characteristic impedance we must substitute $\varepsilon_{\mathrm{eff}}$ in equation (2); both quantities Z_0 and $\varepsilon_{\mathrm{eff}}$ are functions of w/h, and fig. 5*b* shows the calculated variation of Z with w/h. This graph gives us an idea of the values that Z can assume; in practice w/h lies within the region given in the graph, since a strip that is too wide gives rise to undesired propagation modes of higher order, and a strip that is too narrow is difficult to make accurately enough.

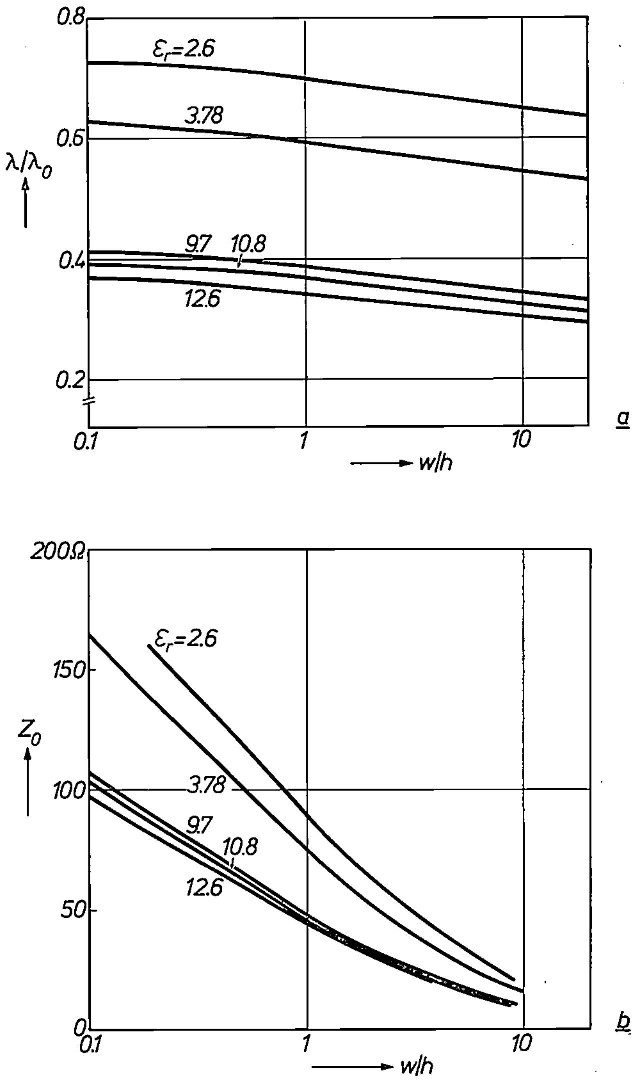


Fig. 5. *a*) The wavelength λ along a microstrip line, expressed on the graph as a fraction of the free-space λ_0, decreases as the width w of the strip becomes greater with respect to the thickness h of the dielectric. The field is then mainly contained in the dielectric. The figure shows curves calculated for the materials PTFE with glass fibre ($\varepsilon_r = 2.6$), quartz glass ($\varepsilon_r = 3.78$), aluminium oxide ($\varepsilon_r = 9.7$ and 10.8) and ferrite ($\varepsilon_r = 12.6$). *b*) The characteristic impedance Z_0 of a microstrip line also decreases as the ratio w/h increases. The figure shows the range of typical values of Z_0 for the same materials as in (*a*).

The deviations from a TEM field found in a microstrip line become greater with rising frequency. One effect is that the phase velocity decreases with rising

[4] J. H. C. van Heuven, *P-I-N* switching diodes in phase-shifters for electronically scanned aerial arrays, this issue, page 405.

[5] H. A. Wheeler, Transmission-line properties of parallel strips separated by a dielectric sheet, IEEE Trans. **MTT-13**, 172-185, 1965.
M. V. Schneider, Microstrip lines for microwave integrated circuits, Bell Syst. tech. J. **48**, 1421-1444, 1969.

frequency, causing the line to be dispersive (see *fig. 6*). When microstrip circuits were first made for frequencies of 10 GHz and higher, results were obtained that because of this effect differed fairly considerably from the static theory mentioned above. Dynamic theories were therefore developed for the microstrip line that take into account the axial field components and the transverse currents in strip and ground plane which are related to the deviations from the TEM field [6]. These theories enable the dispersion to be calculated.

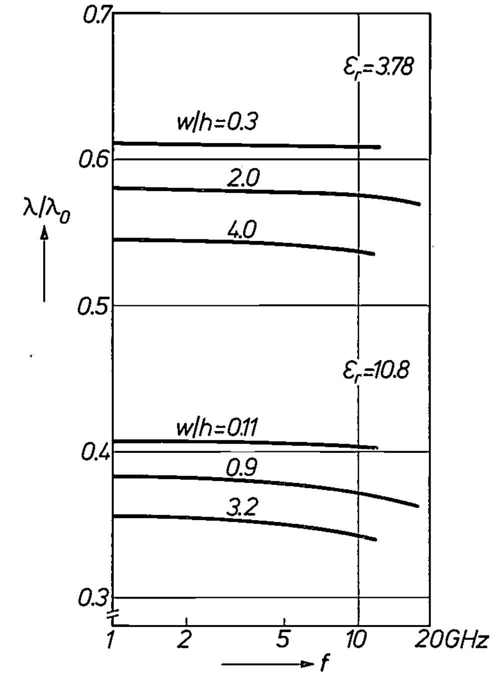


Fig. 6. Measurements at different frequencies *f* of the wavelength λ along a microstrip line. The wavelength has been normalized to the free-space wavelength λ_0. It can clearly be seen that the ratio λ/λ_0 and hence the velocity of propagation along the line decrease at higher frequencies. The measurements were made on lines with different form factors w/h and on substrates of quartz glass ($\varepsilon_r = 3.78$) and of aluminium oxide ($\varepsilon_r = 10.8$).

Losses

Losses in microstrip occur both in the conductors and in the dielectric. The conduction losses are reduced when w and h are both increased in the same ratio, since this has the effect of decreasing the current densities in the conductors. As the ratio w/h remains constant there is no change in the characteristic impedance of the lme. What can be achieved in this way is illustrated in *fig. 7*, which gives the attenuation of a wave propagated over one wavelength, in decibels, as a function of the width of the strip, for several values of the form factor w/h. We have already noted that there are limits to the extent to which the strip width can be increased.

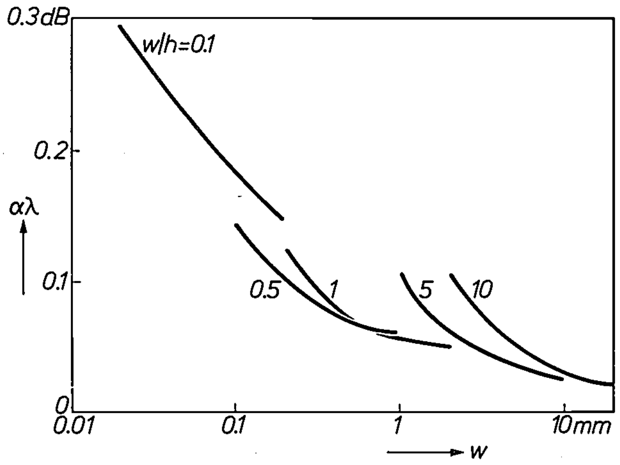


Fig. 7. The reduction of the attenuation $\alpha\lambda$ of a signal in a distance of one wavelength along a microstrip line as the cross-section of the microstrip line is increased in scale. (α is the line attenuation in dB/m and λ the wavelength in metres.) The parameter in each case is a particular form factor w/h, and hence a particular value of the characteristic impedance.

Fig. 8 shows the results of measuring the attenuation in one wavelength at different frequencies. Owing to the increase of the skin effect at higher frequencies, this attenuation does not decrease in proportion to the wavelength. Because of this same skin effect the roughness of the substrate surface becomes important when the surface irregularities approach the same order of magnitude as the depth of penetration of the currents in the conductor.

In addition to the conduction and dielectric losses, a microstrip circuit has other losses that arise because the electromagnetic energy is able to propagate in other ways in the circuit than along a microstrip line, e.g. over the surface of the substrate. These losses are mainly due to discontinuities in the microstrip line, which cause energy to propagate in unwanted directions; in circuits with many discontinuities these losses can be considerably larger than the conduction and dielectric losses. Discontinuities can also introduce undesired coupling between the elements of the circuit.

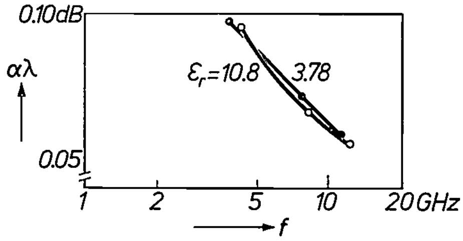


Fig. 8. Measurement of the attenuation $\alpha\lambda$ in one wavelength at different frequencies. The measurements were made on lines with a characteristic impedance of 50 Ω on two materials: quartz glass ($\varepsilon_r = 3.78$; thickness 0.5 mm) and aluminium oxide ($\varepsilon_r = 10.8$; thickness 0.635 mm).

To measure the characteristics of microstrip lines we used a length of line coupled by a slot to a coaxial connector at each end (*fig. 9*). A piece of line of this type is resonant at the frequencies at which its length is equal to an integral number of half-wavelengths (neglecting end-corrections). At resonance the transfer of energy from one connector to the other is a maximum. The phase velocity along the line can be derived from the resonant frequencies, and the attenuation of the line at these frequencies can be found from the Q (quality factor) of the resonances. To prevent the measurements from being affected by losses due to the unwanted modes of propagation mentioned above, the length of line is enclosed in a waveguide whose dimensions are such that wave propagation is possible only along the strip. A cut-away view of the waveguide is shown in the figure.

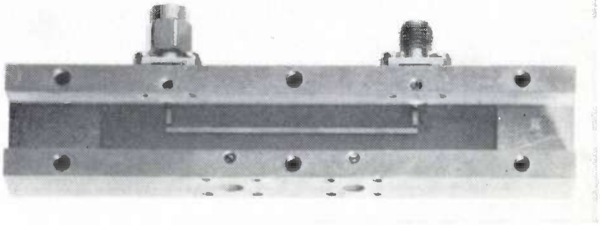


Fig. 9. Microstrip resonator for measurement applications. The length of microstrip line is connected between two coaxial connectors; when it is in resonance there is a maximum in the energy transfer from one connector to the other. The length of the microstrip line is then equal to an integral number of half-wavelengths (neglecting end-corrections). This resonator can be used for determining the phase velocity along the line and the attenuation due to the line. The resonator is built into a waveguide, shown in the cut-away view; the waveguide has dimensions for which wave propagation along the substrate and radiation are impossible, in this case only the conduction losses in the microstrip are measured.

Coupling between microstrip lines

Unless lengths of line are in resonance, there is very little interaction between microstrip lines farther apart than one substrate thickness. This allows a fairly wide latitude in the geometrical design of a microstrip circuit. Coupling between two microstrip lines is obtained by placing them close together. The total field of the two lines then consists of an odd and an even mode, as illustrated in *fig. 10* for the electric field. The system has different values of ε_{eff} for these odd and even modes of propagation, since their fields are not distributed in the same way between air and dielectric, so that the two modes have different phase velocities. This effect must be taken into account when designing components such as coupled sections of line.

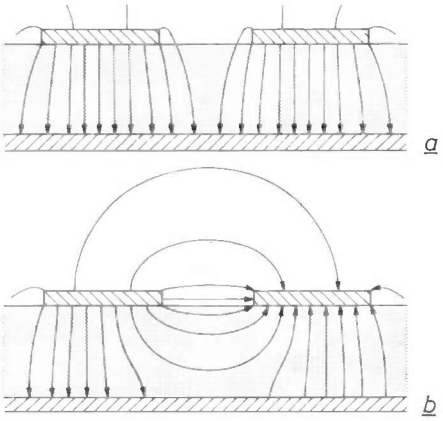


Fig. 10. In the field of two closely adjacent coupled microstrip lines there are two modes of propagation, an even mode (*a*) and an odd mode (*b*). The figure shows the electric field of the two modes.

Electrical characteristics of slotline

Among the other planar transmission lines shown in fig. 2, slotline deserves separate comment, since it has a field configuration (*fig. 11*) that looks particularly suitable for certain applications, though there has been little so far in the way of practical circuits. The field configuration of slotline shows some resemblance to that of a hollow waveguide, and like waveguide slotline gives marked dispersion. The electric field appears across the slot, which makes it very easy to connect impedances in parallel with the line, including semiconductor devices. The presence of longitudinal components in the magnetic field implies that a rotating magnetic-field vector is present at some points, which can be used for coupling to the gyromagnetic effects in ferrites; these effects are used in non-reciprocal microwave devices such as circulators [7]. In general it is most practical to couple into slotline by means of the magnetic field; an example is the coupling to microstrip lines formed on the back of the substrate of the slotline. The combination of slotline and stripline on opposite sides of the same substrate results in circuits of elegant simplicity [8].

Materials

Various materials can be used for the dielectric and conductors of a microwave integrated circuit. The conductors are usually of copper or gold.

Copper has the advantage of being easy to apply with the required accuracy, but it is not very resistant chemically. Copper is therefore coated with a layer of gold, which can also serve for bonding gold wires for

[6] See: E. J. Denlinger, A frequency dependent solution for microstrip transmission lines, IEEE Trans. MTT-19, 30-39, 1971, and the article by H. J. Schmitt and L. Ericsson, based on research at Philips Forschungslaboratorium Hamburg: Fundamentals of microwave integrated circuits, Elektronik (Sweden) 2, 53-57, 1971.

[7] M. Lemke and W. Schilz, Microwave integrated circuits on a ferrite substrate, this issue, page 315.

[8] Some work on this at Philips Research Laboratories, Eindhoven, has been reported in: F. C. de Ronde, A new class of microstrip directional couplers, G-MTT 1970 Int. Microwave Symp., Newport Beach, Cal., U.S.A., pp. 184-189.

connections to semiconductor devices. Semiconductor devices can also be directly alloyed on the gold layer. Gold has the advantage of being chemically resistant, but its resistivity is 1.4 times higher than that of copper.

The choice of the dielectric material depends partly on the nature of the circuit. Low losses are usually specified and in some cases a high ε_r as well, since this permits smaller dimensions. It is also important that the thickness of the dielectric should at no point deviate more than a few per cent from the nominal value, since such deviations will alter the characteristic impedance of the line. The surface must also be smooth, for the reason already mentioned. A survey of the properties of some substrate materials is given in *Table I*.

Table I. Dielectric materials for microwave integrated circuits.

Dielectric	Relative dielectric constant ε_r	Loss factor $\tan\delta$	Thickness variations $\Delta h(\%)$	Surface roughness $R_a(\mu m)$
PTFE with glass fibre	2.6	10^{-3}	3	—
Sintered aluminium oxide	10-11	10^{-4}	10	0.2
Sapphire	10-11	10^{-4}	1	0.01
Ferrite	12-16	$\leqslant 5\times10^{-4}$	1	0.02
Quartz glass	3.78	10^{-4}	1	0.01
Sintered beryllium oxide	6.5	2×10^{-3}	10	0.2

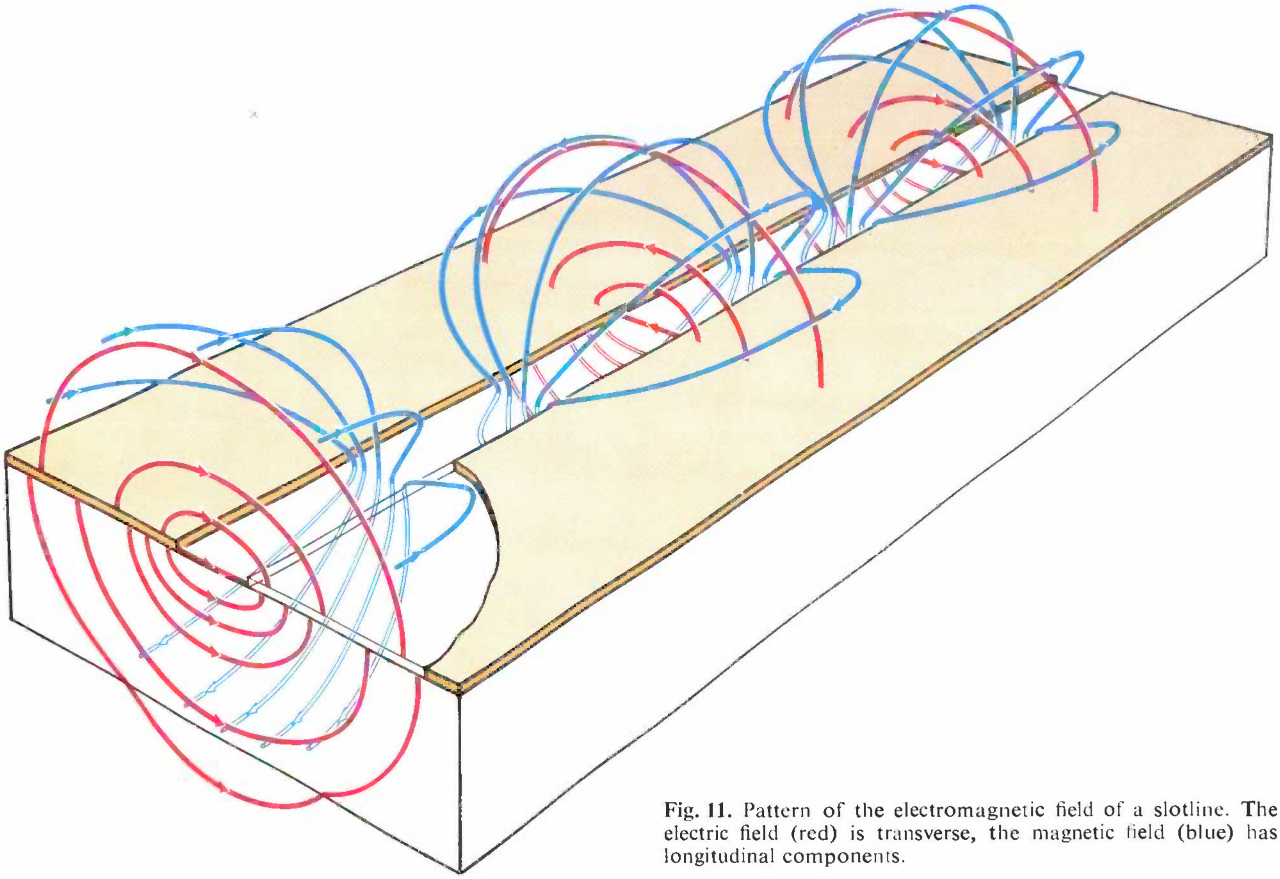


Fig. 11. Pattern of the electromagnetic field of a slotline. The electric field (red) is transverse, the magnetic field (blue) has longitudinal components.

PTFE with glass fibre is not so suitable for frequencies higher than 4 GHz because of its fairly large loss factor $\tan\delta$. Sintered aluminium oxide meets many specifications and is also relatively cheap; a disadvantage is that the sheets now available can vary in thickness or dielectric constant by 10%, either of which can give deviations of 3% in the characteristic impedance of a 50 Ω line. The surface roughness R_a is 0.2 μm, whereas at 30 GHz the depth of penetration of the current in a copper conductor is about 0.4 μm.

Sapphire is used for circuits where very high accuracy is required, because of the close tolerances within which the desired thickness can be obtained and because of its smooth surface. It also has a higher thermal conductivity than the sintered aluminium oxide. Disadvantages are that sapphire is anisotropic, very expensive and obtainable only in small pieces; its dielectric constant can also differ quite considerably.

Ferrimagnetic materials are used for non-reciprocal circuit elements; ferrites have also been developed that are not ferrimagnetic at room temperature and act as excellent dielectrics with a high ε_r. Further information will be found in another article in this issue [7].

Quartz glass has the advantage of a low loss factor and of an accurately known dielectric constant. A disadvantage — particularly for slotline — is its fairly low dielectric constant; another disadvantage is that it is brittle.

Finally, beryllium oxide is used where the dielectric is required to have particularly good thermal conductivity.

Technology

A pattern of conductors has to be formed on the substrate. To ensure firm adhesion of the conductors the surface of the substrate is first provided with an adhesion layer, e.g. a nickel-chromium layer about 5 nm thick. The adhesion layer is deposited by vacuum evaporation, and a 50 nm gold layer is immediately deposited upon it to prevent oxidation. Another method of obtaining good adhesion has been developed at these laboratories. In this method the surface is roughened by a combination of mechanical and chemical treatment; the increase in the conduction losses after treatment is well within the acceptable limits. Electroless deposition is then used to give the roughened surface a thin layer of copper, which acts as the adhesion layer. One of the advantages of electroless deposition is that holes in the substrate can be metallized to give a connection between the two sides of the substrate.

The conducting patterns can be applied to the adhesion layer by a photographic process. Two different procedures are illustrated schematically in *fig. 12.* In the first procedure the adhesion layer is electroplated with gold to a thickness of about 10 μm. A coating of photosensitive lacquer is then applied, which is exposed to ultra-violet radiation through a photomask. After photographic development a pattern is obtained in which lacquer is present at the places where the surface is to be conducting. The uncovered part of the conducting layer is then removed in an etching bath. In the second procedure the photosensitive lacquer layer is applied directly to the adhesion layer. This method differs from the first in that the lacquer is removed at places where the surface is to be conducting. The conductors are then grown by means of an electrolytic process, and the lacquer pattern ensures that conducting material is deposited only at the required places. After removal of the lacquer mask the underlying adhesion layer is etched away.

If an adhesion layer of nickel chromium has been used a resistance can be introduced in a microstrip line by not coating this adhesion layer with copper and gold over a certain length of the line. Resistances cannot be made in this way if the adhesion layer has been formed by the electroless deposition of copper.

The question arises as to how accurately the conducting pattern can be applied. In the first place there are the deviations in the photomask to be considered. A photomask is made by the photographic reduction of a drawing which is 10 to 25 times the true size. The drawing is made by a specially designed drawing machine, called a "coordinatograph", which can be controlled either manually or numerically. The repetition accuracy of a manually controlled coordinatograph is about 50 μm, that of the numerically controlled machine about 15 μm. With a reduction of 10 times the total drawing error is therefore 5 μm with manual control or 1.5 μm with numerical control.

Next, there are the optical-imaging errors caused during photography. In the equipment that we use errors of up to 10 μm can occur in a field of 5×5 cm. For a field of 2.5×2.5 cm this uncertainty is about 2 μm.

Finally, errors will occur during the photochemical forming process. As a rule the deviations due to exposure and development of the photosensitive lacquer are negligible. In the first method, however (fig. 12*a*) the underetching can give rise to considerable errors. *Fig. 13a* shows a strip and a slot made by this method. On the photomask the strip and the slot both have a width of 25 μm. Owing to underetching, however, the strip is 13 μm too narrow and the slot 20 μm too wide. The underetching is reasonably constant in any given etching procedure, and can therefore be taken into account when making the drawing. The pattern made by the second method (fig. 12*b*) is much more accurate, and moreover the error in the nominal value is least near the surface of the substrate, i.e. where deviations have the greatest effect (fig. 13*b*).

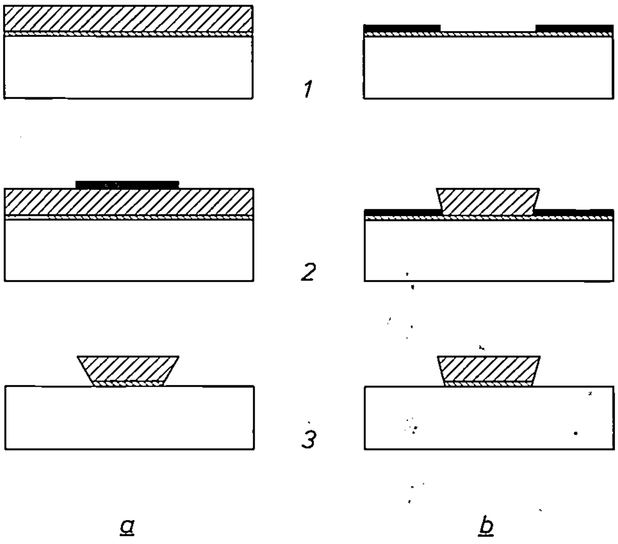


Fig. 12. Two photo-etching processes for applying the pattern of conductors to the adhesion layer that covers the substrate. *a*) 1. Gold is electrolytically deposited on the adhesion layer. 2. Where the surface is to be conducting a protective layer of lacquer is applied by a photographic method. 3. The surplus gold is etched away and the lacquer is removed. *b*) 1. A protective layer of lacquer is applied to the parts of the adhesion layer where the surface is not required to conduct. 2. Gold is electrolytically deposited on the rest of the layer. 3. After removal of the lacquer the adhesion layer is etched away outside the conductors.

Connections, encapsulation, adjustment

The reliability of conventional microwave circuits depends to a great extent on the quality of the connections between the waveguides or coaxial elements. Integration considerably reduces the number of these mechanical connections. Nevertheless some interconnections are still required, and these are often made in coaxial line. Transitions between microstrip and coaxial line are therefore required, and they are also important for connecting test equipment. These transitions can give rise to unwanted reflections and to losses. It has however been found possible to construct microstrip-to-coaxial transitions that give less than 3% reflection and low losses up to a frequency of 12 GHz.

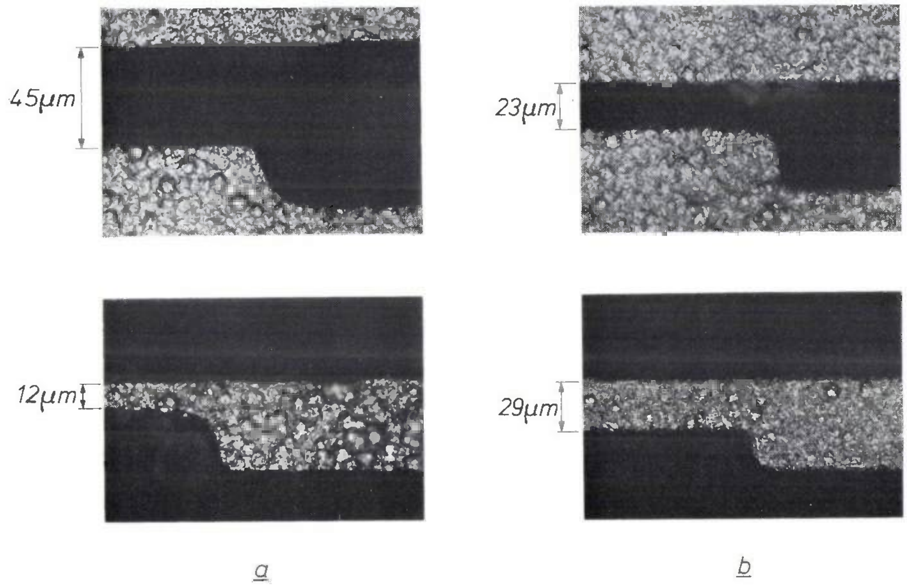


Fig. 13. Dimensional errors in the metal patterns; layer thickness 10 μm. *a*) A slot (above) and a strip (below) made by the process illustrated in fig. 12*a*. Both should have had a width of 25 μm; owing to underetching the slot is 20 μm too wide, the strip 13 μm too narrow. *b*) A slot and a strip made by the process illustrated in fig. 12*b*. Both differ from the specified width of 25 μm in the other direction, but the differences are much smaller.

The manner in which the semiconductor devices widely used in integrated circuits are attached depends partly on their heat dissipation. If the dissipated power is greater than half a watt it is preferable to attach the semiconductor device to the metal encapsulation of the circuit, which gives better cooling. A short wire or strip is then used to make the connection to the rest of the circuit on the substrate. If less power is dissipated the semiconductor chip can be mounted direct on the conductors. Beam-lead elements, which have connections that project outside the edge of the chip, are very suitable for this kind of mounting; *fig. 14* shows a Schottky-barrier diode with beam leads. Direct mounting on the conductors can be effected by the common techniques of ultrasonic or thermo-compression bonding, giving reliable connections with few strays. Resistors and capacitors on a semiconductor chip can also be provided with beam leads and mounted in the circuit as discrete components.

The encapsulation of a circuit calls for special care. Since some of the semiconductor devices are mounted unencapsulated, the entire circuit must be fitted in a moisture-proof and air-tight enclosure. Usually the

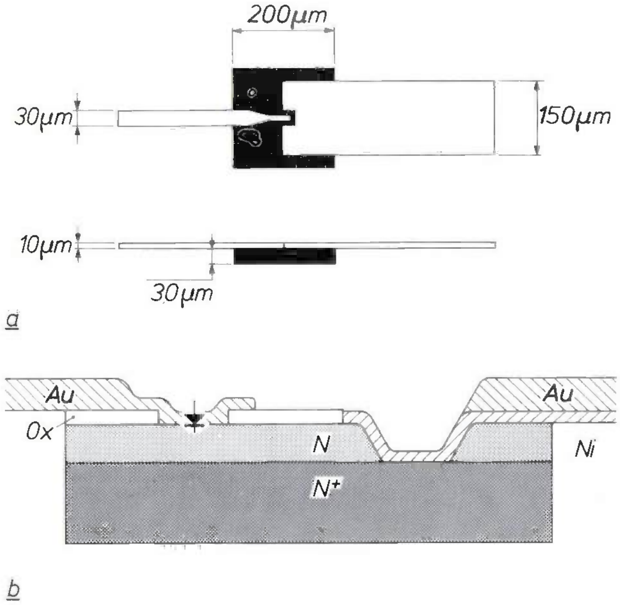


Fig. 14. Schottky-barrier diode with connections projecting outside the edge of the chip ("beam leads"). *a*) Plan and side view. The diode is directly connected to the strip conductors by the flat leads. *b*) Enlarged cross-section. *Ox* silicon oxide. *Au* gold. *Ni* nickel contact electrode (sometimes of titanium or palladium). A diode symbol indicates the polarity of the diode at the position of the rectifying contact.

encapsulation will be of metal to act as a screen against interference. A resonant cavity is thus formed above the circuit, and coupling with the circuit can set up undesired resonances in the cavity. This effect can be countered by the use of absorbing material. The number of possible resonances can be minimized by reducing the distance between the substrate and the screening. However, the screening should not be so close that the field distribution around the lines is affected.

Since there is a spread in the characteristics of the various circuit elements, particularly with the semiconductor devices, it may be necessary to have some method of adjustment. Continuous variables are especially difficult to obtain; with microstrip it is very hard to find any useful equivalent for the tuning screw in the wall of a waveguide. In some cases stepped adjustment may be possible, for example by fitting a number of capacitors and switching in as many of them as measurements show to be necessary. In some cases a variable capacitor can be included in the circuit, as will be illustrated below.

Some actual microstrip circuits

By way of example we shall now discuss some of the microwave integrated circuits that we have made. A common component of these circuits is the hybrid ring (*fig. 15a*). This consists of a ring of stripline with four lengths of stripline *1* to *4* connected to the circumference. The signal applied to port *1* divides into equal parts which go to ports *2* and *4*, if these have matched (non-reflecting) terminations; the signals arrive at these two ports in opposite phase. Port *3* is decoupled from port *1*. The behaviour of the hybrid ring corresponds to that of its low-frequency counterpart, the hybrid transformer (fig. 15*b*). The operation of the hybrid ring depends on the transit time of the signal along the ring. It only behaves in the way described when the circumference of the ring is equal to one and a half times the wavelength of the signal frequency, the path from port *1* to port *3* then being exactly $\frac{1}{2}\lambda$ longer than the other path. Fig. 15*c* shows the attenuation, as a function of frequency, of the signal transmitted from port *1* to ports *2*, *3* and *4*, measured for a hybrid ring on a quartz-glass substrate, designed for operation at 3 GHz.

Other common components of microwave circuits are band filters — both band-stop and band-pass filters — which are built up from resonant sections of stripline a

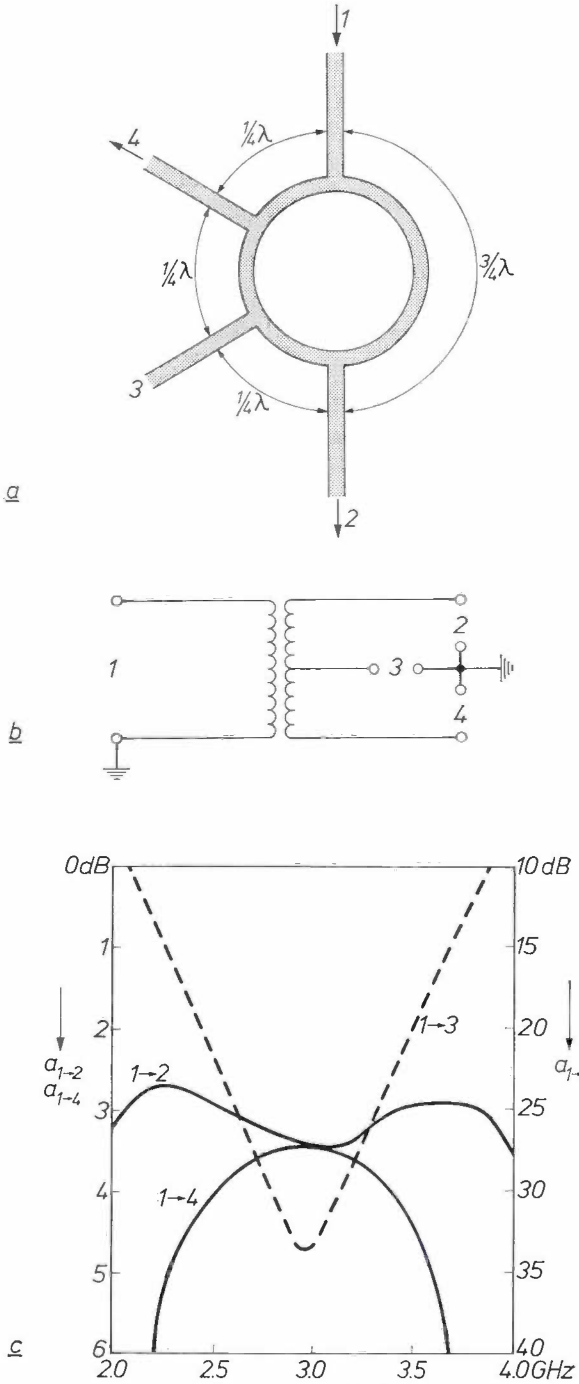


Fig. 15. *a*) Hybrid ring. The signal entering port *1* divides equally between two paths. The path to port *2* is equally long for both signals, so that they arrive in phase at port *2* where they reinforce one another. The path to port *4* is exactly one wavelength λ longer for one signal than for the other, so that both signals also arrive in phase at port *4*. The paths to port *3* differ by half a wavelength, and therefore the two signals arrive there in opposite phase and cancel; ports *1* and *3* are therefore decoupled. The power supplied to port *1* is transmitted in equal amounts to ports *2* and *4*. *b*) Hybrid transformer. The relations between ports *1* to *4* are similar to those in the hybrid ring. *c*) Measurements on a hybrid ring for 3 GHz, showing the insertion loss *a* during the transfer from port *1* to ports *2* and *4* (left-hand scale) and *3* (right-hand scale). The level at ports *2* and *4* is the same for 3 GHz, but the level at port *3* has a minimum of −34 dB compared with the level at port *1*.

quarter of a wavelength long (or a multiple thereof). *Fig. 16a* shows two practical versions of a filter with a quarter-wave resonator. At resonance the short piece of line has zero impedance at the end where it is connected to the stripline; it is then equivalent to a short-circuit across the line at that position, and as a consequence the frequency band around the resonant frequency is almost completely rejected (band-stop filter). Fig. 16*b* shows the attenuation *a* measured at a number of frequencies on a quarter-wave band-stop filter of the type shown in fig. 16*a*. The calculated curve found for such a filter is also shown; here the dispersion, edge effects and losses are not taken into account. This gives an indication of the errors that can be caused by the approximations. If the open end of the stripline resonator is connected to the ground plane the filter has a high impedance at the resonant frequencies but not to signals at other frequencies, which are therefore attenuated (band-pass filter).

If a response is desired with a narrower passband and steeper sides than this very simple filter will give, coupled half-wave resonators can be used. *Fig. 17a* shows a filter of this type for 12 GHz. The response is shown in fig. 17*b*, and for comparison the passband response as calculated from a simplified theory (dashed curve).

Fig. 18 shows a circuit containing the hybrid ring and the quarter-wave band-stop filter. It is a balanced mixer designed as the input stage of a 12 GHz receiver; a version in quartz glass can be seen on the left in the figure, and a version in aluminium oxide on the right. In this mixer stage the aerial input signal *SIG* is mixed with the signal from a local oscillator *LO*, giving a difference-frequency signal *IF*, which is further amplified in the intermediate-frequency amplifier of the receiver.

The aerial and oscillator signals are applied to the two decoupled ports of the hybrid ring *R*. Owing to the decoupling only a small part of the oscillator signal arrives at the aerial. Connected to the two remaining ports of the hybrid ring there are Schottky-barrier diodes *D* (of the type illustrated in fig. 14), in which the difference frequency is obtained because of the non-linearity of the diode characteristic. The band-stop filters F_1 form a short-circuit behind the diodes for the aerial and oscillator signals. The i.f. signal is extracted via the narrow striplines, one of which has to bypass another stripline with a small piece of wire. The fairly strong second harmonic of the oscillator frequency resulting from the non-linearity of the diodes is short-circuited by the resonators F_2, which are connected at the end to the ground plane. *Fig. 19* shows the diagram of an equivalent circuit for lower frequencies, built with lumped components. The sensitivity of the circuits

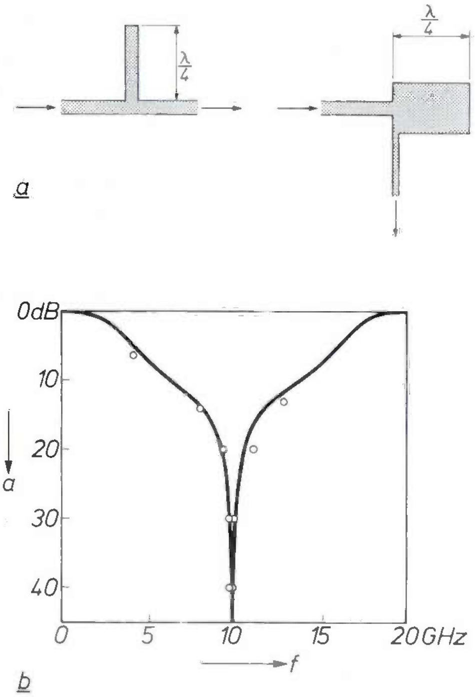


Fig. 16. *a*) Band-stop filters, consisting of quarter-wave resonators which have zero impedance at the resonant frequency at the junction point. *b*) Some measurements of the attenuation *a* of a filter as illustrated top right, for various frequencies. For comparison the dashed curve gives the calculated band-pass response, neglecting losses and strays.

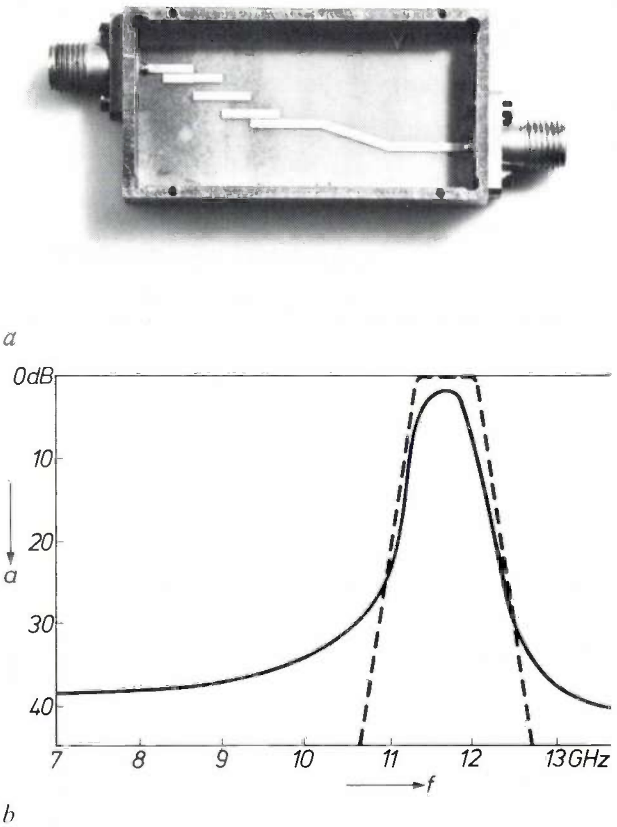


Fig. 17. *a*) Band-pass filter for 12 GHz, consisting of three coupled half-wave resonators. *b*) Band-pass response of this filter. Solid curve: measured response; dashed curve: calculated response, neglecting losses and strays.

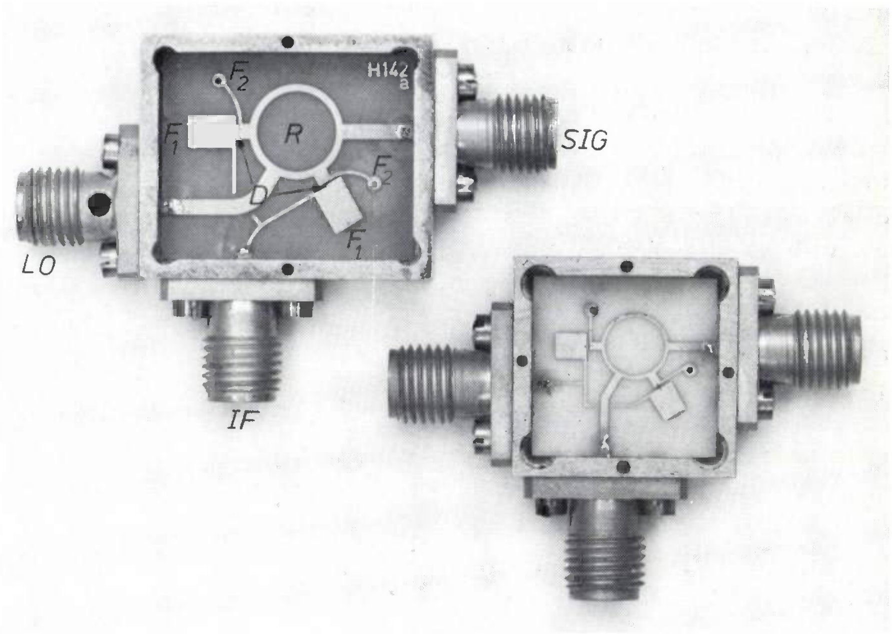


Fig. 18. Balanced mixer for 12 GHz, *left*: on quartz glass, *right*: on aluminium oxide. *R* hybrid ring. *D* Schottky-barrier diodes. F_1 band-stop filters for 12 GHz. F_2 band-stop filters for the second harmonic of the oscillator frequency. *LO* oscillator input. *SIG* aerial input. *IF* intermediate-frequency output.

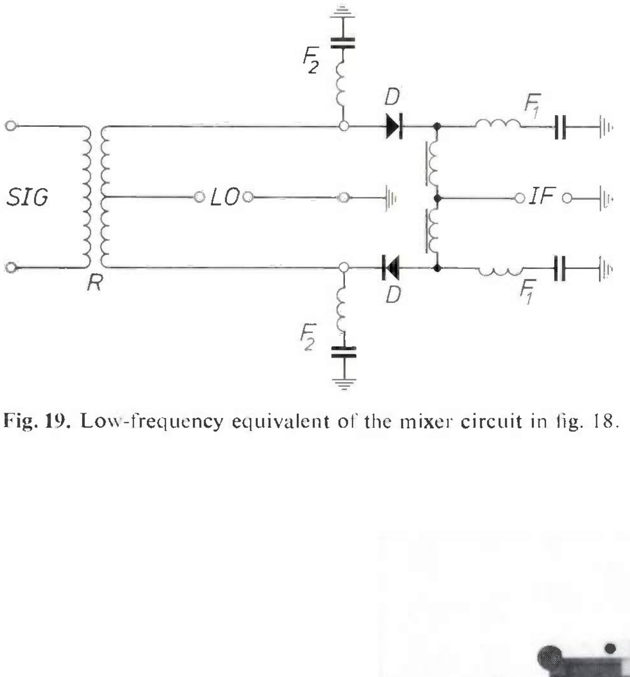


Fig. 19. Low-frequency equivalent of the mixer circuit in fig. 18.

in fig. 18 is comparable with that of the best waveguide mixer circuits; dimensions and weight, however, are much smaller. The measured noise figure is less than 11 dB.

This noise figure applies to the frequency band from 10 tō 12 GHz. The value quoted is the single-sideband noise factor measured with an i.f. amplifier that contributed 4 dB to the noise at an i.f. frequency of 500 MHz.

In *fig. 20* the balanced mixer is contained in a single encapsulation with a frequency multiplier, which supplies the oscillator frequency of 11.7 GHz. The fre-

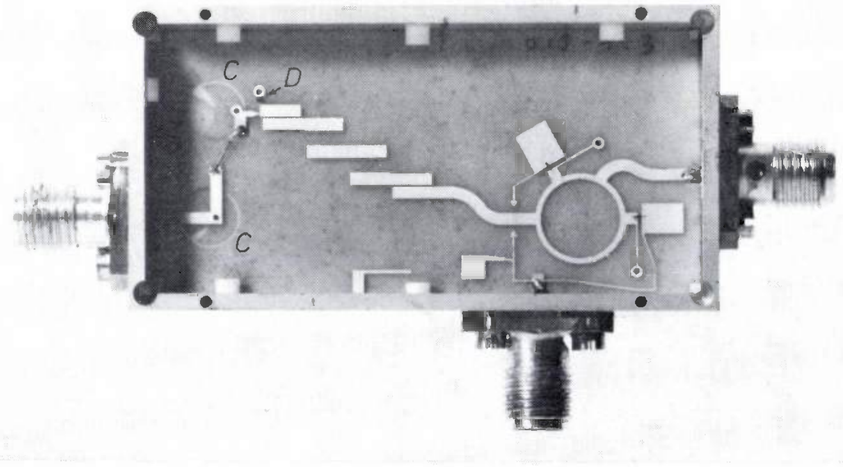


Fig. 20. Input stage of a 12 GHz receiver. The mixer circuit of fig. 18 is combined here with a frequency multiplier comprising a step-recovery diode *D*. A signal at a frequency of 1.17 GHz is applied to diode *D* through the junction on the left and a matching network containing two variable capacitors *C*. The tenth harmonic is filtered out by a band-pass filter and acts as the oscillator signal for the mixer stage.

quency multiplier contains a step-recovery diode *D*, to which a 1.17 GHz signal is applied. The diode produces strong higher harmonics of this signal; the 10th harmonic is filtered out by a selective band-pass filter of the type shown in fig. 17 and applied as oscillator signal to the mixer stage. In front of the diode there is a filter consisting of lumped components — an inductor and two variable capacitors; this filter provides the matching to the 1.17 GHz source. The circuit in fig. 20 is thus the complete microwave section of a 12 GHz receiver, and would be suitable for television receivers for the 12 GHz band now being studied for television transmission. The circuit is a good example of the compact construction that is possible with microwave integrated circuits. This compactness and the fabrication process described, which lends itself well to quantity production, open up prospects for large-scale applications of microwave components in applications such as television.

Summary. Besides waveguides and coaxial cables, flat conductors can also be used for microwave transmission. The most common configuration is the microstrip line, consisting of a narrow metal strip separated from a wide ground-plane conductor. The complete pattern of conductors of a microstrip circuit is formed by means of a photo-etching process on a dielectric substrate. The resultant integrated circuit is much smaller than the same circuit made in waveguide or coaxial line. Recently developed semiconductor diodes for oscillators, for example, are easily mounted on strip conductors, enabling active integrated circuits to be made as well. The characteristics of microstrip circuits are comparable with those of conventional circuits. Because of their small size, and their low cost if produced in large quantity these circuits are of considerable interest for applications where large numbers are necessary, as in television in the 12 GHz band. A circuit suitable for the input stage of a 12 GHz television receiver is discussed.

Lumped components for microwave frequencies

C. S. Aitchison

Introduction

Microwave circuit functions have been performed in the past by combinations of individual components built from transmission lines such as hollow waveguide or coaxial line, and carefully manufactured with precision tolerances. Such components are large compared with the wavelength and are described as "distributed". They contrast with the alternative lumped components, used at much lower frequencies where the component dimensions are very small compared with the wavelength.

Traditionally the transition from the use of lumped to distributed components has taken place in the region of 500 MHz to 1 GHz; but there is no reason why components for microwave frequencies, i.e. above 1 GHz, should not be constructed in lumped form — other than the possible physical inconvenience of so doing. Recent developments in photo-etching and vacuum-deposition techniques have made it possible to deposit lumped-circuit elements on a suitable backing surface (substrate) by evaporation. Both high-conductivity metals such as gold and copper and microwave dielectric materials such as silica can be deposited.

The advantages of fabricating microwave circuits in lumped form are the extremely small dimensions of the circuits thus produced and the rather inexpensive technology — even more so because microwave semiconductor devices are mounted into the lumped circuits as unencapsulated chips. It is interesting to consider the cost of the existing sequence of making a semiconductor chip and combining it with traditional distributed components. Normally the chip, costing a few tenths of a pound sterling, is encapsulated in a special microwave encapsulation which usually costs more than the chip, and it is the encapsulated chip that forms the output from a semiconductor factory and is purchased by the customer. He constructs around it a three-dimensional distributed circuit which normally costs many times more than the encapsulated chip. Thus the typical microwave circuit is expensive — so expensive that its use can only be considered in professional applications where the performance is more important than the price.

Lumped microwave circuit elements are a possible answer to this situation, but not necessarily the only one. Another possibility is the use of microstrip technology, which is dealt with extensively in other articles in this issue [1] [2]. Both techniques use a substrate on to which the conductors are evaporated; however it is important to differentiate between the two to avoid confusion. Microstrip is a distributed transmission-line medium in which a conductor is spaced from a ground plane by a material of low microwave loss and high dielectric constant so that most of the microwave energy is confined to the dielectric substrate. Thus the substrate has to have acceptable electrical and mechanical properties. A typical microstrip substrate will be some tens of square centimetres in area.

In the lumped case the main function of the substrate is to support the evaporated layers. The microwave energy within the substrate is small; thus the importance of the substrate electrical and mechanical properties is considerably reduced. Typical substrate areas are one square centimetre or less.

Viewed in the light of these facts lumped microwave circuits bear promise of an even smaller size and greater economy of production than is offered by microstrip circuits. This article describes a first survey of the field, consisting of the experimental assessment of the properties of individual lumped components — the five basic circuit elements are the inductor, the capacitor, the resistor, the transformer and the gyrator — and of simple circuits built up from these components with the addition of semiconductor devices.

The ultimate object of our work is to demonstrate the feasibility of manufacturing on one side of a glass (or similarly cheap material) substrate a complete microwave circuit with one or more microwave inputs, a number of d.c. connections and an output at a convenient intermediate frequency. Such a circuit should be completely contained on not more than one square centimetre of glass and typical examples might contain: a mixer with one or two Schottky-diode chips, a tunable local oscillator with a Gunn device and a Schottky diode, a preamplifier (parametric amplifier, tunnel diode or transistor amplifier), a number of d.c. con-

C. S. Aitchison, B.Sc., M.Inst.P., A.R.C.S., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] J. H. C. van Heuven and A. G. van Nie, Microwave integrated circuits; this issue, page 292.

[2] M. Lemke and W. Schilz, Microwave integrated circuits on a ferrite substrate; this issue, page 315.

nections isolated through low-pass filters, and an encapsulation for the complete circuit, together with one or more appropriate microwave connectors and a number of d.c. connectors (this encapsulation is not required to have microwave properties). The successful construction of a circuit at the research laboratory should enable the cost of quantity production to be estimated. Preliminary factory estimates of a simple 9 GHz Doppler radar including a circulator and Schottky detector suggest a factory selling price of about one fiftieth of the price with conventional components.

In the following sections the results obtained with individual passive lumped elements, measured at frequencies up to 12 GHz will first be described. Secondly, some simple filter circuits and a few active circuits will be described: the mixer, tunable oscillator and preamplifier mentioned above. Finally an account will be given of the Doppler radar.

Passive lumped elements

Substrate and measurement technique

The components made for the measurements are mounted on a quartz substrate, although the ultimate aim is to use a less expensive material. Quartz was used at the beginning of the investigations to avoid the complications that would arise with a less well known microwave material.

In *fig. 1* the arrangement is shown which has been adopted for examining the microwave characteristics of most of the passive lumped elements. On a quartz disc 9 mm in diameter and 0.5 mm thick, inner and outer gold connections are deposited so that the disc can be placed across the end of a coaxial line. This enables conventional microwave measurements to be made using standard coaxial measuring equipment. The element to be examined is evaporated within the area $AA'B'B$.

Inductor, capacitor and resistor

A typical inductor is shown in *fig. 2a*. It consists of one turn of evaporated metal. Both the diameter and track-width can be varied to give a range of inductance values from less than 1 to more than 3.5 nH.

The lumped capacitor is formed by the fringing field between an interdigital gap as shown in fig. 2*b*. The capacitance values which are obtainable range from 0.01 pF to 1 pF. Larger capacitance values can be obtained by means of a metal/silicon-dioxide/metal sandwich. These are avoided where possible since they involve more processing stages than the simple interdigital capacitor.

The inductance and capacitance of elements produced in this way have values that are particularly convenient for use with existing unencapsulated microwave semiconductor chips since they combine with the intrinsic and stray capacitances and inductances of these chips to give resonances in the range of 4 to 12 GHz, at least.

Lumped resistors are readily formed using evaporated nickel chromium, which is, in any case, used as a seed layer for the evaporation of gold and copper conductors.

Work on the lumped transformer is still in the experimental stage and will therefore not be described here. Initial results are not discouraging.

The characteristics of these lumped components are most conveniently measured by combining them in either a series or parallel circuit resonant between 4 and 12 GHz. *Fig. 3a* shows a series resonant circuit and fig. 3*b* shows a parallel resonant circuit. The parallel circuit is particularly convenient for measurement purposes because it presents a high impedance at resonance and thus is less dependent on contact losses which may

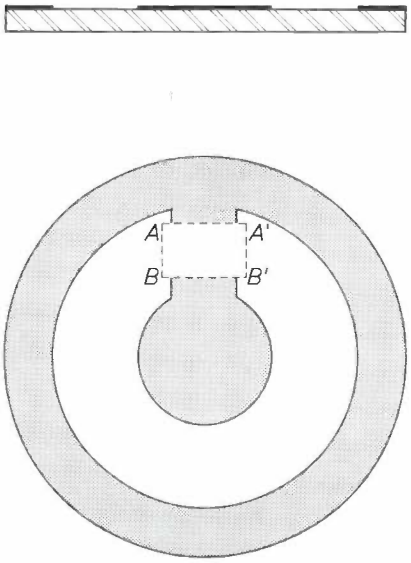


Fig. 1. A quartz disc, 9 mm in diameter and 0.5 mm thick, on which the experimental lumped passive elements are evaporated (inside the rectangle $AA'B'B$) for measurement purposes. The disc is provided with evaporated contact areas which enable it to be placed directly across a coaxial 50 Ω line for connection with the measuring equipment.

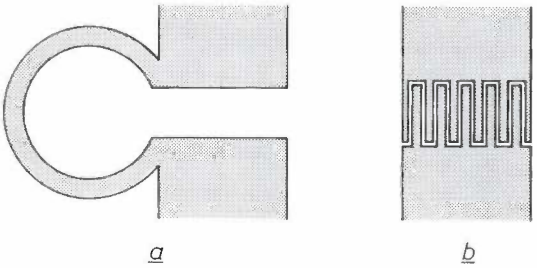


Fig. 2. *a*) Single-turn lumped inductor. *b*) Lumped interdigital capacitor. The gap is 20 μm wide.

occur because the sample is mounted in the coaxial measuring jig.

The parameters of these resonant circuits are determined by measuring the reflections they cause when connected to a transmission line of 50 Ω characteristic impedance. A microwave generator whose frequency is swept from 4 to 12 GHz feeds a signal into the line. The reflection coefficient of the circuit under test is plotted on the well known Smith chart [3]. *Fig. 4* gives an example of a Smith-chart plot for a parallel resonant circuit having a Q (quality factor) of 80 and resonating at 6.7 GHz. The resonant frequency or frequencies can be directly read from the plot as they are the frequencies at which the curve intersects the horizontal axis; fig. 4 shows, besides the parallel resonance at 6.7 GHz, a spurious series resonance at about 10.5 GHz. Additionally, the inductance L or the capacitance C and the Q can be deduced from the plot. Q depends on the losses of the circuit; these losses have been found to be almost exclusively due to the inductor.

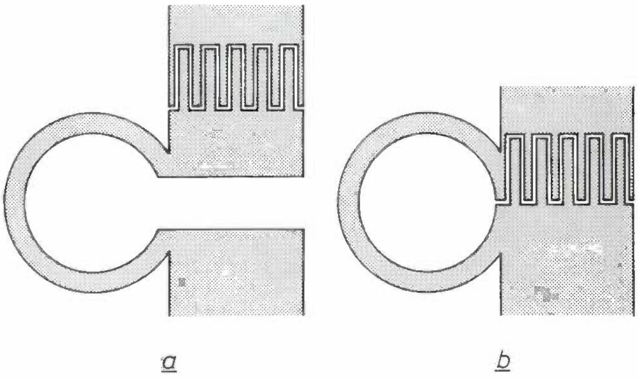


Fig. 3. *a*) Series-resonant circuit. *b*) Parallel-resonant circuit.

For a parallel-resonant circuit the admittance representation is the more convenient one; in this case the admittance is $Y = G + \mathrm{j}(\omega C - 1/\omega L) = G + \mathrm{j}B$. In this expression G is a parallel conductance which expresses the losses of the circuit. It can easily be shown that $(\mathrm{d}B/\mathrm{d}\omega)_{\omega_0} = 2C$; the value of $\mathrm{d}B/\mathrm{d}\omega$ at the angular resonant frequency ω_0 can be found graphically from the Smith-chart plot and gives us the value of C. Since the resonant frequency is known L can now be calculated from $\omega_0{}^2 = 1/LC$. At resonance $Y(\omega_0) = G$, whose value can also be directly read from the plot. Q is then found from $Q = (C/L)^{\frac{1}{2}}G$.

The losses of the circuit may be expressed by a series resistance R for the inductor. R is not equal to $1/G$ but can be easily calculated from L, C and Q.

In *fig. 5* an example is given of a Smith-chart admittance plot of a parallel-resonant circuit in series with a lumped resistor designed to have a value of 50 Ω.

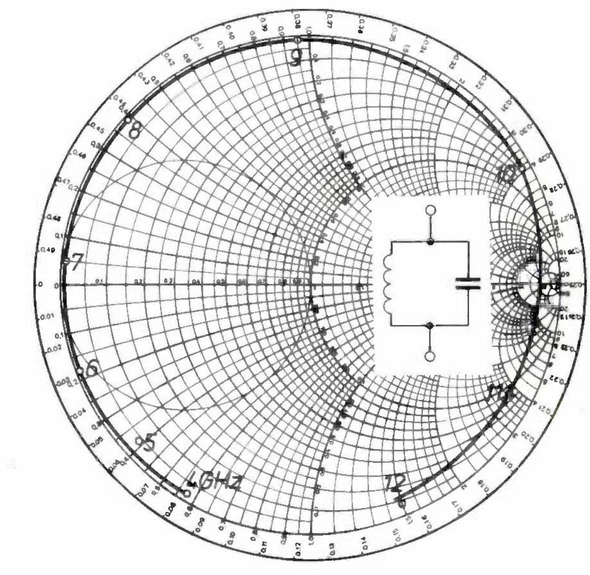


Fig. 4. Admittance plot of a lumped parallel-resonant circuit on a Smith chart. The Smith chart is essentially a polar plot of the reflection coefficient of the circuit, which is what is usually measured; however, it is calibrated in such a way that the real and imaginary parts of the admittance can be read off for any of the frequencies of measurement. The admittance is normalized to the characteristic admittance of the transmission line on to which the circuit has been connected, in this case 0.02 S [4]. On the horizontal line the imaginary part of the admittance is zero; above the horizontal line it is positive, below the line it is negative. At about 6.7 GHz the measured curve intersects the horizontal line at a minimum admittance value, indicating that this is the frequency of the parallel resonance. At this frequency the admittance is about 0.01 × 0.02 S; from this value and the capacitance and inductance, the Q (quality factor) can be calculated to be 80. Between 10 and 11 GHz there is a series resonance with an admittance of more than 20 × 0.02 S.

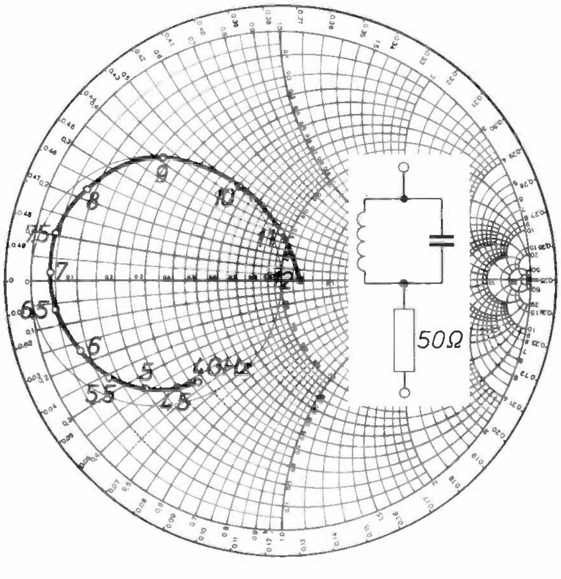


Fig. 5. Smith-chart admittance plot of a parallel-resonant circuit with a lumped series resistor with a design value of 50 Ω; the admittance is again normalized to 0.02 S. The resonant frequency is 6.9 GHz; at high and low frequencies the curve approaches the centre of the diagram, which means that the admittance approaches a pure conductance of 0.02 S, the design value for the resistor.

[3] P. H. Smith, Transmission line calculator, Electronics **12**, Jan. 1939, 29-31; An improved transmission line calculator, Electronics **17**, Jan. 1944, 130 *et seq.*

[4] In this article and elsewhere the siemens (S) has been adopted for mho or Ω^{-1} in accordance with international recommendations.

To assess the losses associated with the lumped inductors 20 parallel-resonant circuits were measured. The capacitance was nominally equal to 0.35 pF in each case and the track-width and outside diameter of the single-turn inductors were varied. Assuming the losses in the resonant circuit to be exclusively due to the inductor, the inductor resistance could be deduced from the Q-factor of the circuits. Results are presented in *figs. 6a* and *b*; in fig. 6*a* the series resistances that were measured at different resonant frequencies are plotted; three classes of inductors with outside diameters of 0.6, 0.8 and 1.2 mm were involved. In fig. 6*b* the resistance is plotted as a function of the inductance and theoretical curves are included for comparison. It can be seen that the resistance achieved is approximately twice as large as theory predicts; it is likely that this difference arises because the calculation is based on the normal bulk values of resistivity for the conductor metals without taking into account the particular structure of the evaporated and plated metal.

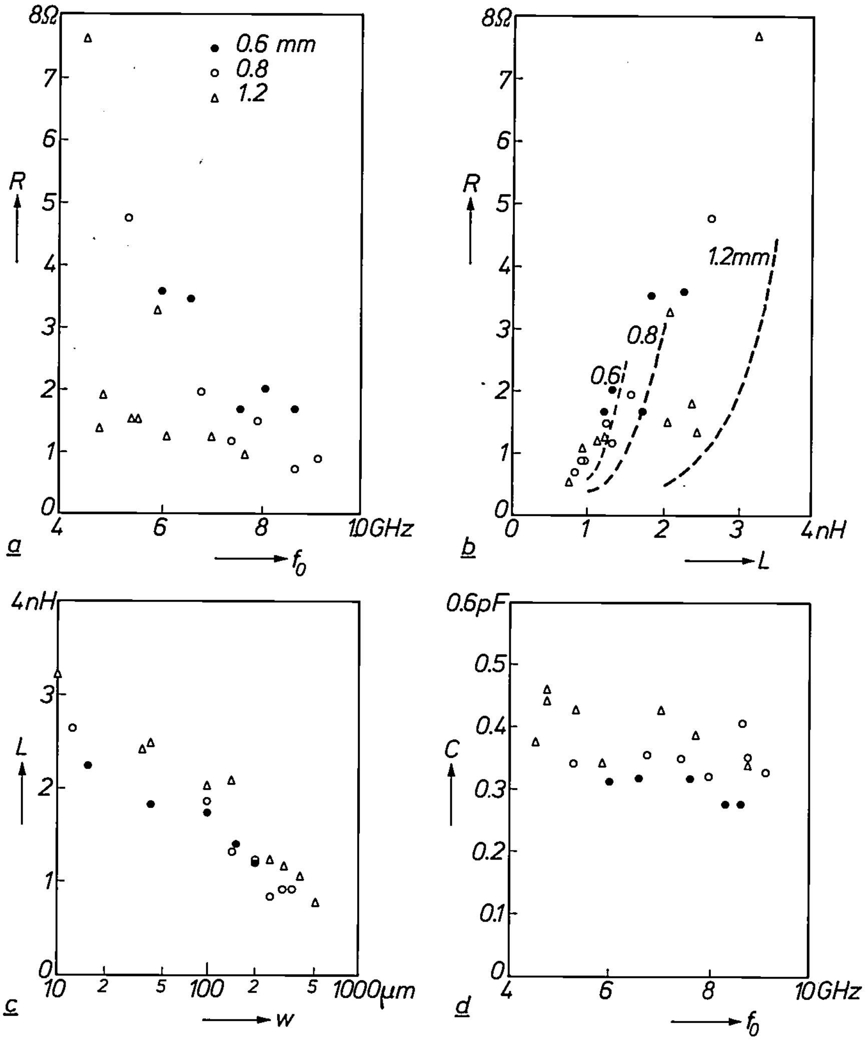


Fig. 6. Measurements on 20 lumped resonant circuits. The inductors of the circuits are divided into three classes with outside diameters of 0.6 mm (*dots*), 0.8 mm (*circles*) and 1.2 mm (*triangles*). *a*) Series resistance R of the inductor for circuits of various resonant frequencies f_0. *b*) Series resistance R for inductors of various inductances L. Computed curves are included for comparison. R turns out on average to have twice the computed value; the large experimental values may be due to the fact that the conductors are evaporated and plated. *c*) The inductance L for different track-widths w. *d*) The measured capacitance C for circuits of different resonant frequency f_0. All the capacitors were nominally identical and the design value was 0.35 pF. A large part of the scatter in the measured values is accounted for by the measurement error which is estimated at ± 25%, so that the reproducibility of the capacitors appears to be fairly good.

The results demonstrate that lumped-element parallel-resonant circuits with single-turn inductors can be fabricated for operation from 5 to 10 GHz with resistance values of about one ohm or less, corresponding to Q values of between 10 and 90. To understand the significance of the resistance values measured for the inductors it should be kept in mind that for applications where a low noise figure is required the loss in the inductor should be small compared with the loss in the semiconductor. This condition is normally met with lumped components. The same applies for conditions in which power handling is of importance.

The measurements on the twenty resonant circuits provide us with more data. In fig. 6*c* the measured dependence of the inductance on the inductor trackwidth *w* is shown; from fig. 6*d* an impression can be gained of the reproducibility of the interdigital capacitors. The scatter of the measured capacitance values is not much larger than the scatter accounted for by the limited measuring accuracy.

In conclusion, capacitors, inductors and resistors have been made which exhibit lumped-element characteristics up to 12 GHz (this is not the limit of the technique). The loss associated with the lumped reactances has been adequately small and is compatible with their use with semiconductor chips.

Gyrator

The gyrator, which is the fifth basic circuit element, appears in practical circuit form as the circulator and the isolator [2]. At microwave frequencies junctions of transmission lines, including ferrite material, form circulators and these are distributed structures. At frequencies below 3 GHz lumped circulators have been made by using a block of ferrite and winding coupled inductances on this with a geometrical spacing of 120°.

At microwave frequencies this design of lumped circulator is not attractive; a corresponding structure with planar conductors has been designed and is shown in *fig. 7*. It consists of a symmetrical pattern of intersecting conductors on ferrite; the gaps at the intersections are bridged by bonded wires to remove the d.c. isolation. This pattern terminates three 50 Ω microstrip lines and is contained within a circle of 0.8 mm diameter.

The performance of this circuit is shown in *fig. 8* as a function of frequency showing that an isolation, $a_{1\to3}$, of 18 dB is obtained with an insertion loss, $a_{1\to2}$, of less than 1.2 dB. The performance is available over the band 5 to 7 GHz by variation of the applied magnetic field.

A simpler unsymmetrical pattern is shown in *fig. 9* where it can be seen that two conductors pass from one port to the other two ports, giving a single intersection. The performance is similar to that obtained with the symmetrical structure.

The details of the operation of these circulator configurations are being studied. Later models of the symmetrical and unsymmetrical circulators have been made without the d.c. isolation so that there was direct contact at the junctions. No significant change in microwave behaviour was observed in either case.

Filters

When lumped-circuit elements and semiconductor chips are combined to form an active circuit it is necessary to supply and remove d.c. or low-frequency a.c. energy without disturbing the microwave energy. This

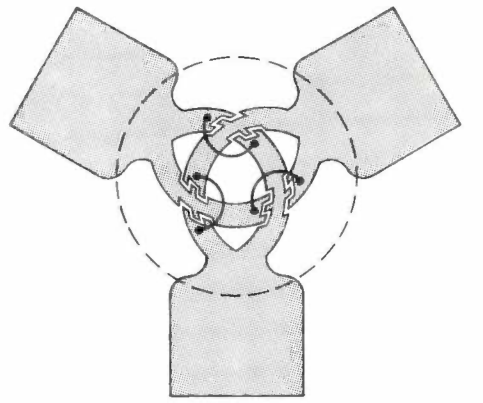


Fig. 7. Lumped circulator consisting of a symmetrical pattern of conductors on ferrite. Outside the circle the conductors are 50 Ω microstrip lines; inside the circle the ground plane has been removed. Bonded wires bridge the gaps for d.c. interconnection.

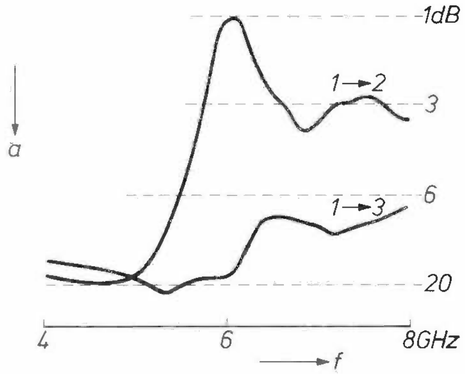


Fig. 8. Pen recordings of insertion loss $a_{1\to2}$ and isolation $a_{1\to3}$ of the symmetrical lumped circulator at frequencies *f* between 4 and 8 GHz. The optimum performance is $a_{1\to2} = 1.2$ dB, $a_{1\to3} = 18$ dB; it is available over the band 5 to 7 GHz by variation of the applied magnetic field.

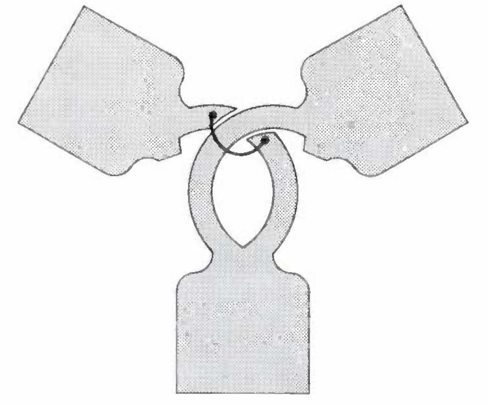


Fig. 9. Unsymmetrical lumped circulator, also for a frequency band centred on 6 GHz. It shows an insertion loss $<$ 1 dB, an isolation $>$ 20 dB and a bandwidth of about 2%.

is conveniently done by means of simple low-pass filters with lumped inductors and capacitors for the reactive components.

A simple three-element filter, as indicated in the inset of *fig. 10,* has been constructed with a series inductance of 1.9 nH and a shunt capacitor of 1.0 pF. The circuit consists of two single-turn inductors and an interdigital capacitor. With a 50 Ω termination this should give a 3 dB cut-off frequency of 5.5 GHz and 20 dB isolation at 10.5 GHz. Fig. 10 shows reasonable agreement between the computed and measured values.

The circuit diagram of a simple band-stop filter is shown in the inset of *fig. 11;* it consists of a cascaded combination of a series-resonant circuit in shunt and a shunt-resonant circuit in series. The 2 Ω series resistances of the inductors result from the loss in these elements. The insertion-loss performance is shown in fig. 11 together with the computed value with the system terminated in 50 Ω. An adequately high value of insertion loss (30 dB) is obtained at 9 GHz.

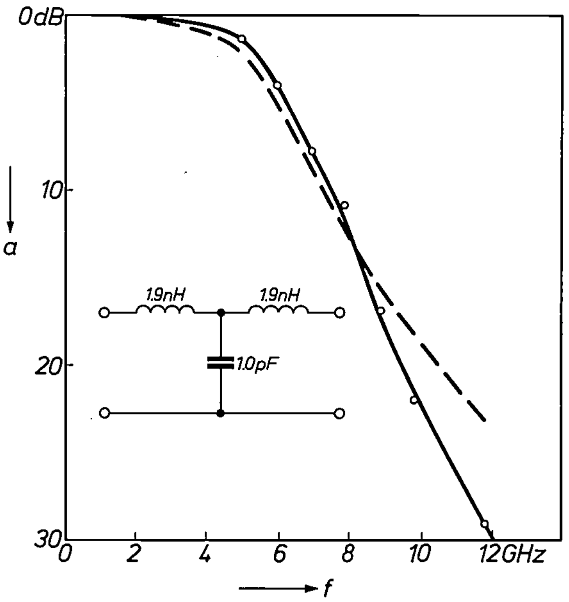


Fig. 10. Insertion loss *a* plotted against frequency *f* for the lumped low-pass filter shown in the inset. The dashed curve gives the computed frequency response.

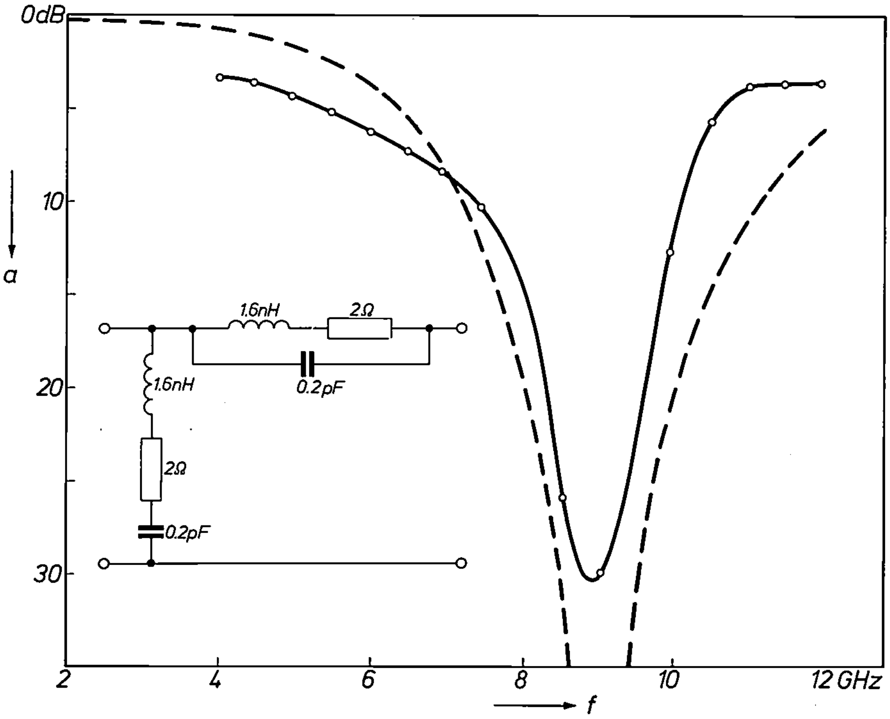


Fig. 11. Insertion loss of the lumped 9 GHz band-stop filter shown in the inset. The dashed curve gives the computed frequency response.

Active circuits

Tunnel-diode amplifier

The first active circuit which was constructed with lumped elements and a semiconductor chip was a 4 GHz tunnel-diode amplifier. This circuit was chosen because of its simplicity.

In *fig. 12* a circuit diagram is shown of a tunnel-diode amplifier connected to a microwave system by means of a circulator. There are three basic circuit blocks that make up the tunnel-diode amplifier. Block *1* is a transforming network which transforms the standard impedance of the transmission line (usually 50 Ω) to the value required to give the desired gain. Block *2*

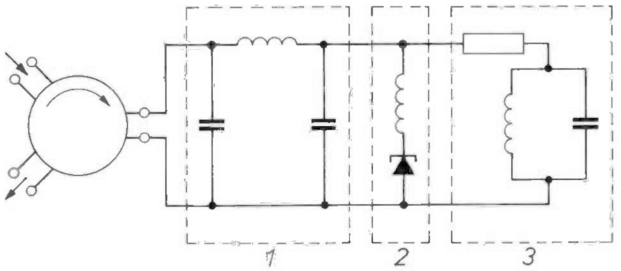


Fig. 12. Circuit diagram of lumped tunnel-diode amplifier connected to a circulator. *1* transforming network matching the amplifier to the line impedance. *2* tunnel diode with resonating inductor. *3* stabilizing network.

is the tunnel-diode chip resonated by an inductor, and block *3* is a stabilizing network which removes the negative conductance of the tunnel diode at all frequencies except in the band over which gain is required. The stabilizing network consists of a resistor of appropriate value in series with a parallel-resonant circuit which resonates at the operating frequency, thus removing the loss from the tunnel-diode circuit at this frequency (the Smith-chart plot of fig. 5 applies to a stabilizing network of this type). *Fig. 13* is a drawing of the amplifier in the coaxial configuration; a photograph is shown in *fig. 14*.

For measurements of the gain-frequency response the amplifier was connected to a broad-band circulator. *Fig. 15* is a pen recording showing the gain variation over the band 2.6 to 4.2 GHz with 3, 9 and 12 dB calibration lines. These are recorded on the chart by introducing a known fixed attenuation. It can be seen that the response is double-humped; this effect is produced by the impedance variation of the wide-band circulator. The peak gain is approximately 12 dB and 9 dB of gain is obtained over almost all of the band from 2.8 to 4.1 GHz. The measured noise figure is 6.2 dB.

It is concluded that the technique of lumped elements and unencapsulated chips is suitable for the construction of practical microwave tunnel-diode amplifiers.

Varactor-tuned Gunn oscillator

The frequency of a Gunn-effect oscillator [5] can be varied by connecting a varactor (a voltage-dependent capacitor) in series with the Gunn device and varying the capacitance of the varactor by means of the bias voltage applied to it. This provides a simple circuit for a tunable oscillator; an equivalent-circuit diagram incorporating the strays is shown in *fig. 16*. Calculation enables the parameters of the varactor to be specified

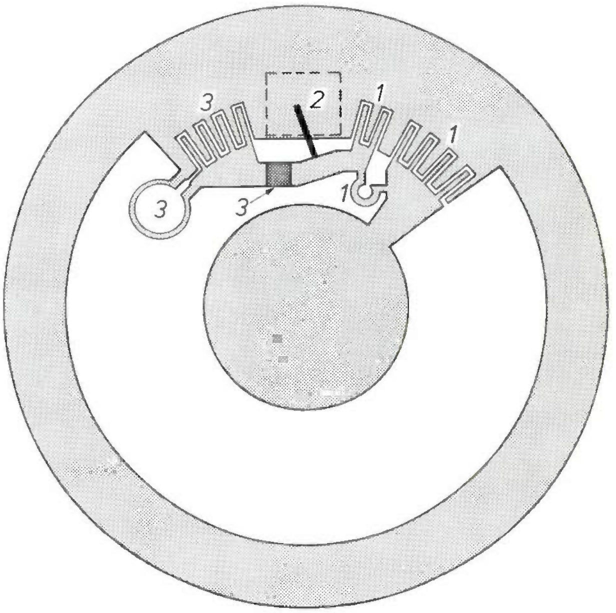


Fig. 13. Lumped tunnel-diode amplifier on standard 9 mm disc. *1* transforming network. *2* area for mounting tunnel diode and bonding wire serving as an inductor. *3* stabilizing network with nickel-chromium resistor.

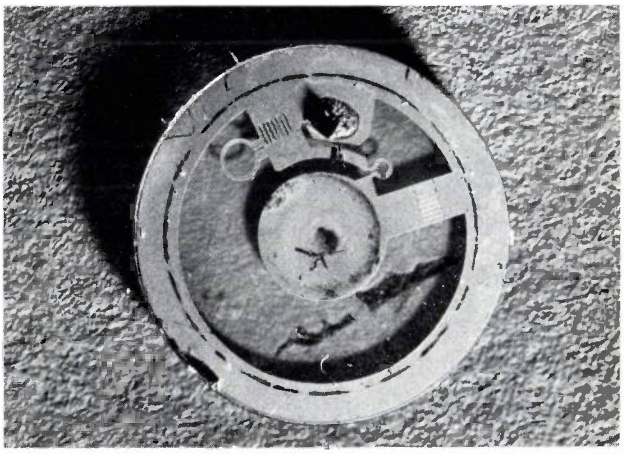


Fig. 14. Lumped tunnel-diode amplifier. In this example the transforming network has only one capacitor.

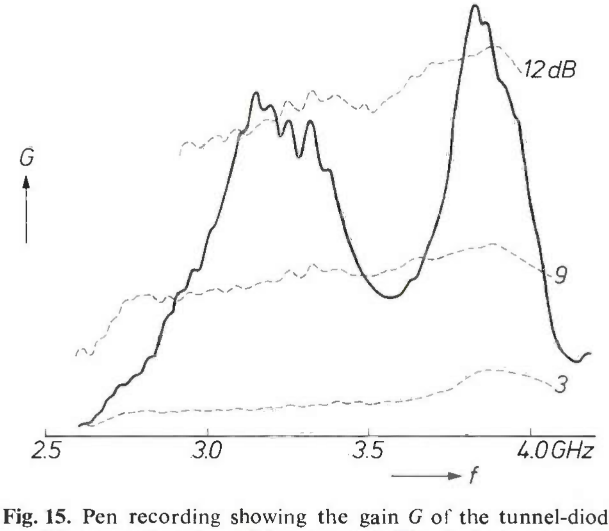


Fig. 15. Pen recording showing the gain *G* of the tunnel-diode amplifier as a function of frequency *f*. The dashed curves are calibration curves recorded for the measuring arrangement.

[5] G. A. Acket, R. Tijburg and P. J. de Waard, The Gunn effect; this issue, page 370.
J. Magarshack, Gunn-effect oscillators and amplifiers; this issue, page 397.

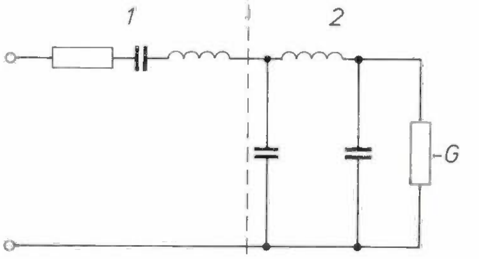


Fig. 16. Equivalent-circuit diagram of varactor-tuned Gunn oscillator, incorporating stray components. *1* varactor. *2* Gunn device. $-G$ is the negative conductance of the Gunn device.

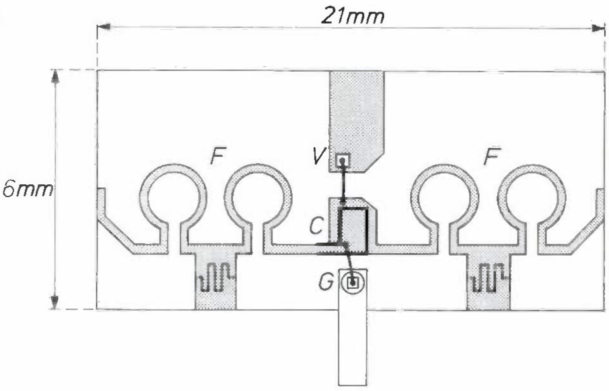


Fig. 17. Varactor-tuned Gunn oscillator with lumped low-pass filters *F* for connection to d.c. bias. *V* varactor. *C* overlay capacitor providing d.c. isolation. *G* Gunn device mounted on heat-sink tab.

for this application. In practice, for frequencies about 9 GHz, existing varactors such as the Mullard CXY 10 and other varactors with a similar high figure of merit are suitable provided they can be mounted in chip form.

Means must be provided for supplying a d.c. bias to both the Gunn device and the varactor chip. The lumped-element low-pass filters previously described are inserted in the bias leads; they enable us to realize the whole tunable oscillator circuit within very small dimensions. *Fig. 17* shows the actual component disposition; the Gunn chip *G* is mounted on a heat-sink tab and a series overlay capacitor *C* is provided for d.c. bias isolation.

For the purpose of measurements this circuit was connected to a microstrip circulator used as a microwave isolator by loading one of the arms with 50 Ω. A photograph of the arrangement is shown in *fig. 18*. This shows clearly the two low-pass filters as well as the Gunn and varactor chips. The circuit arrangement here is slightly different from fig. 17 so that the overlay capacitor can be omitted. The tuning range obtained with this arrangement is 1.1 GHz together with a maximum power output of 20 mW. *Fig. 19* shows the variation of frequency and power with varactor bias.

The conclusion is that the lumped-element technique yields an extremely compact Gunn-oscillator circuit of satisfactory performance.

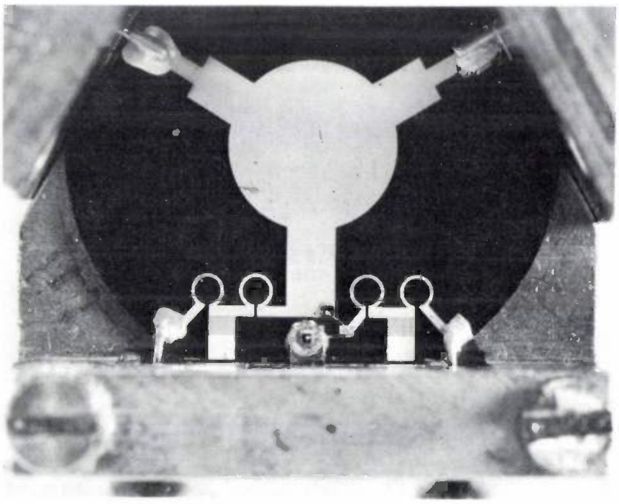


Fig. 18. Varactor-tuned Gunn oscillator for frequencies of about 9 GHz. A lumped oscillator circuit and a microstrip circulator are integrated on the same substrate. The overlay capacitor shown in fig. 17 can be omitted here.

Degenerate parametric amplifier

Another simple circuit that has been fabricated in lumped form is a degenerate parametric amplifier. This consists simply of a series-tuned circuit formed from a varactor and a resonating inductance. *Fig. 20* shows the experimental arrangement which was used: a varactor *C* and a three-turn lumped inductance *L* mounted on the standard 9 mm diameter disc.

The degenerate parametric amplifier requires a source of "pump" power at a frequency which is twice the signal-circuit resonant frequency. The pump voltage periodically changes the capacitance of the varactor by which means amplification is obtained [6] [7]. The

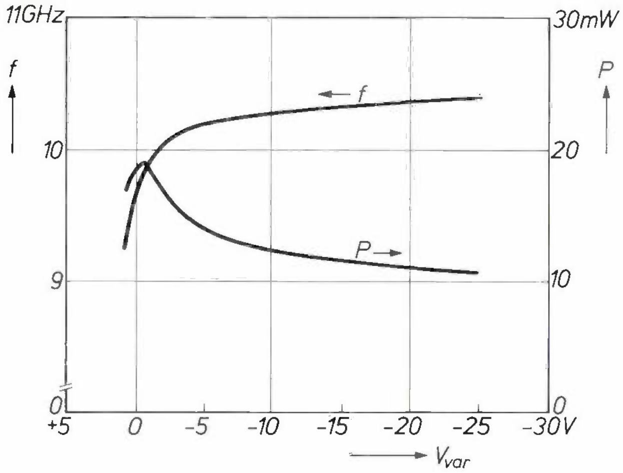


Fig. 19. The variation of frequency *f* and output power *P* of the varactor-tuned Gunn oscillator with varactor bias V_{var}.

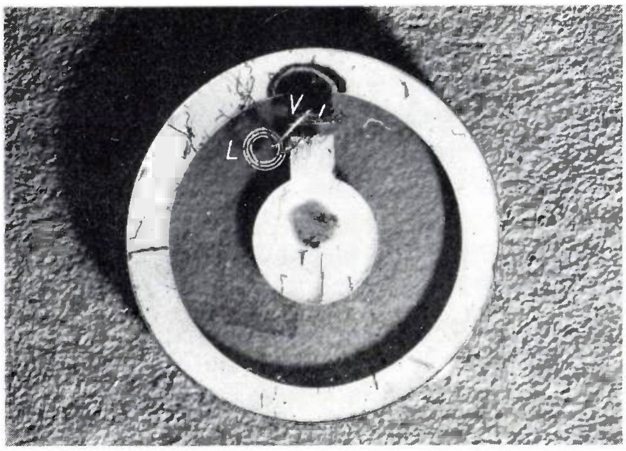


Fig. 20. Experimental configuration for lumped-element degenerate parametric amplifier constructed for standard 9 mm test mount. *C* varactor. *L* three-turn coil.

circuit is said to be degenerate because it is a special and simplified case of the general parametric amplifier which includes an "idling" circuit tuned to the difference frequency of pump and signal source. In the degenerate case this difference frequency coincides with the signal frequency and a separate idling circuit is not required.

The gain of the lumped parametric amplifier was lower than planned; the cause of this was found to lie with the type of varactor used. Nevertheless, the lumped-element technique appeared to be a satisfactory way of making a parametric amplifier.

Our parametric amplifier was equipped with a Schottky-barrier varactor. After it was mounted in the circuit this varactor turned out to have only a moderate figure of merit (3 to 10 GHz, which is low compared with 40 GHz for the CXY 10 mentioned above). As a result of this the gain was lower than planned; 8.0 dB of gain was obtained at 3.1 GHz with a bandwidth of 210 MHz. The measured amplifier noise temperature [7] was approximately 200 K. A conclusion resulting from this experiment is that the figure of merit obtainable from Schottky barriers is low and it would therefore be sensible to use a mesa diode which is bonded to the circuit and etched to the required value on the circuit.

Doppler radar

A Doppler-radar circuit operating in the band 8 to 11 GHz was the first attempt to make a complete subsystem. It was chosen because of its simplicity and also because of its market potential.

A Doppler radar transmits a beam of continuous-wave energy. Energy reflected by objects in the path of the beam is received by the aerial and mixed with part of the transmitter signal in a mixer circuit. If the objects move either towards or from the aerial the frequency of the reflected wave is shifted towards higher or lower values respectively. This is known as the Doppler effect. In this case a difference frequency proportional to the velocity of the moving object in the direction of the aerial appears in the mixer circuit. Doppler radar is thus very suitable for the detection of moving objects and for the measurement of their speed; a common application is in the measurement of the speed of motor vehicles [8].

Our Doppler-radar circuit consists of three parts (*fig. 21*): a Gunn oscillator *G*, a mixer stage with a

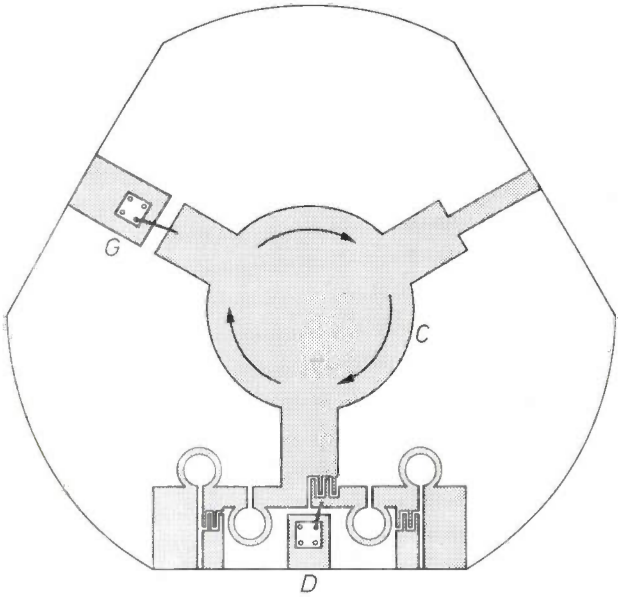


Fig. 21. Lay-out of Doppler-radar circuit. *G* Gunn device. *C* circulator. *D* Schottky-barrier diode of mixer stage. The right-hand port of the circulator is connected to the aerial.

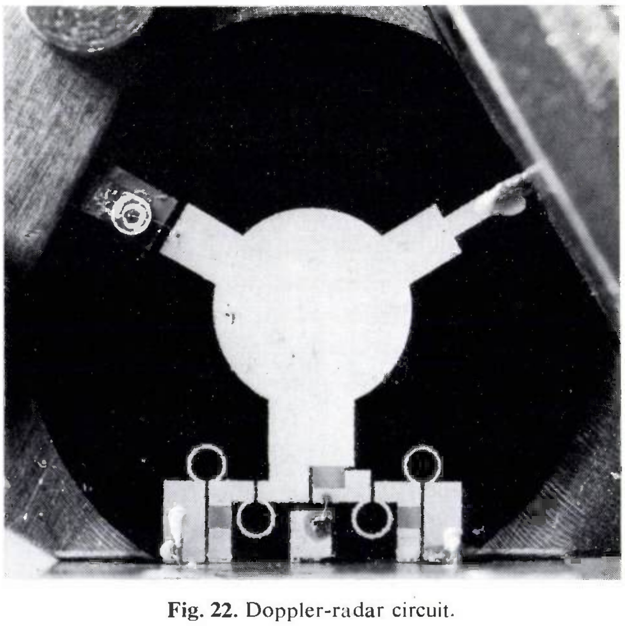


Fig. 22. Doppler-radar circuit.

[6] B. Bollée and G. de Vries, Experiments in the field of parametric amplification, Philips tech. Rev. **21**, 47-51, 1959/60.

[7] C. S. Aitchison, Low noise parametric amplifiers, Philips tech. Rev. **28**, 204-210, 1967.

[8] K. L. Fuller and A. J. Lambell, Traffic-flow analysis by radar, Philips tech. Rev. **31**, 17-22, 1970.

Schottky-barrier diode *D*, and a circulator *C*. The first two are constructed from lumped elements; the third was made in microstrip, even though lumped-element circulators were available. The decision to proceed in this way resulted from the wide bandwidth available from microstrip circulators and the advantage in the initial research work of using such a wide-band circulator.

The Gunn oscillator at this frequency band has already been described. This is incorporated directly though the varactor tuning facility has been removed and the d.c. filter is placed on a different port of the circulator.

The Schottky-barrier mixer circuit was specially designed for this application and is matched to the circulator by a reactive network consisting of a capacitor and an inductor which is connected in series with the Schottky diode. In fig. 21 the bonding wire connecting the Schottky diode *D* to the circuit serves as an inductor. Inductor and diode form a shunt branch; following this shunt branch is a low-pass filter consisting of a capacitor-inductor T-network.

A small amount of the output power from the Gunn device is required as a local oscillator signal for the mixer; this is supplied through the circulator in the isolation direction. The attenuation thus obtained has been optimized for maximum sensitivity of the mixer which receives about 100 μW of local oscillator power; the transmitted power is typically 10 mW at 10 GHz. The system is sufficiently sensitive to detect the return signal from a walking man at 30 metres. A photograph of the Doppler-radar circuit is shown in *fig. 22.*

The satisfactory performance of this complete subsystem confirms what was concluded from the previous examples. It not only turns out to be feasible to make lumped capacitors, inductors and resistors for frequencies up to 12 GHz, but the technique of combining these lumped circuit elements with semiconductor chips also appears to be a viable method of making active microwave circuits.

Summary. In electrical circuits the use of lumped components, small compared with the wavelength, has traditionally been confined to frequencies below 1 GHz. Nowadays planar techniques are available that enable lumped components to be made for frequencies up to at least 12 GHz. These components, and the circuits built from them, are considerably smaller and cheaper than the microwave components in current use, which are built from hollow waveguide or coaxial line, and smaller even than circuits made in microstrip. Active circuits can be made by mounting microwave semiconductor diodes in the lumped circuits as unencapsulated chips. Individual components and both passive and active circuits have been fabricated and give satisfactory performance; among the examples described are low-pass and band-stop filters, a tunnel-diode amplifier, a tunable Gunn oscillator and a Doppler-radar circuit.

Microwave integrated circuits on a ferrite substrate

M. Lemke and W. Schilz

Introduction

In addition to passive components, like filters and couplers, and active elements such as semiconductor oscillators, many microwave systems use non-reciprocal elements. They are called non-reciprocal because their characteristics depend on the direction in which the signals propagate in them. For example, propagation is strongly attenuated in the one direction but weakly attenuated in the other, so that the non-reciprocal element passes the signals in one direction only. Such devices may be used for one-way couplings between generator and load to prevent the load from reacting upon the generator (isolators), or in radars for separating transmitting and receiving systems connected to a common aerial (circulators). Recently non-reciprocal devices have also come into use as variable phase-shifters for controlling directional aerials without the use of mechanical systems ("phased arrays").

These non-reciprocal microwave elements are based on a magnetic material; in practice a special class of ferrite is used. This ferrite possesses ferrimagnetic properties; it is given a magnetic anisotropy by an external static magnetic field. The anisotropy is used for producing the non-reciprocal effect.

The operation of the ferrite components is controlled by means of the external magnetic field, for both analog and digital applications. In digital use the external field serves only to reverse the remanent magnetization of the ferrite, with the significant advantage that no holding power is needed in the time between reversals. Other interesting microwave applications of ferrites result from using the non-linear effects in these materials, for example in the construction of power limiters.

In introducing microwave integrated circuits into microstrip technology [1] it is also necessary to consider non-reciprocal or non-linear ferrite devices in planar-miniaturized form. The ferrite can be used as the substrate material for the microstrip circuit, the ferrite element then being an integral part of the microwave transmission line. This is a practical possibility since polycrystalline sintered ferrite substrates have outstandingly good electrical and mechanical properties. The magnetic and dielectric losses are low, and the mechanical density of the material is so high that optically polished surfaces can be produced. Both these features contribute towards low propagation losses in a microstrip system.

Ferrite can also be used as a non-ferrimagnetic material. By making appropriate chemical substitution the Curie point (below which the material is ferrimagnetic) can be shifted to low temperatures, so that in the working temperature range the material is a pure dielectric that can be generally applied as a high-quality substrate. In the laboratory it has also been found possible to sinter magnetically active zones into the non-magnetic ferrite substrate. This gives in one operation a "composite substrate" on which a microstrip circuit with reciprocal and non-reciprocal elements can be produced in a single photo-etching process [2].

Ferrimagnetic and non-magnetic substrates are made in our laboratories. Their use underlies the investigations on microstrip components and subsystems described in this article. Before going on to examine a number of typical non-reciprocal devices we shall first take a closer look at wave propagation in a ferrite medium. The explanation of non-reciprocity is to be found in the interaction between the r.f. magnetic field and the electron spins in ferrimagnetic ferrite.

Wave propagation in ferrites

An electromagnetic wave is able to propagate in ferrites of the type considered here because, unlike other magnetic materials, they are insulators in which no eddy currents can occur, and because their magnetic and dielectric losses are low (magnetic loss factor $\tan \delta_\mu \leqslant 10^{-3}$, electric loss factor $\tan \delta_\varepsilon \leqslant 2\times10^{-4}$).

When a ferrite is magnetized to saturation by applying a static magnetic field, the elementary magnetic moments, i.e. the electron spins responsible for the magnetism of the material, are oriented parallel to the magnetic field. If they are deflected from this parallel orientation, for example by an r.f. magnetic field nor-

Ing. M. Lemke and Dr. W. Schilz are with Philips Forschungslaboratorium Hamburg GmbH, Hamburg, Germany.

[1] A general treatment of microstrip circuits is given in: J. H. C. van Heuven and A. G. van Nie, Microwave integrated circuits, this issue, page 292.

[2] P. Holst and M. Lemke, Ferrite substrates for microwave integrated systems, IEEE Trans. MAG-5, 478-480, 1969. W. Tolksdorf and P. Holst, Gemeinsames Sintern von Ferriten mit unterschiedlicher Sättigungsmagnetisierung und Curie-Temperatur für integrierte Mikrowellensysteme, Ber. Dtsch. Keram. Ges. 47, 670-673, 1970.

mal to the static field, the electron spins do not simply assume the instantaneous direction of the resulting field but precess around the direction of the static field (*fig. 1*) [3]. The angular frequency ω of the precession is given uniquely by the local magnetic field H:

$$\omega = 2\pi\gamma H; \gamma = 35.2 \text{ kHz A}^{-1}\text{m } (= 2.80 \text{ MHz Oe}^{-1}).$$

In this way an electromagnetic wave propagating in a magnetized ferrite medium can interact with the electron spins. This is most clearly seen for a circularly polarized wave, which has a rotating magnetic field vector. When this vector rotates in the same direction as the precessing spins the interaction with the ferrite medium is strong. When the vector rotates in the opposite direction the interaction is weak. In each case the wave has a different phase velocity and attenuation. It follows from this that the interaction between the ferrite medium and a microwave signal is dependent on the direction of propagation of the wave; the propagation is non-reciprocal.

In practice non-reciprocal microstrip components are frequently made by exciting a circular r.f. magnetic field in a waveguide arrangement, so as to obtain the maximum coupling to the ferrite. However, it is also possible to produce a non-reciprocal device (e.g. the field-displacement isolator shown in fig. 3*a*) in which the electron spins are excited by a linearly polarized r.f. magnetic field. It becomes clear that a linearly polarized field will also interact with the electron spins if this kind of field is considered as the sum of two circularly polarized components, one right-handed (clockwise) and the other left-handed (anti-clockwise). New magnetic-field components arise because of the precession of the electrons, and these cause the field to lose its original linear polarization. The propagation can then be made non-reciprocal in an asymmetric waveguide structure.

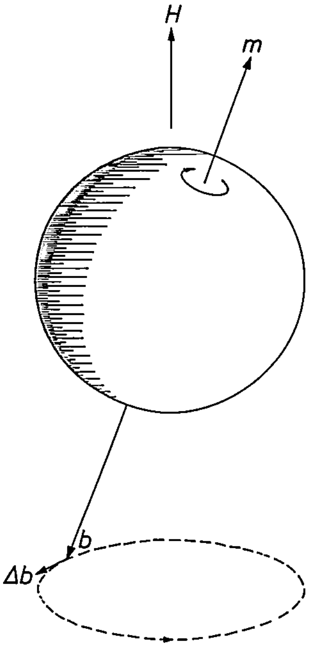


Fig. 1. A spinning electron (angular momentum $\boldsymbol{b}$, magnetic moment $\boldsymbol{m}$) precessing in a magnetic field $\boldsymbol{H}$. The magnetic field attempts to align the magnetic axis of the electron parallel to the field and exerts on the electron a couple that in every infinitesimal time interval adds a small increment $\Delta\boldsymbol{b}$ to the angular momentum $\boldsymbol{b}$. Under the influence of the field the apex of the vector $\boldsymbol{b}$ describes a circle around the direction of the magnetic field; $\Delta\boldsymbol{b}$ is always tangential to this circle. A spinning top precesses in a similar way because of the effect of the gravitational field.

At high r.f. fields there are non-linear effects. These arise because the cone of precession becomes wider, and give extra losses. Such effects are made use of in limiters.

Microstrip components on a ferrimagnetic ferrite substrate

Circulator

Microstrip circulators for frequencies between 3 GHz and 18 GHz have been studied in our laboratories. Purely ferrimagnetic substrates have been used as well as sintered composite substrates, and gave comparable results. *Fig. 2a* shows a typical circulator for 16 GHz. Three ports are coupled to a circular metal disc which acts as a resonator. The disc rests on a ferrimagnetic ferrite substrate magnetized in a direction perpendicular to the surface by a permanent magnet. This non-reciprocal device always allows energy to be transferred from input *1* to *2*, from *2* to *3* and from *3* to *1*, but not in the reverse direction. Thus, if we connect *1* to a transmitter, *2* to an aerial and *3* to a receiver, the transmitted signal will go only to the aerial and the incoming signal only to the receiver. In this configuration the circulator permits the same aerial to be used for transmitting and receiving. The sensitive receiver is protected from the transmitter power by the circulator. A good circulator should therefore combine maximum isolation in one direction with minimum attenuation (insertion loss) in the other. Typical curves of isolation $a_{1\to3}$ and insertion loss $a_{1\to2}$ plotted against frequency are shown in fig. 2*b*. The isolation curve has the characteristic shape of a resonance curve. The bandwidth is primarily determined by the coupling of the circulator ports to the resonator. Given appropriate coupling the reflections at the circulator disc are weak, as appears from the small standing-wave ratio (the voltage standing-wave ratio S is smaller than 1.2 at the centre frequency [4]).

The operation of this type of circulator depends on the dissimilarity in the wavelengths of the waves circulating clockwise and anti-clockwise along the circumference of the circular disc. A standing-wave field is set

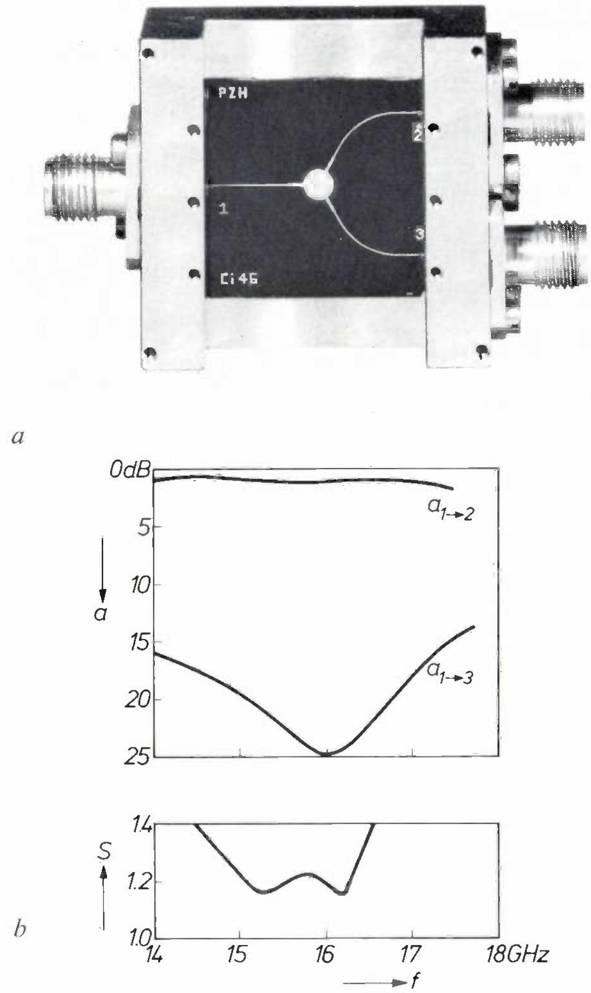


Fig. 2. *a*) A circulator for 16 GHz. Three microstrip lines are coupled to a resonator in the form of a circular disc. The circuit is mounted on a ferrimagnetic ferrite substrate, which is magnetized in a direction perpendicular to the surface by a permanent magnet. Energy transfer is only possible from input *1* to *2*, from *2* to *3* and from *3* to *1*. *b*) Insertion loss ($a_{1\to2}$), isolation ($a_{1\to3}$) and voltage standing-wave ratio (*S*) as a function of the frequency *f*.

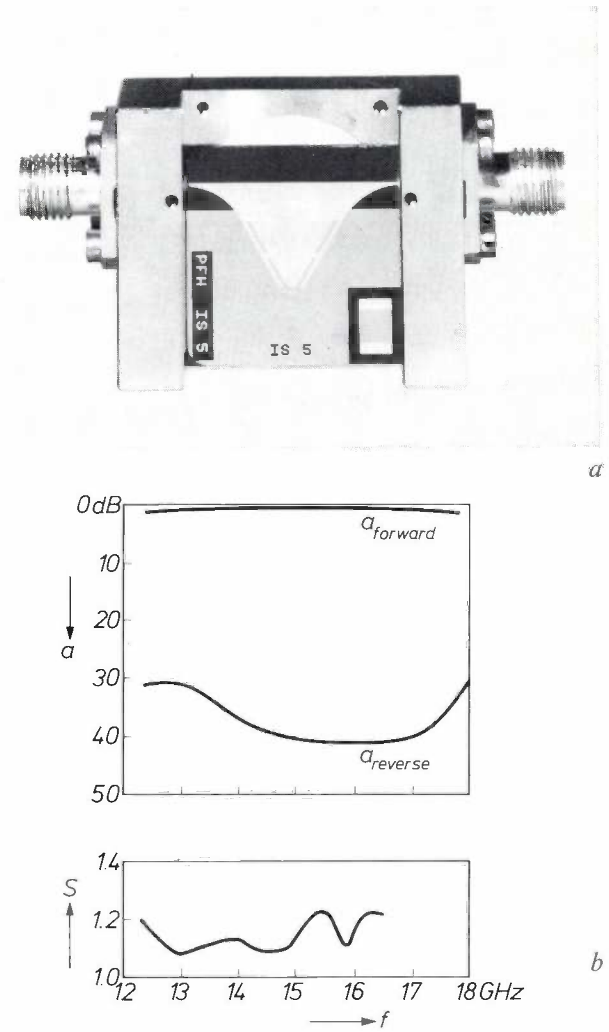


Fig. 3. *a*) Field-displacement isolator. A permanent magnet gives a magnetic field perpendicular to the surface of the substrate. The microwave signal is conducted, depending on the direction of propagation, either along the straight edge of the microstrip to the output, or along the other edge to curve sideways and be absorbed in the vacuum-evaporated nickel-chromium film. *b*) Insertion loss (a_{forward}), isolation (a_{reverse}) and voltage standing-wave ratio (*S*) in the band from 12 to 18 GHz.

up in the circular resonator; the position of the voltage minimum of this standing wave is determined by the magnitude of the applied magnetic field. By adjusting this field the voltage minimum can be made to appear at one of the output arms, which therefore becomes completely decoupled. The total input energy leaves the circulator via the third coupled port.

Isolator

An isolator passes the signal in one direction and absorbs it in the other. The reaction of the load upon the generator can therefore be removed without appreciably attenuating the useful signal. In some applications, an isolator can be produced from a circulator by using port *1* as the input, port *2* as the output and terminating port *3* in a matched load. Planar thin-film matched loads have been developed in microstrip, and these will be discussed below.

Another type of isolator is shown in *fig. 3a*. Here again the static magnetic field is perpendicular to the

[3] A detailed treatment is given in: H. G. Beljers and J. L. Snoek, Gyromagnetic phenomena occurring with ferrites, Philips tech. Rev. **11**, 313-322, 1949/50, and H. G. Beljers, The application of ferroxcube in unidirectional waveguides and its bearing on the principle of reciprocity, Philips tech. Rev. **18**, 158-166, 1956/57.

[4] Whenever signals are reflected at a discontinuity in the transmission path, they interfere with the oncoming signals to form standing waves, and a stationary voltage pattern with maxima and minima at alternate quarter-wavelengths is set up on the transmission line. The ratio of the maximum and the minimum voltage is called the voltage standing-wave ratio (*S*). If the signal is totally reflected the minimum voltage is zero and consequently $S = \infty$; if there is no reflection at all no maxima and minima occur and $S = 1$. In matching a microwave element to a circuit, *S* is usually made as close to unity as possible.

surface of the substrate. On the substrate there is a microstrip line that is considerably broadened out to one side, half-way between the connectors; the added part of the width overlaps an area of vacuum-evaporated nickel chromium. In this field-displacement isolator the microwave signal is conducted, depending on the direction, either along one edge of the stripline to the output, or along the other edge to be absorbed by ohmic losses in the nickel-chromium film [5]. Since an isolator of this type does not include any form of transmission-line resonator it is essentially a broad-band device. Measurements on an isolator designed for 12-18 GHz are given in fig. 3*b*. The isolation was higher than 40 dB for a bandwidth of 2 GHz and more than 30 dB over the whole frequency band at a static magnetic field of 0.52 tesla (5200 gauss). The insertion loss in the transmission direction was always less than 1.5 dB. The good match evident from the small voltage-standing-wave ratio ($S \leqslant 1.23$), in conjunction with the large bandwidth, would make this type of isolator eminently suitable for applications in microwave measurement techniques using a swept frequency, and in other broad-band systems.

Phase-shifters

A typical non-reciprocal microstrip phase-shifter on a ferrite substrate is shown in *fig. 4*. On the substrate there is a microstrip "meander line" (fig. 4*a*) in which the loops are a quarter wavelength long ($\lambda/4$); the distance between them is comparable with or smaller than the substrate thickness. Consequently the lines of force of the r.f. magnetic fields of neighbouring strips intersect each other in the substrate approximately at right angles (fig. 4*b*); at the centre of the strips the phase difference is 90°. This implies that the r.f. field in this region is circularly polarized. With static magnetization normal to the plane of the picture in fig. 4*b* there is thus non-reciprocal interaction with the electron spins. This static magnetization is applied by means of a wire or a few turns through the hole in the centre of the ferrite substrate. The change in the phase velocity of the microwave signal on this structure depends on the magnitude of the applied magnetic field. This type of device can therefore be used as a magnetically tunable phase-shifter.

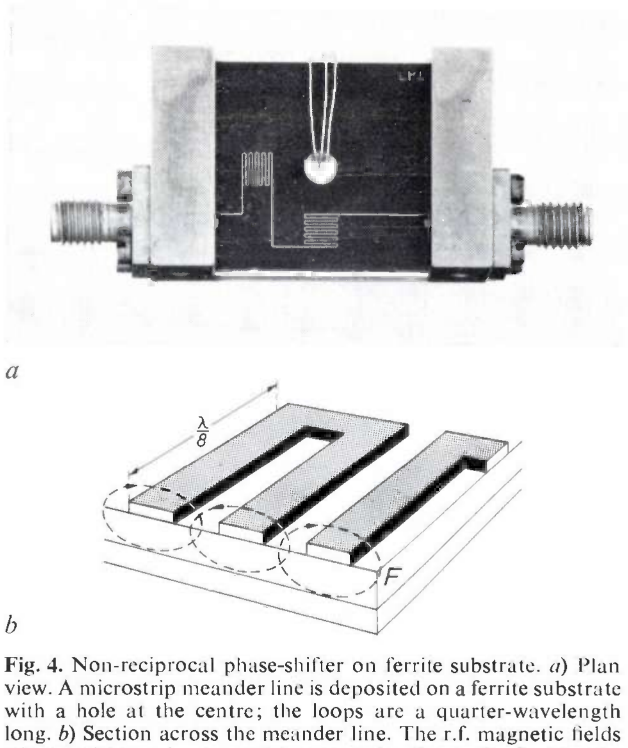


Fig. 4. Non-reciprocal phase-shifter on ferrite substrate. *a*) Plan view. A microstrip meander line is deposited on a ferrite substrate with a hole at the centre; the loops are a quarter-wavelength long. *b*) Section across the meander line. The r.f. magnetic fields of two neighbouring loops intersect each other in the ferrite substrate *F* approximately at right angles and differ by 90° in phase, giving circular polarization, for maximum interaction with the electron spins.

If a square-loop ferrite is used in the system shown in fig. 4 a digital phase-shifter is obtained. Saturation switching of the toroidal ferrite is effected by a wire through the hole in the toroid. Use is made of the difference in phase shift between the two opposite remanence points of the magnetization curve ("latching device"). Groups of such digital phase-shifters can be used in phased arrays for radar [6]. The figure of merit of these devices is the maximum phase shift per dB loss. We have achieved values of about 200°/dB at 9 GHz.

Microstrip components on non-ferrimagnetic ferrite substrates

Passive components

The propagation of microwave signals on microstrip and slotline with non-ferrimagnetic ferrite substrate is similar to that on alumina-based transmission lines, as described in a previous article [1]. We shall give an account of some measurements on ferrite-based microstrip and slotline and compare the results with calculations based on a theory of these lines that has been developed at our laboratories [7]. This theory includes the dispersion in microstrip at higher frequencies; this has been neglected in earlier theories [8]. The higher dispersion of slotline is also calculated. The results are given in *fig. 5a* and *b*, which show the relative wavelength λ/λ_0 as a function of frequency (λ is the wave-

[5] M. E. Hines, A new microstrip isolator and its application to distributed diode amplification, G-MTT 1970 Int. Microwave Symp. Digest, pp. 304-307.

[6] Digital phase-shifters based on a different operating principle and using semiconductor diodes are dealt with by J. H. C. van Heuven, *P-I-N* switching diodes in phase-shifters for electronically scanned aerial arrays, this issue, page 405.

[7] H. J. Schmitt and K. H. Sarges, Wave propagation in microstrip, Nachrichtentechn. Z. 24, 260-264, 1971.
H. J. Schmitt and L. Ericsson, Fundamentals of microwave integrated circuits, Elektronik (Sweden) 2, 53-57, 1971.

[8] H. A. Wheeler, Transmission-line properties of parallel strips separated by a dielectric sheet, IEEE Trans. **MTT-13**, 172-185, 1965.

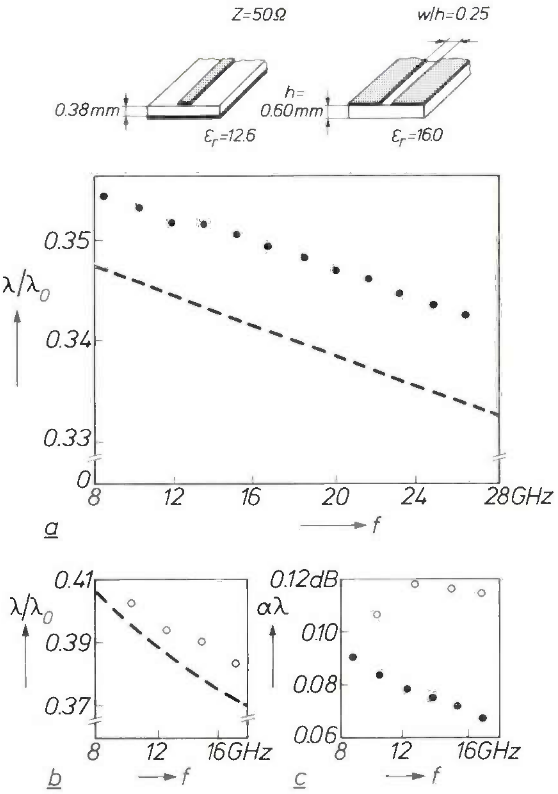


Fig. 5. Measurements and calculations of wavelength and propagation losses for a microstrip line and a slotline made on a non-ferrimagnetic ferrite substrate. The calculations for the microstrip line are based on a theory that takes into account the dispersion ensuing from the deviations from TEM waves. *a*) The wavelength λ on the microstrip line, divided by the wavelength λ_0 in free space, as a function of frequency. H. A. Wheeler's theory [8] assuming TEM waves yields a frequency-independent value of $\lambda/\lambda_0 = 0.359$. *b*) The quantity λ/λ_0 for the slotline. Calculated values are represented by the curve. *c*) Losses over one wavelength $\alpha\lambda$ measured for microstrip line (dots) and slotline (circles).

length on the line and λ_0 the wavelength in free space). The small systematic deviations in absolute value between theory and experimental data are due to an inaccurate determination of the dielectric constant ε_r. Losses for transmission lines on the non-magnetic ferrite substrate are of the same order of magnitude as on the alumina substrate, since the magnetic loss factor tan δ_μ (see page 315) does not apply here (see fig. 5*c*).

The impedance of the microstrip transmission line is essentially determined by the thickness and dielectric constant of the substrate and the width of the conducting strip. In microstrip on a ferrite substrate the characteristic impedance is restricted to the range from about 20 to 100 ohms. Lower impedance values are not acceptable because they would require strip widths which would no longer be small compared with the wavelength, so that the transmission-line characteristics would be lost. The upper impedance limit is set by the increasing losses of extremely narrow lines.

Numerous microstrip components have been developed by combining sections of transmission line. Two band-pass filters and a 50-ohm termination will now be discussed in more detail.

The two filters are shown in *fig. 6*; they are designed

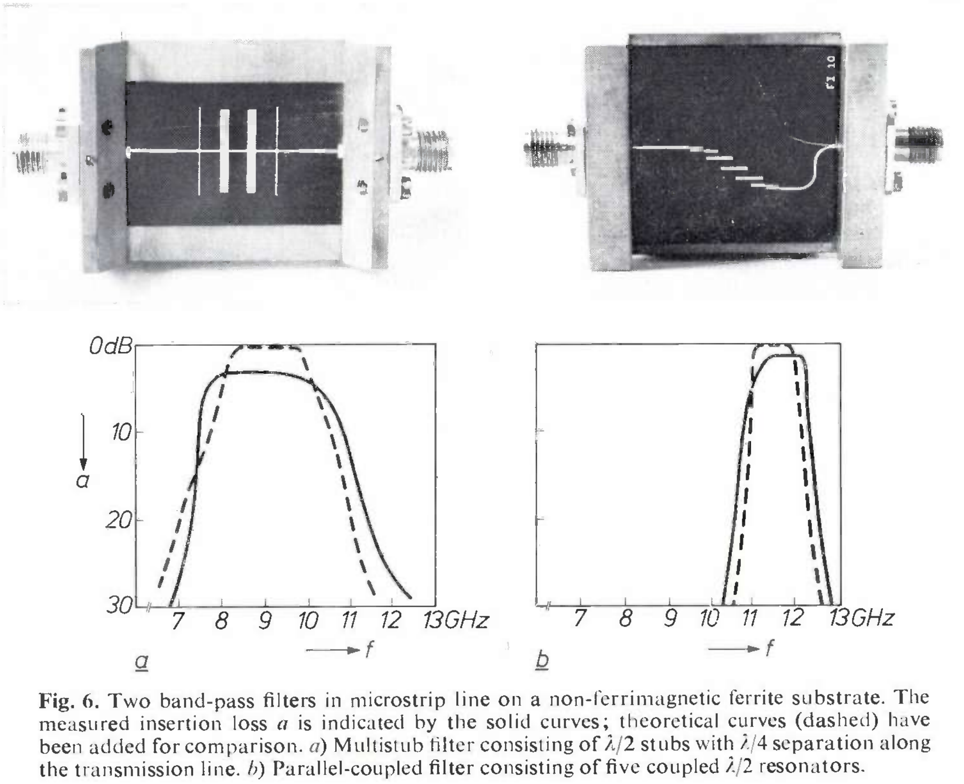


Fig. 6. Two band-pass filters in microstrip line on a non-ferrimagnetic ferrite substrate. The measured insertion loss *a* is indicated by the solid curves; theoretical curves (dashed) have been added for comparison. *a*) Multistub filter consisting of $\lambda/2$ stubs with $\lambda/4$ separation along the transmission line. *b*) Parallel-coupled filter consisting of five coupled $\lambda/2$ resonators.

for maximally flat response at frequencies between 8 and 12 GHz. The filter shown in fig. 6*a* is formed from a series of $\lambda/2$ stubs with $\lambda/4$ separation along the main transmission line. Since high stub admittances are not feasible [9], the minimum bandwidth of this filter is about 10%. A second type of filter is shown in fig. 6*b*. It consists of a number of parallel coupled $\lambda/2$ resonators [10], five of them in the case shown. Here the minimum bandwidth is determined by the microstrip losses, i.e. by the maximum Q-factor of a linear resonator. A Q of about 300 can be obtained, corresponding to a bandwidth of about 3%.

To produce a matched load, there has to be a gradual transition between the loss-free transmission line and an attenuating system. In the spiral load (*fig. 7*) a gradual increase in the absorption is achieved by a gradual transformation of the microstrip mode into even and odd modes propagating between adjacent turns of the conducting spiral. The electric-field vector of the odd mode lies in the surface of the substrate and is attenuated in the thin nickel-chromium film evaporated on to the spiral. The accuracy of the match can be deduced from the voltage standing-wave ratio (S): typical values for 50-ohm terminations are shown in fig. 7.

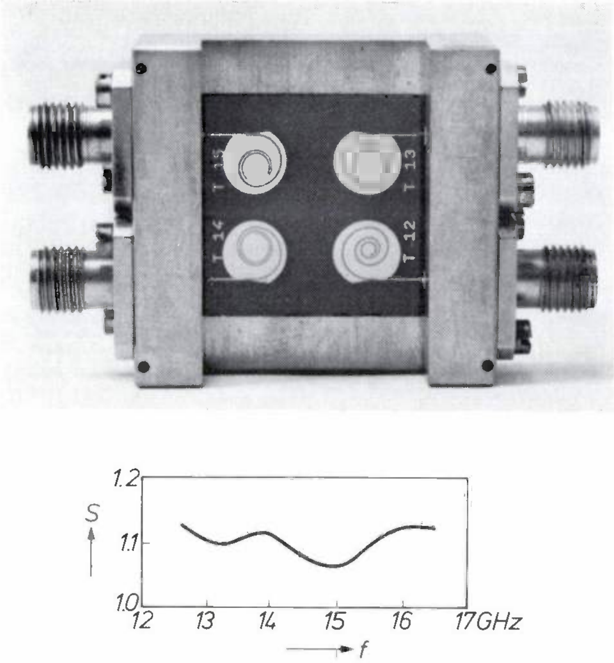


Fig. 7. *Above:* four 50 Ω matched loads consisting of a spiral microstrip line upon which a thin nickel-chromium film has been evaporated. The electric field between the adjacent turns of the spiral has a component in the plane of the resistive nickel-chromium layer that gives ohmic losses. *Below:* accuracy of the match is expressed by the low standing-wave ratio S.

Subsystems

As examples of microstrip subsystems made on a non-magnetic ferrite substrate we shall now discuss a microstrip Gunn oscillator [11] and a balanced mixer using beam-lead diodes. Conventionally a Gunn oscillator for a fixed frequency is used with a microstrip resonator $\lambda/2$ long. An improved version of a microstrip oscillator for 9.2 GHz is shown in *fig. 8* and experimental results are shown in *fig. 9*. This type of oscillator has two circular resonators of the same resonant frequency. The resonant frequency of a circular resonator is determined by the diameter of the metal disc and the dielectric constant of the substrate. The diameter of the disc on the right in fig. 8, which contains the Gunn element, is slightly decreased to compensate for the reactance of the Gunn device. The Q of this resonator is relatively low and the frequency stability of the system is determined by the left-hand resonator ($Q \approx 400$) coupled directly to the first disc. The oscillator is more stable and has a narrower linewidth than an oscillator with $\lambda/2$ resonator (fig. 9).

A microstrip balanced mixer designed for a centre frequency of 11.5 GHz is shown in *fig. 10*. This mixer was developed for future television reception in the 12 GHz band. The local-oscillator and signal frequencies are combined in the hybrid ring (which is of a different design from the one shown in a previous article [1] but operates on similar principles) and fed to the two Schottky-barrier diodes D [12]. These non-encapsulated diodes (beam-lead diodes) are incorporated in the mixer structure by thermocompression bonding. As can be seen from fig. 10, the distance between the diodes and the hybrid is different, the difference in path-length is $\lambda/4$. This introduces an additional 180° phase shift for the signal reflected at one of the diodes. Because of this phase shift the signals reflected at the diodes add up in such a way that they return entirely to the signal input, and the local-oscillator signals reflected at the diodes return to the local-oscillator input.

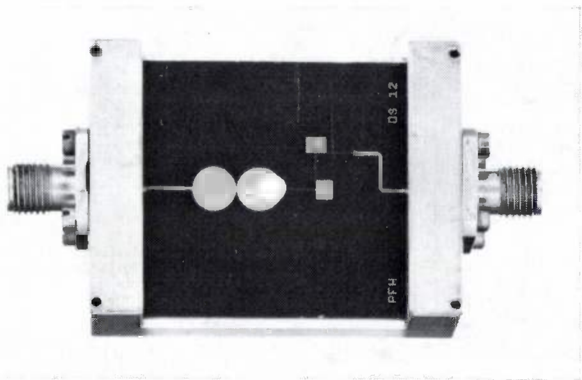


Fig. 8. Gunn-oscillator circuit for 9.2 GHz with two circular resonators. The Gunn device is mounted on the right-hand resonator, and receives its d.c. supply through a low-pass filter. The left-hand resonator is coupled to the output.

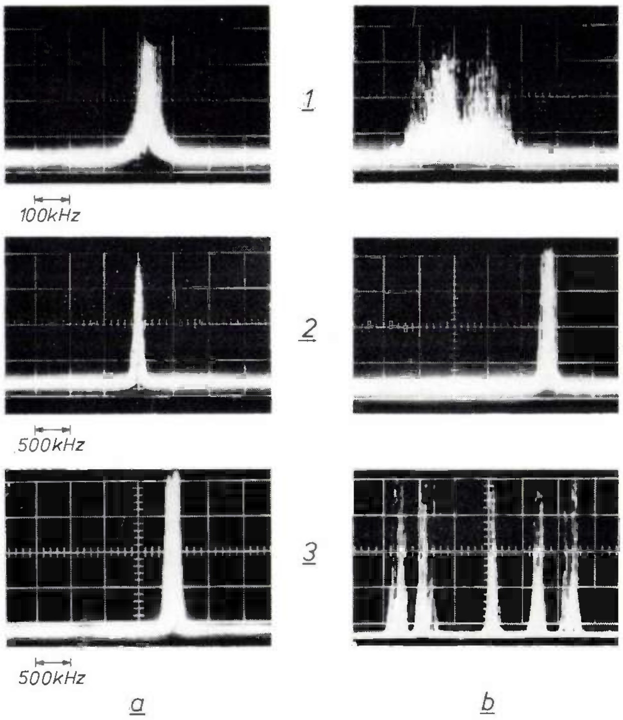


Fig. 9. *a*) Frequency spectrograms of signal generated by the Gunn oscillator of fig. 8, showing the frequency stability. *1*: 100 kHz per division; *2* and *3*: 500 kHz per division. *3* shows five superimposed exposures, taken during a period of one minute. *b*) Frequency spectrograms of signal generated by a Gunn-oscillator circuit with a $\lambda/2$ resonator. Here the linewidth is much greater than for the spectrograms in (*a*). The five exposures made in one minute shown in *3* do not coincide as in (*a 3*).

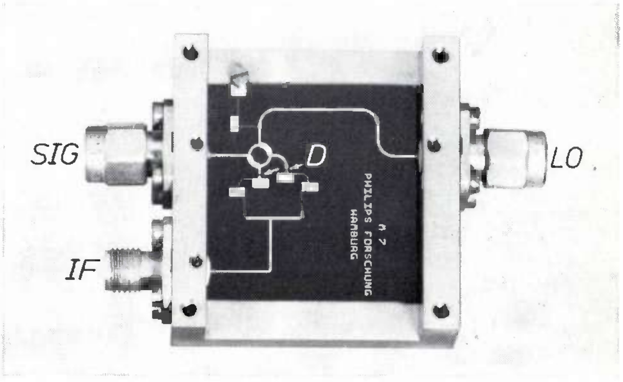


Fig. 10. Balanced mixer on a non-magnetic ferrite substrate for 12 GHz television reception. *SIG* antenna input. *LO* local oscillator input. *IF* intermediate-frequency output. *D* Schottky-barrier diodes.

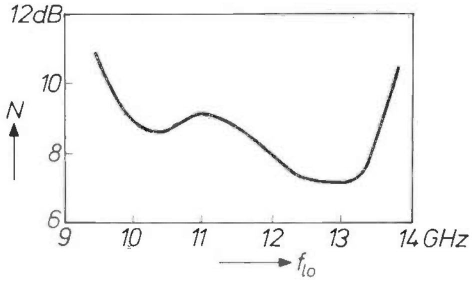


Fig. 11. Single-sideband noise figure of the balanced mixer of fig. 10 at various local-oscillator frequencies $f_{\rm lo}$. This noise figure includes a contribution from a 30 MHz i.f. amplifier with a noise figure of 1.5 dB.

This permits both these inputs to be matched externally to the mixer circuit and provides more than 20 dB of isolation between them for a 10% bandwidth.

The measured noise figure is shown in *fig. 11*. It does not seem possible to achieve much lower values than those shown here. These values would be too high for the direct reception of satellite-based television transmitters, and additional pre-amplification would be required. The noise figure is however low enough for application of the mixer as the first stage of a television receiver for terrestrial transmitters on 12 GHz.

These two examples show that microwave integrated circuits of high quality can be made on a substrate of non-ferrimagnetic ferrite. Ferrite thus appears to be an all-round substrate material: it is indispensable for non-reciprocal elements, and in the non-magnetic form it is of great value for general application.

Summary. In many microwave systems non-reciprocal elements are used (circulators and isolators). These non-reciprocal devices can be produced in the form of integrated circuits using microstrip on a substrate of ferrimagnetic ferrite. When an external magnetic field is applied the wave propagation in the ferrite becomes non-reciprocal, owing to the interaction of the r.f. magnetic field with the precession of the electron spins around the external magnetic field. The article describes a circulator, a broad-band isolator and a meander-line phase-shifter. By chemical substitution the Curie temperature of the ferrite can be reduced to below the working temperature. The ferrite is then a very suitable general substrate material for microstrip circuits with a high dielectric constant ($\approx$ 12) and low loss factor ($\leqslant 2\times10^{-4}$). Some filters, a 50-ohm matched load, a Gunn oscillator and a low-noise mixer are described. If the circuit contains a non-reciprocal element a ferrimagnetic zone can be sintered in, producing a composite substrate.

[9] W. W. Mumford, Tables of stub admittances for maximally flat filters using shorted quarter-wave stubs, IEEE Trans. **MTT-13**, 695-696, 1965.

[10] See for example: S. B. Cohn, Parallel-coupled transmission-line-resonator filters, IRE Trans. **MTT-6**, 223-231, 1958.

[11] See for example: J. Magarshack, Gunn-effect oscillators and amplifiers, this issue, page 397.

[12] The silicon Schottky-barrier diodes were supplied by Dr. D. de Nobel of Philips Research Laboratories, Eindhoven.

YIG filters

P. Röschmann

Introduction

A new group of microwave ferrite components have attracted a good deal of attention in recent years because of their unique and very nearly ideal characteristics. These components, known as "YIG" devices, are now employed in various applications. Small polished samples of single-crystal yttrium iron garnet (YIG, $Y_3Fe_5O_{12}$), operated at the ferrimagnetic resonance, are used as resonators for tunable filters [1], for tuning oscillators and for low-level limiters. Single-crystal YIG has an extremely small ferrimagnetic resonance linewidth, which gives high unloaded Q-factors (up to 10 000).

Fig. 1 shows the principle of operation of a YIG resonator in a typical coupling configuration giving a band-pass-filter response. The electron spins in the YIG crystal that are unpaired and are the origin of its ferrimagnetism are aligned by a static or quasistatic magnetic field $\boldsymbol{H}_0$. An r.f. signal at the input coupling loop builds up an r.f. magnetic field $\boldsymbol{H}_{\text{rf}}$ perpendicular to the static field. Because of the gyroscopic property of the electron spins the resultant magnetization $\boldsymbol{M}$ in the YIG crystal will precess around $\boldsymbol{H}_0$; the precession increases in amplitude if the signal frequency coincides with the precession frequency, which is also known as the ferrimagnetic resonance frequency, and is determined solely by the fundamental constants of the electron and by the applied magnetic field H_0 [2]. As a result of the precession an r.f. magnetic-field component arises perpendicular to the plane of $\boldsymbol{H}_0$ and $\boldsymbol{H}_{\text{rf}}$ (fig. 1); this perpendicular component can be coupled out by means of the second half-loop. It follows that only signals at the precession frequency are coupled from the input loop to the output loop by the precessing magnetization in the YIG sphere. Signals at other frequencies are unaffected by the YIG sphere and no r.f. power is transferred by the YIG filter because there is no coupling between the two orthogonal loops. Thus a YIG filter will give a single unambiguous response, which can be tuned over a range of more than ten to one in frequency simply by varying the applied static magnetic field. The electrical behaviour of YIG filters is in many aspects the same as that of conventional microwave filters with transmission-line resonators or cavities, but YIG resonators do have some unique features (including certain limitations) which will be described in the following section. Several completely assembled YIG devices are shown in *fig. 2*. Their size is mainly determined by the electromagnet used for tuning.

Resonance; lower cut-off frequency

In applications of single-crystal YIG samples as ferrimagnetic resonators for microwave filters the uniform-precession resonance mode (UPR mode) is utilized. This is the fundamental resonance mode in which all electron spins in the YIG sample precess with the same amplitude and phase angle about the static magnetic field. Higher-order resonance modes also exist in which the amplitude and phase of the precession are not the same all over the YIG sample (usually a sphere) and vary in a regular geometrical pattern. The

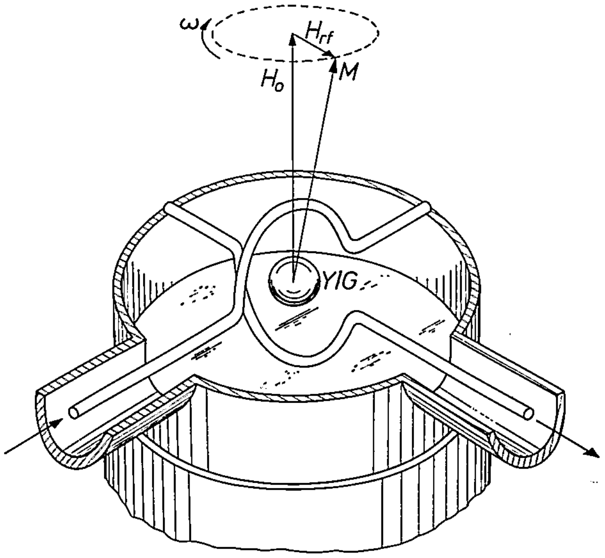
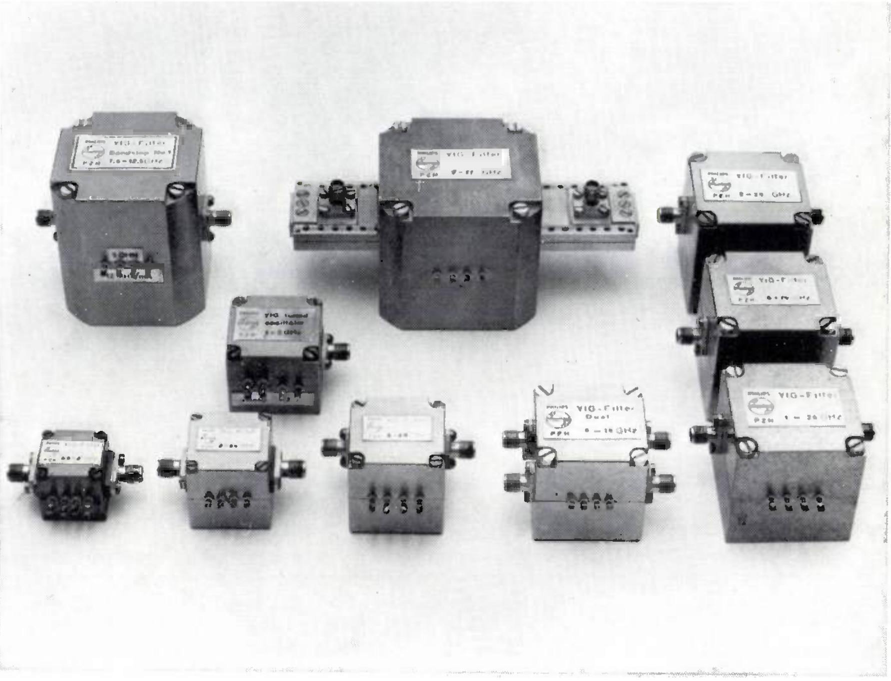


Fig. 1. Band-pass filter with YIG resonator. Input and output are coupled to the YIG sphere *YIG* by two orthogonal half-loops. The sphere is magnetized by a static external field $\boldsymbol{H}_0$; if an r.f. magnetic field $\boldsymbol{H}_{\text{rf}}$ is coupled in, the resulting magnetization $\boldsymbol{M}$ precesses about the direction of $\boldsymbol{H}_0$ because of the gyroscopic property of the electron spins in the material. The precession introduces magnetic-field components that can be coupled out by the output semiloop. The precession angular frequency ω increases linearly with the magnetic field strength H_0, which is adjusted to tune the filter to the frequency required. Other frequencies do not excite the precession and are therefore not transmitted by the filter.

Ing. P. Röschmann is with Philips Forschungslaboratorium Hamburg GmbH, Hamburg, Germany.

Fig. 2. Several tunable YIG devices. The electromagnets, whose connections can be seen on the front of the housings, determine the size of the devices.



resonance modes of YIG samples sufficiently small compared with the wavelength for the effects of wave propagation through the sample to be neglected have been calculated by L. R. Walker [3]; these are called magnetostatic modes or Walker modes.

The resonant frequency of the UPR mode for axially magnetized spheroids with circular symmetry about $\boldsymbol{H}_0$ is given by:

$$f_0 = \gamma \; \{H_0 + H_a + (N_t - N_z)M_s\}, \qquad (1)$$

where γ is the gyromagnetic ratio, which is equal to 35.2 kHz/(A/m) (2.8 MHz/Oe). In the relation (1) H_0 is the external static or quasistatic magnetic field and H_a is the crystal-anisotropy field, which may be regarded as an additional external field whose magnitude and sign depend on the crystallographic orientation of the YIG resonator with respect to $\boldsymbol{H}_0$. The quantities N_t and N_z are the demagnetizing factors for the YIG sample in the transverse and axial directions respectively. Both vary with its geometric shape; for a sphere N_t is equal to N_z, which means that the last term in (1) vanishes. Consequently the UPR resonant frequency of a sphere does not depend on the saturation magnetization M_s, which is a constant of the material and has a value of 1.42×10^5 A/m (1780 Oe) for YIG.

In some applications the YIG resonator cannot be made very small compared with the wavelength and wave-propagation effects have to be taken into account. For a sphere in which N_t is equal to N_z, the resonant frequency is given by [4]:

$$f_0 = \gamma \left\{ H_0 + H_a - \frac{4\pi^2}{90} M_s(\varepsilon_r + 5) \left(\frac{d}{\lambda_0}\right)^2 \right\}, \qquad (2)$$

where ε_r is the relative dielectric constant ($\varepsilon_r = 16$ for YIG) d the sphere diameter and λ_0 the free-space wavelength at resonance. Unlike that of the UPR mode, the resonant frequency is not independent of the dimensions of the sample. The propagation term remains smaller than 50 MHz if the ratio of the diameter of the sphere to the wavelength is less than 1 : 30. This ratio is exceeded, for typical YIG resonators with a diameter of 0.3 mm, at frequencies above 30 GHz.

The resonant frequencies f_n of the higher-order magnetostatic modes are located in a frequency band around the resonant frequency f_0 of the UPR mode:

$$f_0 - \gamma N_t M_s < f_n < f_0 + \gamma \, (0.5 - N_t) M_s. \qquad (3)$$

Fortunately only a few of the higher-order modes are excited in a homogeneous r.f. magnetic field, and only weakly. In multistage filters the level of interfering higher-order modes can be suppressed further by using spheres with a slightly different M_s. This will not change the resonant frequency of the UPR mode but the higher-order modes will have slightly different resonant frequencies owing to their dependence on M_s. Thus the main response remains unchanged whereas the interstage coupling of the unwanted higher-order modes is considerably reduced.

It also follows from eqs. (1) and (2) that there are no resonances at higher harmonics of the frequency f_0; this means that a YIG resonator gives only a single response between d.c. and millimetre wavelengths, which can be tuned linearly through a frequency band of more than a decade by an external static or quasistatic magnetic field.

[1] See for example: G. L. Matthaei, L. Young and E. M. T. Jones, Microwave filters, impedance-matching networks, and coupling structures, McGraw-Hill, New York 1964.

[2] See also: M. Lemke and W. Schilz, Microwave integrated circuits on a ferrite substrate; this issue, page 315.

[3] L. R. Walker, Resonant modes of ferromagnetic spheroids, J. appl. Phys. 29, 318-323, 1958.

[4] J. E. Mercereau, Ferromagnetic resonance g factor to order $(kR_0)^2$, J. appl. Phys. 30, 184S-185S, 1959.

Single-crystal YIG has the smallest known ferrimagnetic resonance linewidth (of the order of 1 MHz [5]). Imperfections in the surface and impurities in the crystal give losses because of scattering of r.f. energy into the crystal lattice; a highly polished surface and a high degree of purity and homogeneity of the YIG crystal are therefore required to obtain the high Q_0 (quality factor) which is possible with YIG resonators. Curves showing the measured Q_0 plotted against frequency for YIG and YGaIG (a gallium-substituted YIG) spheres and for a YIG disc are shown in *fig. 3* [6]. The Q_0 decreases rapidly to zero at the lower frequencies. This occurs when the static magnetic field H_0 corresponding to low frequencies becomes so weak that the YIG resonator is no longer magnetically saturated and the internal magnetic field approaches zero, because the spins are then no longer aligned parallel and r.f. energy couples from the UPR mode to the crystal lattice.

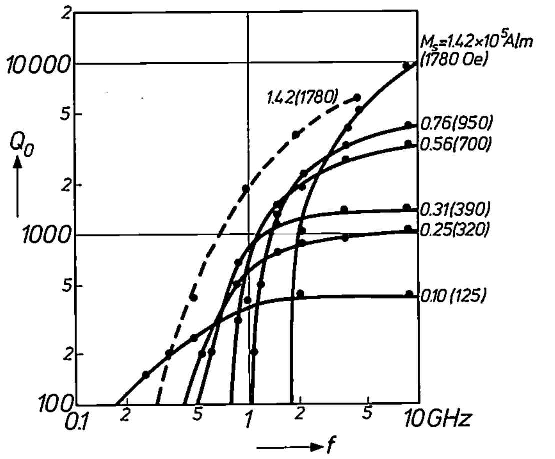


Fig. 3. Unloaded Q_0 (quality factor) plotted against frequency f, measured for YIG and YGaIG spheres (*solid lines*) and a YIG disc (*dashed*) [6].

The cut-off frequency f_c of a YIG resonator that is obtained when the internal magnetic field H_{iz} approaches zero can be found from (1) when we neglect H_a and put $H_{iz} = H_0 - N_zM_s = 0$:

$$f_c = \gamma N_t M_s. \tag{4}$$

For spheres, with $N_t = 1/3$, the cut-off frequency given by (4) is 1660 MHz; measured values are somewhat higher (around 1800 MHz) because not all the electron spins are aligned as soon as H_{iz} becomes greater than zero (this can be seen from the rounded corners of the hysteresis loop).

The cut-off frequency of a ferrimagnetic resonator can be lowered by using thin axially magnetized discs which have an N_t close to zero, or by reducing the saturation magnetization of the resonator material. The saturation magnetization can be reduced either by substituting gallium at iron sites in the YIG system ($Y_3Ga_xFe_{5-x}O_{12}$, with typical x-values between 0.05 and 0.9) or by heating the YIG resonator and making use of the natural decrease of M_s with increasing temperature. All these methods can of course be combined to extend the application of YIG resonators to lower frequencies; owing to poor Q_0 and weak coupling, applications are not feasible below 200 or 100 MHz. The upper limit of the useful frequency range is only determined by the high tuning fields required; a practical limit lies at about 56 GHz, requiring a field of 1.6 MA/m (20 kOe).

Temperature dependence

YIG resonators can be operated from liquid-helium temperatures up to a little below the Curie temperature, which is at about 280 °C. In the usual temperature range for applications, —40 °C to +80 °C, the effect of the temperature on Q_0 and f_0 can be entirely accounted for by the temperature dependence of M_s and H_a. It has been found experimentally that at frequencies well above f_c the quality factor Q_0 is approximately proportional to M_s; the resulting variation in Q_0 with temperature is not large, of the order of ± 10% between —40 °C and +80 °C.

The requirements for the temperature stability of the resonant frequency are quite strict since YIG filters are narrow-bandwidth devices. Optimum performance is obtained with spherical resonators, because N_t is equal to N_z and eqs. (1) or (2) show that H_a is the chief temperature-sensitive parameter. Depending on the crystallographic orientation with respect to the tuning field, the temperature coefficient of the resonant frequency may be adjusted between a positive or a negative value of about 1 MHz/°C, and a very low value (20 kHz/°C) can be obtained with a carefully adjusted sphere.

In other shapes of resonator M_s has a direct effect on the resonant frequency. Thin discs, which are sometimes used because of their low f_c and high Q_0, have a temperature coefficient of the order of 10 MHz/°C.

Non-linearity; power limiting

Depending on the applied r.f. power level YIG resonators have either a linear or a non-linear response. Marked non-linearity appears suddenly above a sharply defined r.f. power level which is different for different materials and for different signal frequencies. This power level is very low (about 10 μW) if an r.f. signal at the UPR-mode frequency f_0 can parametrically excite spin waves at a frequency $f_0/2$. This is possible in a frequency band whose upper limit can be shown to be $2\gamma N_t M_s$ [7]. This happens to be twice the cut-off frequency f_c, so that these spin waves can occur when the resonant frequency of the filter lies between f_c and $2f_c$.

Signals at frequencies above $2f_c$ mainly couple to degenerate spin waves and the threshold for nonlinearity is of the order of 10 mW to 100 mW. The nonlinearity can be utilized for making passive r.f. power limiters, but it also reduces the range of application of linear YIG devices to systems operating at r.f. power levels below 100 mW.

Coupling; practical examples

The r.f. magnetic field, which couples the r.f. signals to a YIG resonator, should be perpendicular to the tuning field to achieve the maximum coupling. The coupling is proportional to M_s and the volume of the YIG resonator and it also depends on the dimensions and impedance of the transmission line. Unlike transmission-line resonators YIG resonators will only give a relatively weak coupling, which means that these resonators will not be heavily damped by the transmission lines coupled to them. As a consequence the bandwidth of a YIG filter remains small; the maximum achievable bandwidth (between the −3 dB points) is of the order of 1 or 2 per cent of the resonant frequency.

A coupling section that is frequently used is the orthogonal-semiloop arrangement shown in fig. 1, which will give a single-stage *band-pass* filter. Multistage filters, which are often required for increased selectivity, are obtained by cascading such coupling sections. A two-stage band-pass filter using this principle is shown in *fig. 4*; the upper half of the tuning magnet has been removed. This particular filter can be continuously tuned from 1 GHz to 20 GHz and has a 3 dB bandwidth varying from 20 to 45 MHz and passband losses between 1.5 dB and 3 dB. *Fig. 5* shows a typical response for a filter of this type tuned to 9 GHz; a spurious response due to coupling of higher-order magnetostatic modes can also be seen; at other frequencies the rejection is more than 50 dB.

Band-stop filters are obtained by inserting YIG resonators into a transmission line at approximately a quarter-wavelength spacing; a symmetrical stripline is usually chosen, so that the tuning magnet can be brought close to the stripline conductor without disturbing the r.f. field (*fig. 6a*). This configuration may be represented by an equivalent circuit with lumped elements. Let us first consider the case in which there is only one YIG sphere. This may be represented by a parallel-resonant circuit coupled to a coupling winding in the transmission line (fig. 6*b*). In the same way as the magnetically coupled parallel circuit at resonance considerably reinforces the magnetic field inside the

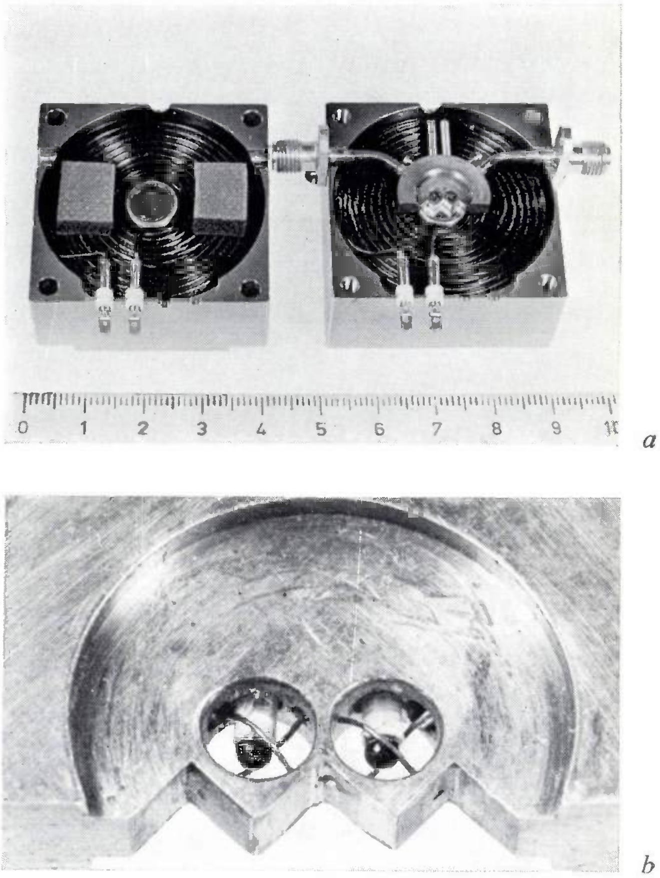


Fig. 4. *a*) Two-stage band-pass filter designed on the principle outlined in fig. 1. The upper half of the electromagnet has been taken off and is shown on the left. This filter can be tuned from 1 GHz to 20 GHz, the 3 dB bandwidth increasing from 20 MHz to 45 MHz. *b*) The two YIG resonators with coupling loops.

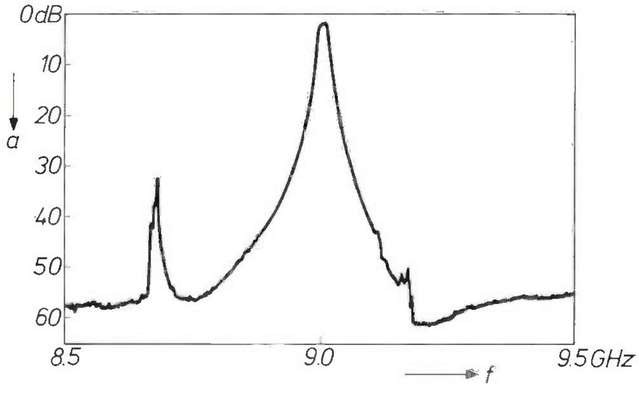


Fig. 5. Attenuation *a* of the filter shown in fig. 4 as a function of frequency. The filter is tuned to 9 GHz. Apart from a spurious response due to a higher-order magnetostatic resonance the rejection is greater than 50 dB outside the passband.

[5] Since the resonance measurements are usually performed at a fixed frequency while subjecting the YIG sample to a varying magnetic field, linewidths are generally given in magnetic units in the literature; 1 MHz corresponds to 28.4 A/m or 0.357 Oe.

[6] The single crystals were grown from a solution of YIG (or YGaIG) in molten lead oxide and lead fluoride, in a process described in W. Tolksdorf, Growth of yttrium iron garnet single crystals, J. Crystal Growth 3/4, 463-466, 1968.

[7] H. Suhl, The nonlinear behavior of ferrites at high microwave signal levels, Proc. IRE 44, 1270-1284, 1956.

coupling winding, the precession field of the resonant YIG sphere locally reinforces the magnetic field of the stripline. The resultant effect is that the line has a high impedance at the location of the resonator.

Owing to the distributed nature of the stripline a second YIG sphere spaced a quarter-wavelength from the first presents a low impedance at the reference plane of the first sphere; this may be represented in the equivalent circuit by a series-resonant circuit connected in parallel. If the coupled parallel-resonant circuit of fig. 6*b* is replaced by an equivalent parallel circuit connected into the line, the equivalent circuit represented in fig. 6*c* is obtained for the two-stage band-stop filter of fig. 6*a*. The corresponding band-stop-filter response is given in fig. 6*d*; this figure also shows the voltage standing-wave ratio S. The tuning range is limited to about one or two octaves because of the $\lambda/4$ separation of the resonators.

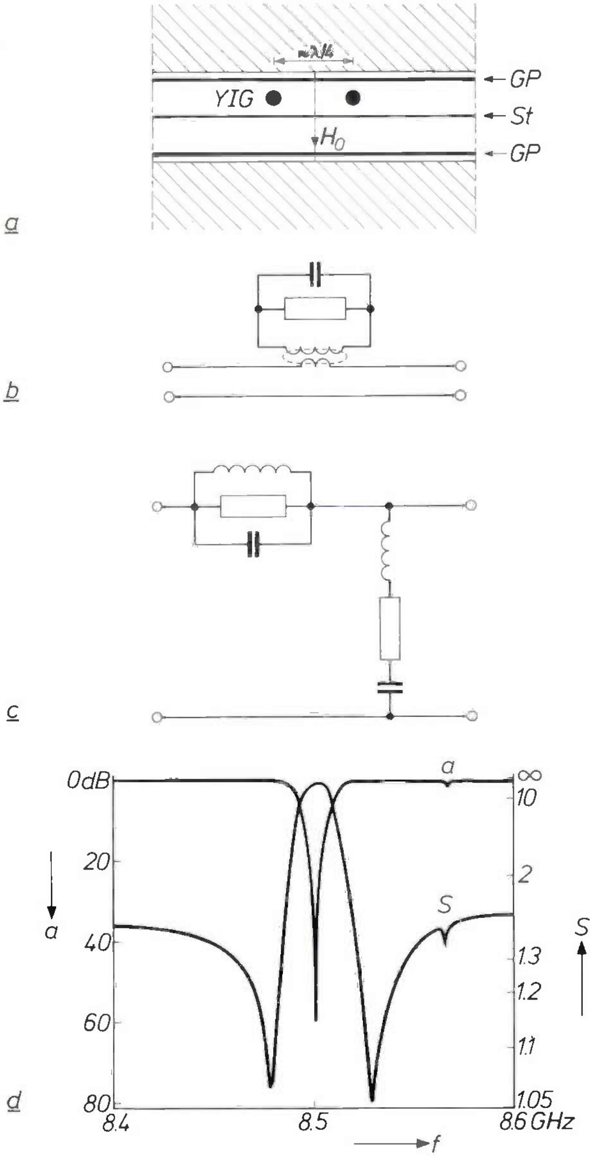


Fig. 6. Two-stage YIG band-stop filter. *a*) Cross-section. Two YIG spheres (*YIG*) are embedded in the dielectric of a symmetrical stripline (*St* strip conductor, *GP* ground planes) at approximately a quarter-wavelength spacing. The whole arrangement is placed between the poles of a magnet giving a static magnetic field H_0. *b*) A single YIG resonator behaves like a parallel-resonant circuit coupled to the line (or appropriately transformed and inserted into the line). *c*) Equivalent circuit of filter containing two resonators at $\lambda/4$ spacing. *d*) Attenuation a and voltage standing-wave ratio S in the frequency band including the resonance.

A *nonreciprocal-attenuation* response is most effectively obtained by coupling the YIG resonator to circularly polarized r.f. magnetic fields. These are present in a waveguide and can be set up in a coaxial or microstrip line by using special techniques. Nonreciprocal YIG filters may be applied as narrow-band tunable isolators, circulators or directional filters. We should note here that a nonreciprocal phase shift is given by the orthogonal semiloops shown in fig. 1, although the r.f. field is linearly polarized here. Since the precession of the magnetization in the YIG resonator has a fixed sense of rotation the direction of wave propagation determines whether the phase shift will be + 90° or −90°.

It has been stated above that the size of YIG devices is mainly given by the required tuning magnet. Since the highest required tuning field is determined by the highest frequency to be tuned, tuning power and magnet size can only be kept small by paying careful attention to the height required for the air gap when designing the r.f. section. Typical air-gap dimensions given by different designs are 1 mm to 2 mm for the orthogonal half-loop, 1.5 mm to 3 mm for symmetrical stripline and 2 mm to 4 mm for waveguides. The diameter of the pole faces should be at least five times the height of the air gap to provide a homogeneous tuning field.

Fixed-tuned YIG filters with permanent magnets are much smaller than filters in waveguide or coaxial line,

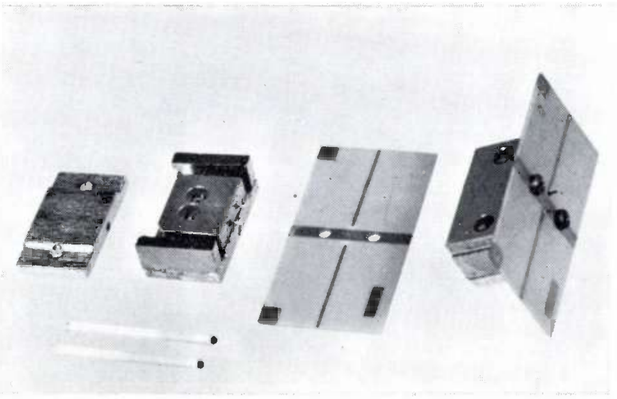


Fig. 7. Fixed-frequency two-stage YIG band-pass filter with permanent magnets, to be inserted into a microstrip line. *From left to right:* upper pole piece, coupling structure for YIG resonators with lower pole piece and two permanent magnets, microstrip substrate and assembled filter. The narrow strip on the substrate is the microstrip line; the filter is connected to it by two contact pins. *Foreground:* two YIG resonators glued to tiny rods for manipulating them and mounting them in position.

particularly for the lower microwave frequencies, where transmission-line resonators become increasingly large. YIG resonators are indeed small enough to be used in microstrip circuits as high-Q resonators. An experimental two-stage YIG band-pass filter with permanent tuning magnet for microstrip application is shown in *fig. 7*; it can be fixed to the substrate of an integrated microstrip circuit by screws. *Fig. 8* shows the performance of this filter. Band-pass or band-stop filters of this kind are feasible for frequencies from abòut 1 GHz to more than 12 GHz.

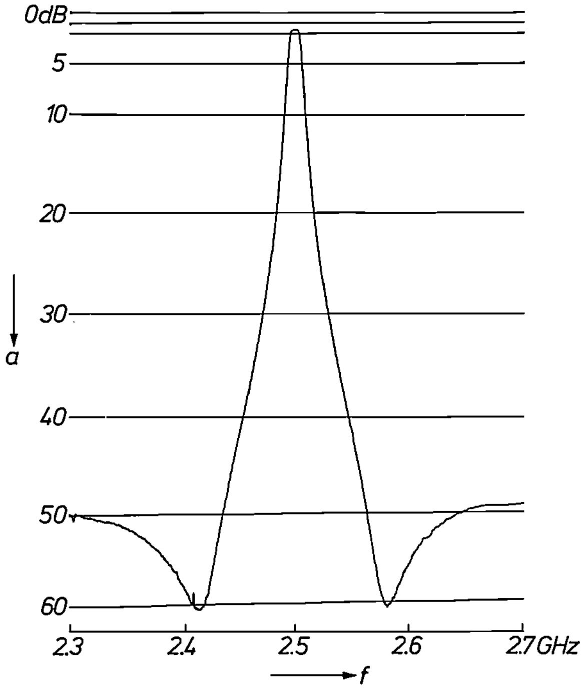


Fig. 8. Measured frequency response of the fixed-frequency YIG band-pass filter shown in fig. 7.

Many other devices can be made with YIG resonators, e.g. fast switchable filters, tunable frequency discriminators, frequency meters, magnetic-field measuring probes, tunable transistor and Gunn oscillators and tunable harmonic generators with varactor multipliers.

Summary. Single crystals of the ferrite material yttrium iron garnet (YIG) give the smallest known ferrimagnetic-resonance linewidth. Since microwave signals are efficiently coupled to the ferrimagnetic resonance, single-crystal YIG samples can be used as magnetically tunable microwave resonators with unloaded Q-factors up to 10 000 and a linear tuning range of more than ten times in frequency. YIG resonators are used in devices such as tunable band-pass or band-stop filters. The resonators can be used for signal frequencies from about 100 MHz up to 60 GHz, depending on material characteristics and shape. Spheres of YIG will give a low temperature coefficient for the resonant frequency (about 20 kHz/°C). The resonance effect is non-linear above an r.f. power level of between 10 and 100 mW (above 10 μW at the lowest frequencies); this is of use in power limiters.

Philips tech. Rev. 32, 328-344, 1971, No. 9/10/11/12

IMPATT diodes

D. de Nobel and M. T. Vlaardingerbroek

Introduction

An oscillator or amplifier circuit always contains an "active" element that delivers power at the operating frequency. The a.c. current through an active device and the voltage across it are more than 90° and less than 270° out of phase: the real part of the impedance (the resistance) is negative. In about 1930 J. Müller [1] and later F. B. Llewellyn and A. E. Bowen [2] attempted to give a thermionic diode a negative resistance by making use of the transit time of the electrons. If the electron transit time is large enough the a.c. component of the current — averaged over the transit space — lags in phase by more than 90° behind the a.c. voltage. Although these attempts were successful, they did not lead to practical applications. The output power and efficiency of such an oscillator are too low, because the current density obtainable in a thermionic diode is very small and the cathode current responds too weakly to an a.c. field. In 1954 W. Shockley drew attention to the possible usefulness of *solid-state devices* with a negative resistance caused by transit-time effects [3]. In 1958 his associate W.T. Read suggested combining the transit-time effect with the avalanche ionization that occurs in the breakdown caused when a high reverse-bias voltage is applied to a *P-N* junction [4]. The first structure proposed was a *P-N-I-N* configuration. The transit-time effect occurs in the virtually intrinsic *I* region, the "drift region". Although the avalanching *P-N* junction fundamentally affects the phase relations, it acts primarily as an electron source, i.e. as a cathode. This implied a considerable advance over the thermionic diode, because apart from the high breakdown-current densities obtainable, the avalanching process reacts very strongly to field variations.

It was not until 1965 that mention was made, in a Bell Laboratories publication [5], of an experimental realization of Read's idea. At the same time it was shown that an ordinary *P-N*-junction diode — without a nearly intrinsic region — is capable of similar operation, i.e. as an "avalanche transit-time diode" [6]. It later appeared that avalanche transit-time diodes had previously been made by A. S. Tager and his associates in Russia [7]. In 1964 Tager made an important contribution to the understanding of this diode with a simple analytical formulation of the large-signal behaviour of the diode. In the literature the diode is referred to variously as an "avalanche diode", "ATT diode" and now more usually as an "IMPATT diode" (ATT for avalanche transit time; IMPATT for impact-ionization avalanche and transit time).

Since 1965 a great deal of work has been done on IMPATT diodes by many research teams in many countries. Among the various solid-state microwave sources now being investigated, the IMPATT diode is potentially the most powerful. It has already proved capable of generating continuous powers of 8 W at 6 GHz and 1 W at 50 GHz.

Fig. 1 gives an example of a microwave oscillator in which the active device is an IMPATT diode. In this article we shall not be concerned with IMPATT diode oscillators as such [8], but will confine ourselves to the diode itself. We shall first describe the operation of the diode, following the theories given by Read and Tager [9]. Some of our experimental results relating to output power, impedance and noise will then be discussed. The article concludes with an examination of some technological aspects of the fabrication of these devices.

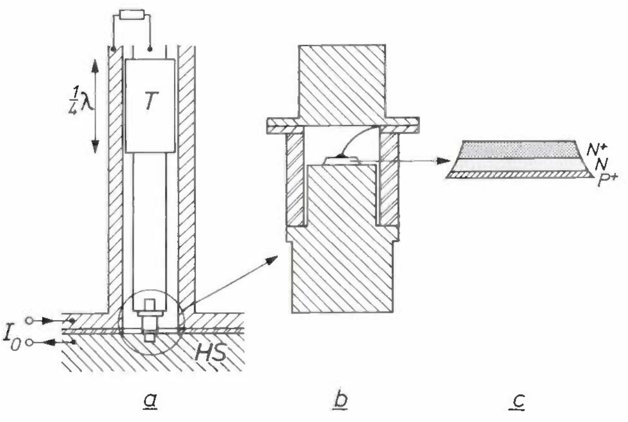


Fig. 1. Coaxial oscillator with IMPATT (impact-ionization avalanche and transit-time) diode, schematic (*a*). The "varactor package" (*b*) contains the IMPATT diode (*c*). The diode is supplied with a d.c. current I_0. The coaxial output lead (the upper part of *a*) is matched to the diode with a quarter-wave transformer *T*. The heat generated in the diode is dissipated through the heat sink *HS*.

Dr. D. de Nobel and Dr. Ir. M. T. Vlaardingerbroek are with Philips Research Laboratories, Eindhoven.

Operation of the IMPATT diode

D.C. behaviour

The a.c. currents and voltages in an IMPATT diode are superimposed on a d.c. situation, which we must first consider.

A reverse-bias voltage is applied to a silicon P-N junction (*fig. 2*). The N region then has a positive voltage with respect to the P region, and we initially take this voltage to be lower than the breakdown voltage. The electrons of the N region and the holes of the P region have moved some way away from the junction, resulting in a depletion layer in which the donors and acceptors are no longer neutralized. There is thus a space charge, which is positive in the N region and negative in the P region (fig. 2*a*). The gradient of the electric field E is proportional to the space-charge density (fig. 2*b*).

In this situation a weak current I_s, called the leakage or saturation current, flows through the diode. The carriers of this current in the depletion layer are holes and electrons that are thermally generated either inside the depletion layer as hole-electron pairs, or just outside the depletion layer — within a diffusion distance from it — as "minority carriers" (electrons in the undepleted part of the P region, or holes in the undepleted part of the N region). In fig. 2 the holes move to the left, the electrons to the right, each with their own "drift velocity".

If we now increase the applied voltage, the depletion layer at first only becomes thicker; the shape of the field profile (fig. 2*b*) remains the same and the current remains low. Finally, however, there is breakdown. Thermally generated charge carriers, in a thin layer where the field is strongest, then take up so much energy from the field that they may form hole-electron pairs by impact ionization. In their turn the newly formed holes and electrons take up energy from the field and generate new pairs, and the process snowballs to produce an avalanche of charge carriers, and hence a large current in the diode. The value of the electric field E_0 at which this process is initiated, called the breakdown field, is about 400 kV/cm in silicon. The differential resistance in an avalanching P-N diode is very low, which implies that the diode must be supplied with power from a current source to obtain well defined operating conditions.

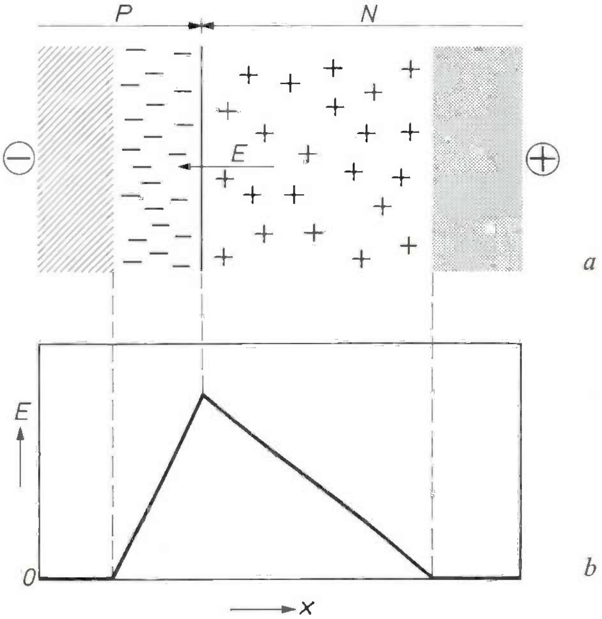


Fig. 2. *a*) In a reverse-biased P-N junction the applied voltage causes the formation of a depletion layer at the boundary between the P and N regions, with unneutralized acceptors (—) in the P region and unneutralized donors (+) in the N region. The electric field is directed to the left; a free electron formed or arriving in the depletion layer therefore travels to the right, and a free hole to the left. *b*) The value of the electric field E as a function of the distance x. The gradient of E is proportional to the space-charge density.

To keep the description straightforward we shall take as our starting point a highly simplified d.c. situation, as characterized by the field profile of *fig. 3*. The

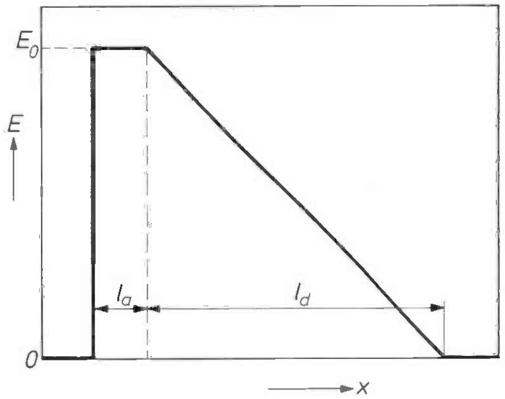


Fig. 3. The field profile in a P^+-N junction at breakdown (schematic). Impact ionization gives rise to avalanching in a region of length l_a, the "avalanche region"; the field here is equal to E_0, the breakdown field-strength. The electrons generated in this region travel at a constant velocity u_s through the "drift region", of length l_d, to the undepleted part of the N region. Transit-time effects in the avalanche region will be neglected in this article. The drift region for the holes is on the left of the avalanche region; it is assumed to be so thin that it can also be neglected.

[1] J. Müller, Hochfrequenztechnik und Elektroakustik **43**, 195, 1934.

[2] F. B. Llewellyn and A. E. Bowen, Bell Syst. tech. J. **18**, 280, 1939.

[3] W. Shockley, Bell Syst. tech. J. **33**, 799, 1954.

[4] W. T. Read, Jr., Bell Syst. tech. J. **37**, 401, 1958.

[5] C. A. Lee, R. L. Batdorf, W. Wiegmann and G. Kaminsky, Appl. Phys. Letters **6**, 89, 1965.

[6] R. L. Johnston, B. C. De Loach, Jr., and B. G. Cohen, Bell Syst. tech. J. **44**, 369, 1965.

[7] A. S. Tager, Soviet Phys. Uspekhi **9**, 892, 1967. V. M. Val'd-Perlov, A. V. Krasilov and A. S. Tager, Radio Engng. Electronic Phys. **11**, 1764, 1966.

[8] These are the subject of the article by K. Mouthaan, IMPATT-diode oscillators, on page 345 of this issue.

[9] These ideas have been expressed in a particularly clear and simplified form by D. Delagebeaudeuf, Onde électr. **48**, 722, 1968. For a detailed treatment of the same subject, with special emphasis on the occurrence of higher harmonics, see: K. Mouthaan, Philips Res. Repts. **25**, 33, 1970.

depletion layer here is divided into an "avalanche region" and a "drift region" for electrons. Impact ionization and avalanching only occurs in the avalanching region, where $E = E_0$. The electrons generated here flow through the drift region towards the undepleted part of the N region. Their transit time is such that transit-time effects are perceptible at microwave frequencies. Transit-time effects in the avalanche region are assumed to be negligible.

It is further assumed in fig. 3 that the drift region for holes, which is on the left of the avalanche region, is so thin that it can be neglected. An approximation to this situation is found in a "P^+-N"junction, where there is an abrupt transition from a very highly doped P region (i.e. one in which the field gradient is very steep) to a much less highly doped N region. Similar field profiles can be obtained in an N^+-P junction or in a Schottky barrier (the potential barrier at a semiconductor/metal interface).

Finally we assume that the electrons move through the entire drift region at a *constant* drift velocity u_s. This is a good approximation in avalanching silicon diodes. Experiments have shown that the drift velocity u of electrons in silicon reaches a saturation value u_s of about 10^7 cm/s at a field value of about 20 kV/cm (*fig. 4*). This field-strength is exceeded in practically the entire depletion region (see fig. 3, $E_0 \approx 400$ kV/cm), and therefore the drift velocity may be assumed to have the value u_s throughout the depletion region. It is not surprising that u does not continue to increase linearly with E at higher fields, since more scattering mechanisms then come into play.

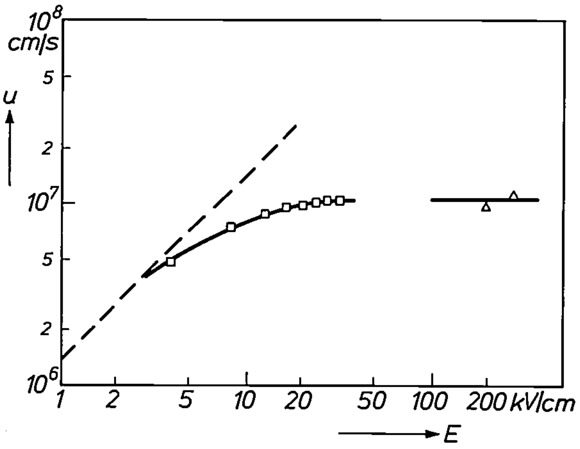


Fig. 4. Drift velocity u of electrons in silicon as a function of the field-strength E. (The squares are after C. B. Norris and J. F. Gibbons, the triangles after C. Y. Duh and J. L. Moll [10].) The dashed line is an extrapolation of the drift velocity at low fields. Nearly everywhere in the depletion layer of an avalanching IMPATT diode the value of E is greater than 20 kV/cm, and u is therefore approximately at its saturation value (about 10^7 cm/s).

The thickness l_a of the avalanche layer is directly related to the value of the ionization coefficient α of the holes and electrons in the avalanche layer. This coefficient is defined as the average number of pairs of charge carriers generated by an electron or hole per unit distance travelled in the direction of the current. The coefficient is zero for $E \ll E_0$, but in the vicinity of E_0 it rises steeply with E [11].

Assuming for simplicity that α is the same for holes and electrons and has a constant value throughout the avalanche layer, and assuming further that the number of thermally generated charge carriers is negligible, then the following simple relation must exist in the steady state:

$$\alpha l_a = 1. \qquad (1)$$

This is because the hole and electron of a given pair formed by impact ionization have between the two of them covered a distance l_a before leaving the avalanche region, the hole going to the left, the electron to the right. On the way they have thus created αl_a new pairs. If the avalanche is to be self-sustaining, the disappearing pair must be compensated by the creation of one new pair, which is expressed in equation (1).

A.C. behaviour

Starting from the d.c. situation of fig. 3 we now show in *fig. 5* what happens in the diode at microwave frequencies.

The diode is supplied from a d.c. source that keeps the time average of the current through the diode at the value I_0. In the drift region the field gradient depends not only on the charge of the donors but also on the charge of the electrons carrying the current. The slope of the d.c. profile (the dashed line in fig. 5) decreases for increasing I_0.

Let us suppose that for some reason or other there is a surplus of electrons in the middle of the drift region (fig. 5*a*). The field gradient here is then smaller, which has the result that the field in the avalanche region is lower than E_0. (This assumes that the thickness of the drift and avalanche regions remain unchanged.) The production of charge carriers in the avalanche region thus falls below par, resulting in a deficit of electrons in the current of electrons moving to the right, and therefore, compared with the steady state, a positive charge moving to the right (fig. 5*b*; note that the positive charges indicated are not holes in the usual sense of the word). Every surplus and deficit of electrons moves to the right, with the entire electron current, at the velocity u_s. In fig. 5*c* the electron deficit has arrived in the middle of the drift region, and the process is repeated in the reverse direction. What this amounts to is that charge variations in the drift region cause field variations in the avalanche region, and *vice versa*, and this

[10] C. B. Norris, Jr., and J. F. Gibbons, IEEE Trans. **ED-14**, 38, 1967.
C. Y. Duh and J. L. Moll, Solid-State Electronics **11**, 917, 1968.

[11] See for example: S. M. Sze and G. Gibbons, Appl. Phys. Letters **8**, 111, 1966.

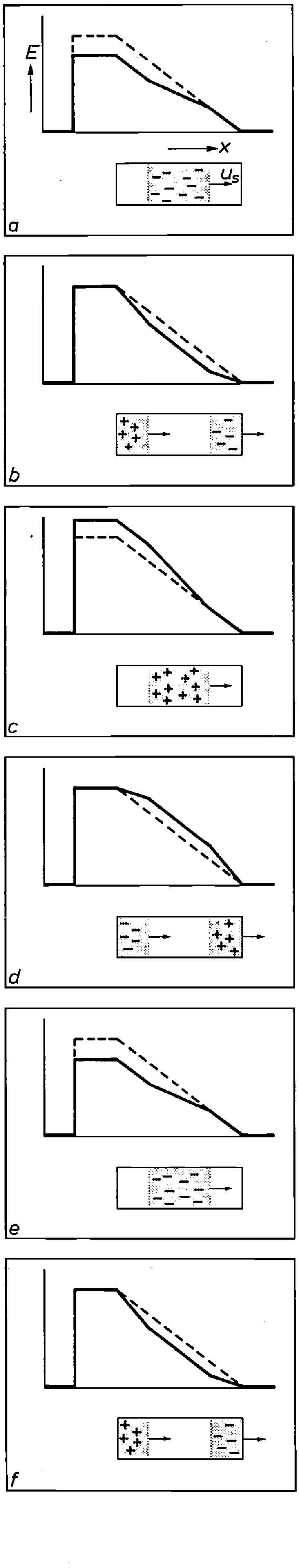


Fig. 5. Field and charge variations in an IMPATT diode at microwave frequencies. Each figure gives the field profile, and below it the additional charge in the drift region (compared with the steady state). The dashed line is the steady-state field profile. Each pair of successive figures corresponds to a time difference of a quarter of a period. *a*) Owing to an excess of electrons in the drift region (compared with the steady state) the field in the avalanche region is too low. Consequently too few electrons are produced, and a deficit of electrons enters the drift region (*b*). As a result of this the field in the avalanche region is too high (*c*), and so on. Between stages *b* and *d* a transit time τ has elapsed, between *b* and *f* a period *T*.

interaction is such that periodic variations in field and charge can reinforce one another.

There are two conclusions to be drawn from this very simplified picture. In the first place the extra negative charge must have largely disappeared from the drift region when the extra positive charge arrives (fig. 5*b*): the charge variations in the drift region must show the appropriate differences of phase. This means that the transit time τ, and hence the length of the drift region l_d must be matched to the angular frequency ω at which the diode is required to operate. In fig. 5 a transit time τ has elapsed between *b* and *d*, and a period *T* between *b* and *f*. It would therefore follow from fig. 5 that $T = 2\tau$, and therefore $\omega\tau = 2\pi\tau/T = \pi$. As we shall see later, the best value for $\omega\tau$ is in fact 0.74π.

In the second place the charge variations must not be immediately corrected by the avalanche process; the correction mechanism must lag in phase sufficiently to allow the charge variations to form and escape. For this to be possible the "avalanche frequency" ω_a, a frequency which will be defined later, must be lower than ω.

Though the above model gives some idea of how the diode works, we have not yet proved that the diode can in fact deliver microwave power. To do this we must analyse the relations between the a.c. voltages and currents in the avalanche region and in the drift region (see *fig. 6*). The diode is usually part of a resonant microwave circuit tuned to a particular angular frequency ω. The higher harmonics of the voltages and currents — which are bound to arise because of the highly non-linear nature of the avalanche process — will then be suppressed by the circuit, and therefore in the following we shall only consider the components with angular frequency ω. In the next article it will be shown that higher harmonics can be put to some use by tuning the circuit to a harmonic as well as to the fundamental.

In fig. 6 the electrical capacitance of the avalanche and drift regions is taken into account: the capacitance of the avalanche region is C_a and that of the drift region C_d. We must now distinguish between *conduction currents*, which are carried by holes and electrons, and *capacitive currents*, which represent field variations. In the avalanche region we thus have a conduction current i_{ca} and a capacitive current $i_{cap\ a}$. The total current measured in the external circuit is:

$$i_t = i_{ca} + i_{cap\ a}. \qquad (2)$$

The contribution i_{ca} to the current in the external circuit is a direct continuation of the charge current in the avalanche region, and the remainder, $i_{cap\ a}$, is used for periodically charging and discharging the avalanche region, regarded as a capacitor. Owing to the transit-time effect the situation in the drift region is somewhat more complicated. In this region we cannot speak of *the* conduction current i_{cd}, because at any given instant it is different for different cross-sections (see fig. 5). What is observed of the charge movement in the drift region at any given instant in the external circuit is the conduction current *averaged over the drift region*, $\overline{i_{\mathrm{cd}}}$. In other words, the charge movement in the drift region induces a current $\overline{i_{\mathrm{cd}}}$ in the external circuit: this current is therefore referred to as the *induction current* i_{i}. The sum of this induction current and the current $i_{\mathrm{cap\ d}}$, which charges and discharges the "drift-region capacitor", is again equal to the total current:

$$\begin{aligned} i_{\mathrm{t}} &= i_{\mathrm{i}} + i_{\mathrm{cap\ d}}, \\ i_{\mathrm{i}} &= \overline{i_{\mathrm{cd}}}. \end{aligned} \qquad (3)$$

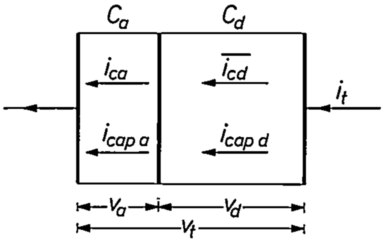


Fig. 6. A.C. voltages and currents in an IMPATT diode. In the avalanche region and the drift region, i_{ca} and i_{cd} are the respective conduction currents, and $i_{\mathrm{cap\ a}}$ and $i_{\mathrm{cap\ d}}$ the capacitive currents. For the conduction current in the drift region, only the average over this region, $\overline{i_{\mathrm{cd}}}$, is important. The total current is i_{t}. The a.c. voltages across the avalanche region, the drift region and the whole diode are v_{a}, v_{d} and v_{t}.

When we speak of capacitive currents *inside* the dielectric, we are really translating every change of field into a displacement current (i_{D}). By definition the displacement-current *density* is simply $\dot{D}$, the time-derivative of the dielectric displacement. Now the field change at a given instant in the drift region, and hence $\dot{D}$, differ at different cross-sections (see fig. 5). The average of i_{D} over the drift region (or, more generally, over an arbitrary piece of dielectric) is in fact equal to the capacitive current because if x is the coordinate along the length of the dielectric, l the length and S the cross-sectional area, the average of i_{D} is given by:

$$\overline{i_{\mathrm{D}}} = \frac{1}{l}\int_0^l i_{\mathrm{D}}\mathrm{d}x = \frac{1}{l}\int_0^l \dot{D}S\mathrm{d}x = \frac{\varepsilon S}{l}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^l E\mathrm{d}x = C\mathrm{d}V/\mathrm{d}t = i_{\mathrm{cap}}.$$

Here $\varepsilon S/l$ is equal to the capacitance C of the piece of dielectric, and $\int_0^l E\mathrm{d}x$ is equal to the voltage V across it.

If we define the total current i_{t} as the sum of the conduction and the displacement current in a cross-section at the location x:

$$i_{\mathrm{t}} = i_{\mathrm{c}} + i_{\mathrm{D}}, \qquad (4)$$

then the divergence of the total current density is equal to zero. This follows immediately from Maxwell's first equation: curl $\boldsymbol{H} = \boldsymbol{J} + \dot{\boldsymbol{D}}$. For a simple circuit without branches this means that the total current through each cross-section is of equal magnitude. If the coordinate refers not only to the distance along a piece of dielectric but also to the distance along the entire circuit, then i_{c} and i_{D} in equation (4) may be functions of x, but i_{t} is independent of x. Equation (2) is now none other than equation (4) applied to the avalanche region, while equation (3) is obtained by averaging equation (4) over the drift region.

The analysis of the behaviour of the diode now depends on two main points. First, we want to know how exactly the avalanche layer responds to rapid fluctuations in voltage; in other words, we want to know the avalanche conduction current i_{ca} that results from the a.c. voltage v_{a}. Secondly, we want to know the average conduction current that arises in the drift region as a result of the avalanche conduction current i_{ca}. The main point here is that $\overline{i_{\mathrm{cd}}}$ has a phase shift with respect to i_{ca}. Once we have the answers to these two questions we can easily calculate the impedance of the diode, and in particular determine when its real part is negative.

The avalanche region

We assume that a sinusoidal a.c. voltage v_{a} of angular frequency ω is superimposed on the d.c. voltage $V_{0\mathrm{a}}$ across the avalanche region, this d.c. voltage being equal to $l_{\mathrm{a}}E_0$ in the situation shown in fig. 3. We now want to determine the resultant conduction current in the avalanche region (see *fig. 7*). During the half-period in which v_{a} is positive, i.e. has the same polarity as $V_{0\mathrm{a}}$, the concentration of charge carriers increases, and since the drift velocity is constant, the conduction current increases accordingly. During the half-period in which v_{a} is negative, the conduction current decreases. We thus have a varying conduction current which reaches its maximum at the end of each half-period of positive v_{a}, and thus lags a quarter of a period behind v_{a}. (In fig. 5 this has been taken into account.)

At sufficiently small values of the amplitude V_{a} of v_{a}, the behaviour of the avalanche layer is linear. The conduction-current variation is then also sinusoidal and the amplitude I_{ca} is proportional to V_{a}. At large values of V_{a} the total conduction current takes the form of a series of short pulses. The average current remains constant at I_0, the value determined by the current supply. The Fourier expansion of such a series of very short pulses — whose repetition frequency is $\omega/2\pi$ and whose average current is I_0 — is

$$I_0 + \sum_{n=1}^{\infty} 2I_0 \cos n\omega t. \qquad (5)$$

The amplitude of each Fourier component for large V_{a} is thus equal to $2I_0$, irrespective of the values of V_{a} and ω. This applies in particular to $i_{\mathrm{c\grave{a}}}$, the component of angular frequency ω, the only one we shall consider. We therefore see that I_{ca}, the amplitude of the first Fourier component of the avalanche conduction current, plotted as a function of V_{a}, initially increases in proportion to V_{a}, but at large values of V_{a} it asymptotically approaches the value $2I_0$. *Fig. 8* shows I_{ca} as a function of V_{a} in accordance with the more detailed theory. An outline of this theory is given in the Appendix to this article.

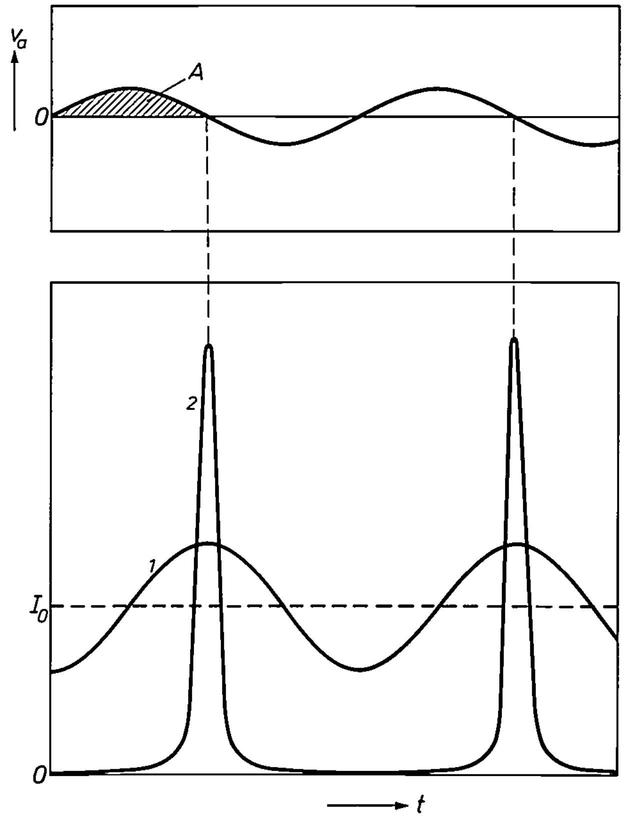


Fig. 7. The lower figure gives the total conduction current in the avalanche layer as a function of time, which results when an a.c. voltage v_a (upper figure) is superimposed on the d.c. voltage across the avalanche layer. If V_a, the amplitude of v_a, is relatively small, the total conduction current is the sum of the d.c. current I_0 and a sinusoidal current i_{ca} (curve *1*); if V_a is large, the conduction current consists of a series of pulses (curve *2*). In both cases the conduction current lags a quarter of a period behind v_a. The amplitude I_{ca} of i_{ca} is proportional to the area A.

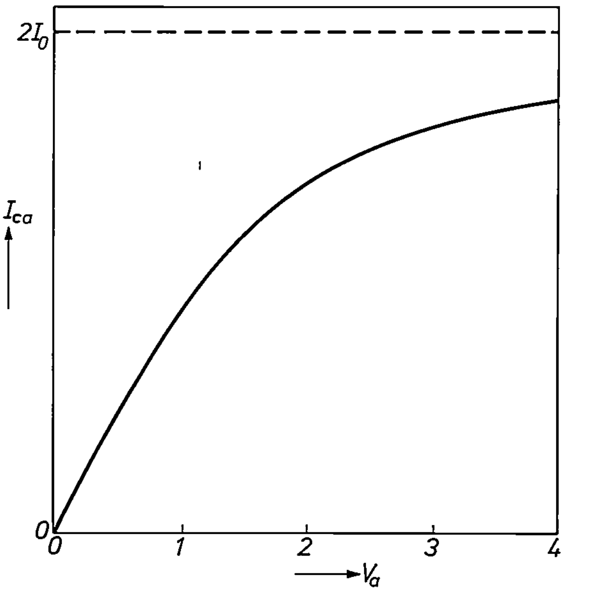


Fig. 8. The amplitude I_{ca} of i_{ca}, the first Fourier component of the avalanche conduction current, as a function of V_a. The equation for this curve is derived in the Appendix (see eq. 19). The unit for V_a (not given in the figure) is the value of V_a at which the transition takes place from a virtually linear to a distinctly non-linear behaviour. This is more explicitly defined in the Appendix.

The avalanche frequency

Let us now compare the a.c. conduction current i_{ca} with the capacitive current $i_{cap\,a}$ for the linear case (V_a small). As we saw in fig. 7, i_{ca} lags 90° in phase behind v_a; on the other hand $i_{cap\,a}$ (equal to $C_a dv_a/dt$) leads v_a by 90° in phase. The two currents are thus in opposition. If we let the frequency increase from low initial values, i_{ca} will decrease in inverse proportion to ω, because the production of extra charge carriers during a positive half-period of v_a is proportional to the area A in fig. 7. The amplitude $I_{cap\,a}$ of $i_{cap\,a}$, on the other hand, increases in proportion to ω. The ratio $i_{cap\,a}/i_{ca}$ is thus proportional to ω^2, and negative: $i_{cap\,a}/i_{ca} = -k\omega^2$, where k is a positive proportionality factor. The *avalanche frequency* ω_a is now defined as the frequency at which $i_{cap\,a}$ and i_{ca} exactly compensate each other, and at which the *total a.c. current i_t is therefore zero*. The factor k is therefore equal to $1/\omega_a^2$, and we have:

$$i_{cap\,a}/i_{ca} = -\ \omega^2/\omega_a^2. \tag{6}$$

The total a.c. current i_t is given by:

$$i_t = i_{cap\,a} + i_{ca} = (1 - \omega^2/\omega_a^2)i_{ca}. \tag{7}$$

The quantity ω_a introduced in equation (6) is of course independent of ω. It is also independent of V_a, since I_{ca} and $I_{cap\,a}$ are both proportional to V_a. Both these statements only hold true, however, as long as the avalanche layer behaves linearly. At larger values of V_a the value of I_{ca} is no longer proportional to V_a (fig. 8), and even $i_{cap\,a}/i_{ca}$ is no longer proportional to ω^2. Nevertheless, we shall continue to use equations (6) and (7) at larger values of V_a. The quantity ω_a in these equations is then however a function of V_a and ω. Since I_{ca} then increases less than proportionately with V_a (see fig. 8), ω_a decreases with increasing V_a. Graphs of ω_a as a function of V_a, derived from the more detailed theory, are given in fig. 11.

Finally we note that in the avalanche process the production of extra charge carriers is always proportional to the number already present. In the linear approximation this means that I_{ca} is proportional to I_0, and since $I_{cap\,a}$ is independent of I_0, it follows from equation (6) that ω_a^2 *is proportional to* I_0. This is found to remain valid for higher V_a.

The expression that will presently be given for the impedance Z and the resistance R of the diode do not explicitly contain the signal level (V_a) and the supply current I_0, but they do contain the avalanche frequency ω_a. The avalanche frequency ω_a can therefore be considered as a parameter, giving the dependence of Z and R on V_a and I_0.

In what follows we shall use the complex representation of a.c. quantities to calculate the relations between the a.c. voltages and currents of angular frequency ω. We shall not introduce any new symbols, but will assume that the variation with time of the quantities v_a, i_{ca}, etc. is given by exp $j\omega t$.

The drift region

The conduction current is always proportional to the charge-carrier concentration, and its value in the drift region thus follows the charge variations (see fig. 5). The current i_{cd} therefore lags farther in phase behind the avalanche current i_{ca} the farther away the cross-section in question is from the avalanche layer. To be more exact, the value of i_{cd} at a distance x from the avalanche layer at the time t is equal to i_{ca} at the time $t - x/u_s$. Using the complex representation we thus have $i_{cd}(x) = i_{ca} \exp(-j\omega x/u_s)$. Integrating this over x from 0 to l_d, the length of the drift region, and dividing by l_d, we find the average we are looking for, $\overline{i_{cd}}$, and hence the induction current i_i. The result is:

$$i_i = \Phi i_{ca}, \tag{8}$$

where Φ is the complex function

$$\Phi = \frac{e^{-j\theta} - 1}{-j\theta} = \frac{\sin\theta}{\theta} - j\,\frac{\sin^2 \frac{1}{2}\theta}{\frac{1}{2}\theta}. \tag{9}$$

Here θ is equal to $\omega\tau$, the transit angle. The transit time τ is directly related to l_d by the relation $l_d = u_s\tau$. As mentioned earlier, the main point is that i_i has a phase shift with respect to i_{ca}, so that Φ is not real. It will later be shown that the optimum value of θ — the 0.74π mentioned earlier — is in fact the value at which the imaginary part of Φ is a maximum.

The impedance of an avalanche transit-time diode

If we make a vector diagram of the various a.c. voltages and currents in the complex plane (*fig. 9*), we see readily that the impedance of the diode has a negative real part when $\omega_a < \omega$ [12].

We plot the voltage across the avalanche layer v_a along the positive imaginary axis. The conduction current i_{ca}, the capacitive current $i_{cap\ a}$ (which is equal to $j\omega C_a v_a$) and the total current i_t through the avalanche region then appear along the real axis, i_{ca} along the positive axis, $j\omega C_a v_a$ along the negative axis, and i_t also appears along the negative axis if $\omega_a < \omega$. The current $i_i = \Phi i_{ca}$, which is induced in the external circuit by the electrons in the drift region, appears in the lower half of the complex plane (see equation (9), θ is positive). The capacitive current in the drift region $i_{cap\ d}$ (equal to $j\omega C_d v_d$) is obtained by vector subtraction of i_i from i_t, and appears in the upper half. The voltage across the drift region v_d thus appears in the right-hand half, as does also the total voltage across the diode $v_t = v_a + v_d$. The impedance of the diode $Z = v_t/i_t$ appears in the negative half, because i_t is negative and real, and thus has a negative real part.

It is evident that two things are essential for Z to have an appreciable negative real part:

1) The ratio i_t/i_{ca} must be negative, i.e. ω_a must be smaller than ω.
2) The transit time in the drift region must be sufficiently long. If this is not the case, then i_i is almost completely real and consequently $j\omega C_d v_d$ is also almost completely real. But then v_d is almost purely imaginary, as are v_t and Z.

We can of course also calculate the small-signal impedance directly. Combining the equalities $Z = v_t/i_t = (v_a + v_d)/i_t$, $v_a = i_{cap\ a}/j\omega C_a$, $v_d = i_{cap\ d}/j\omega C_d$ with equations (3), (6), (7) and (8), we find:

$$Z = \frac{1}{j\omega C_a\,(1 - \omega_a^2/\omega^2)} + \frac{1}{j\omega C_d}\left(1 - \frac{\Phi}{1 - \omega^2/\omega_a^2}\right). \tag{10}$$

The real part of the impedance is given by:

$$R = \mathrm{Re}\,Z = \mathrm{Re}\,\frac{-\Phi}{j\omega C_d(1 - \omega^2/\omega_a^2)} = \frac{1}{\omega C_d\,(1 - \omega^2/\omega_a^2)}\;\frac{\sin^2 \frac{1}{2}\theta}{\frac{1}{2}\theta}. \tag{11}$$

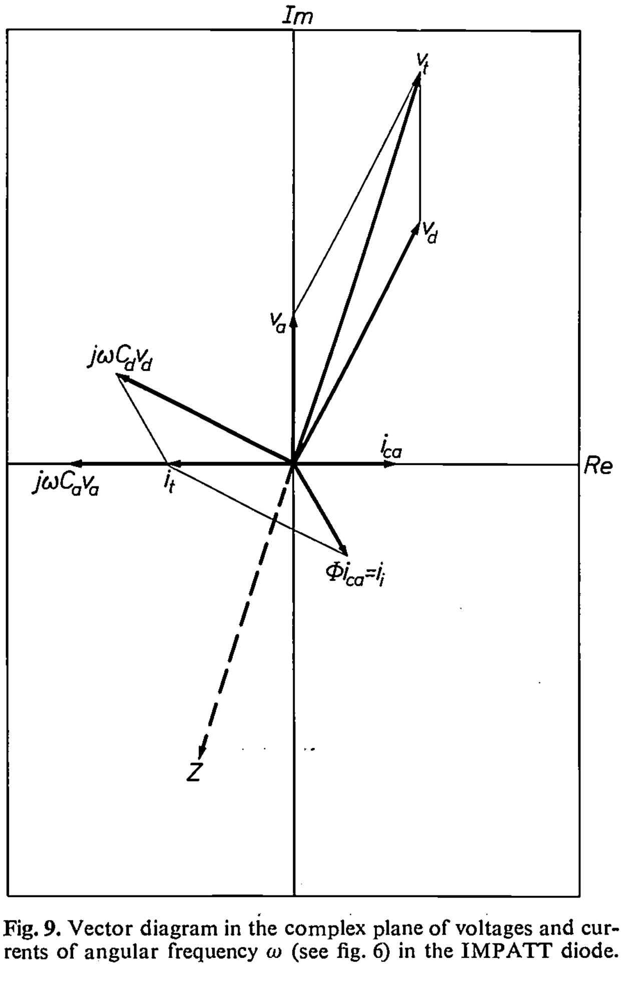


Fig. 9. Vector diagram in the complex plane of voltages and currents of angular frequency ω (see fig. 6) in the IMPATT diode.

In *fig. 10* $-R$ is shown as a function of ω_a, for a particular value of ω. Here again it is clearly seen that ω_a must be smaller than ω to obtain a negative resistance.

Operating point and performance of the diode

To determine the operating point of a diode in accordance with the model described, and to establish its theoretical performance, we assume that it is connected to an impedance $Z_L = R_L + jX_L$, forming a circuit that will oscillate. We assume that losses in the diode itself, which up to now have been neglected, can be taken into account by means of a series loss resistance R_s.

With the oscillator in the steady state the total impedance Z_t is zero:

$$Z_t \equiv Z + R_s + Z_L = 0. \qquad (12)$$

This equation for complex quantities is equivalent to two equations for real quantities:

$$\mathrm{Re}\, Z_t \equiv R + R_s + R_L = 0, \qquad (12a)$$

$$\mathrm{Im}\, Z_t \equiv X + X_L = 0. \qquad (12b)$$

From these we can determine the frequency and the level of the oscillation.

For a given diode we now have three parameters available: the external parameters X_L and R_L, and the d.c. current through the diode I_0. This current, like V_a, is contained in ω_a in equation (10). In approximate terms X_L tunes the circuit to a particular frequency, R_L matches the external circuit to the diode (to allow the total available power to be extracted), and I_0 controls the level of oscillation. These processes are not in fact independent. It does however follow, from (12) and (10), that ω is only a function of X_L and R_L and not of I_0 and V_a [13]. Within the scope of our model, therefore, *controlling the oscillation level by I_0 does not affect the frequency.*

The independence of ω from I_0 and V_a can be seen from the following. Substituting (10) in (12) yields an equation from which $(\omega_a/\omega)^2$ can be found:

$$\left(\frac{\omega_a}{\omega}\right)^2 = g(R_L, X_L, \omega)\,. \qquad (13)$$

From this equation ω and ω_a can be determined by equating the real and the imaginary parts. Since ω_a is real, one of the equations:

$$\mathrm{Im}\, g = 0,$$

contains only ω, R_L and X_L, but not ω_a, nor therefore I_0 and V_a.

The operating point of the diode is now easily found from fig. 10 and from *fig. 11*, which shows ω_a calculated as a function of V_a for different values of I_0 from the theory outlined in the Appendix. With the aid of X_L, we choose a particular oscillating frequency ω. Since $-R$ must have the value $R_s + R_L$ (eq. 12a), we read off the required value $\omega_{a\,\mathrm{osc}}$ of ω_a directly from fig. 10, and we find from fig. 11 the oscillation level V_a that follows for a given supply current I_0. We also see that oscillations are not possible until I_0 has exceeded a critical starting value, I_{start}.

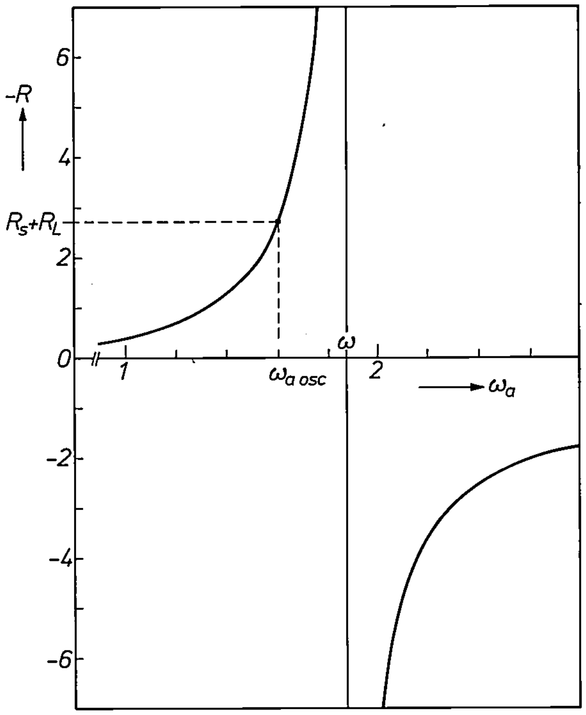


Fig. 10. The negative resistance of the diode, $-R$, as a function of the avalanche frequency ω_a, as given by equation (11), for a particular signal frequency ω. During oscillation, $-R$ has the value $R_s + R_L$ (see eq. 12a). This gives the value of ω_a at oscillation, $\omega_{a\,\mathrm{osc}}$, and from this the oscillation level is found (see fig. 11). The unit for R is chosen such that the value along the vertical scale is in fact the value of $1/(1 - \omega^2/\omega_a^2)$. The unit for ω_a is explained in the caption to fig. 11.

At a given ω we can also derive from figs. 10 and 11 the direct relationship between $-R$ and V_a (for different I_0). The result can be seen in *fig. 12.* From this figure we can read off directly the combinations of I_0 and V_a at which the condition $-R = R_s + R_L$ is satisfied.

If at the beginning of an experiment ($V_a = 0$) I_0 is suddenly given the value of curve *3* in fig. 12 (point *A*), then since $-R$ is greater than $R_s + R_L$ the situation is *unstable.* Oscillations of increasing amplitude V_a occur, and this increase continues until the point *B* is reached. If we then increase I_0 successively to the values *4*, *5*, . . . , the oscillation adjusts itself to the points *C*, *D*, . . . ; for each I_0 we read off the corresponding V_a.

A remarkable situation is found if at the beginning of the experiment ($V_a = 0$) I_0 is suddenly given the value of curve *5* (point *E*). All resistances in the circuit are positive, every fluctuation is damped out, and the

[12] G. A. Acket and M. T. Vlaardingerbroek, Festkörperprobleme **9**, 280, 1969.

[13] M. T. Vlaardingerbroek and J. J. Goedbloed, Philips Res. Repts. **25**, 452, 1970.

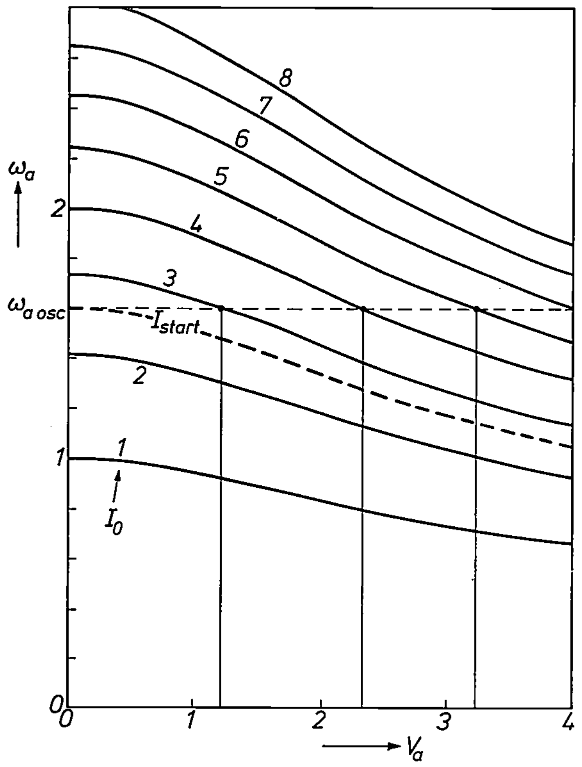


Fig. 11. The avalanche frequency ω_a as a function of the amplitude V_a of the a.c. voltage across the avalanche region (as given by eq. (20) in the Appendix). For each value of I_0 the oscillation level follows from $\omega_{a\,osc}$, the value found for ω_a in fig. 10. For each $\omega_{a\,osc}$ there is a starting current I_{start}, which is the value of I_0 at which the circuit begins to oscillate as I_0 is increased. The unit for V_a is the same as in fig. 8. An arbitrary unit is used for I_0, and the unit for ω_a is the small-signal value which ω_a assumes when I_0 has its unit value (see Appendix).

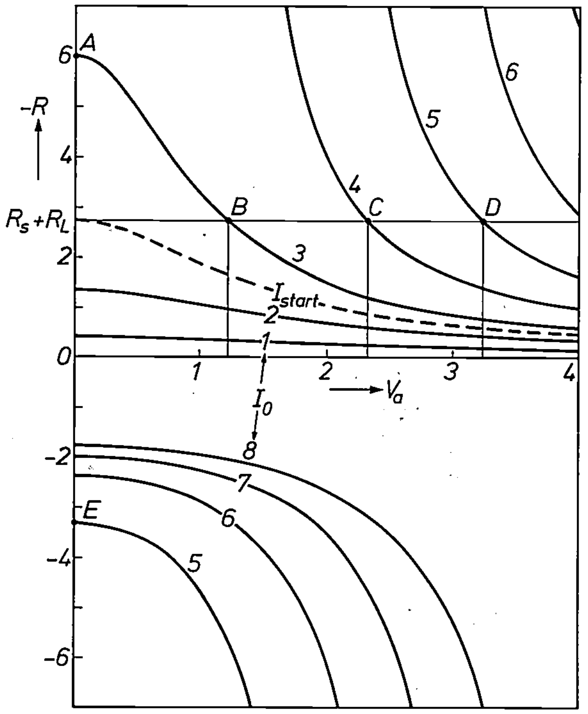


Fig. 12. The negative diode resistance $-R$ as a function of V_a. These curves are obtained by combining figures 10 and 11. The oscillation level can be read off directly given $R_s + R_L$ and I_0.

situation is stable at zero signal level. If however it is possible to introduce a signal so large that the vertical asymptote of curve *5* is exceeded, then the oscillations stabilize at the point *D*. This often happens automatically as a result of the switching transient.

The power that the diode delivers to the external load R_L is $\frac{1}{2}I_t^2R_L$, where I_t is the amplitude of the total a.c. current i_t. With the aid of equations (12a), (11) and (7) this output power P_o is found to be [7] [9]:

$$P_o = \tfrac{1}{2}I_{ca}^2 \frac{1}{\omega^2 C_d^2}\left(\frac{\sin^2 \frac{1}{2}\theta}{\frac{1}{2}\theta}\right)^2 \frac{R_L}{(R_s + R_L)^2}. \tag{14}$$

P_o depends on I_{ca}, and hence on the signal level V_a (see fig. 8), which has yet to be determined for a given supply current I_0 in the manner described above.

If the load is matched to the diode, i.e. R_L is made equal to R_s, the diode will deliver the "available power" P_a at a given I_0. If V_a is so large that we can put $I_{ca} = 2I_0$ (see fig. 8), then P_a is given by:

$$P_a = \frac{I_0^2}{2\omega^2 R_s C_d^2}\left(\frac{\sin^2 \frac{1}{2}\theta}{\frac{1}{2}\theta}\right)^2 \qquad (V_a \to \infty). \tag{15}$$

To find the efficiency we must divide P_a by the d.c. power I_0V_0, where V_0 is the d.c. voltage across the diode.

Some experimental results

Table I gives the output power and efficiency of diodes of various types made at Philips Research Laboratories.

Equation (15) would seem to suggest that an arbitrarily high output power might be obtained by stepping up the supply current I_0 indefinitely. In practice, of course, there are limits: if I_0 becomes too high the diode becomes too hot to function properly, since the heat generated cannot be dissipated. Our diodes could handle current densities up to about 1000 A/cm² without getting too hot. There is yet another effect that limits the power obtainable even more severely, and this will now be discussed.

Table I. Output power P_o and efficiency η of some IMPATT diodes made at Philips Research Laboratories. The "maximum permissible voltage swing" (see p. 337) of some types is expressed as a fraction γ of the breakdown voltage.

Type of diode	f	P_o	η	γ
Si-P^+N	5 GHz	1.5 W	7%	
Si-P^+N	6 GHz	1.4 W	7%	
Si-N^+P	6 GHz	1.3 W	6%	
Si-Schottky-b.	7 GHz	0.6 W	5%	
Si-P^+N	9 GHz	0.9 W	7%	0.37
Si-N^+P	9 GHz	0.8 W	6%	
Si-Schottky-b.	9 GHz	0.3 W	5%	
Ge-N^+P	9 GHz	0.5 W	14%	0.50
GaAs-Schottky-b.	9 GHz	0.7 W	15%	0.45

At small currents (30 to 40 mA) good agreement is found between theory and experiment. *Fig. 13* shows an example. Here the output power of a large series of diodes, showing marked differences in series resistance R_s and capacitance C, is plotted against $(R_sC^2)^{-1}$. The supply current and frequency were the same in all cases. As can be seen, the experimental points closely follow a straight line, as is predicted by the theory [14].

However, at higher currents (but not such high currents that the thermal effect is significant) the maximum output power is noticeably lower than the theory predicts. This was particularly evident in measurements at higher currents on a series of diodes with lower values of the resistance R_s ($\frac{1}{2}$ Ω instead of 1 Ω). The higher power that would have been expected from equation (15) was not reached. We note that both a reduction in R_s and an increase in I_0 lead to a greater voltage swing (as can immediately be seen in fig. 12). And indeed the failure of the theory does appear to be directly related to the magnitude of the voltage swing: there is a *maximum permissible voltage swing* which is a certain fraction of the breakdown voltage for any kind of diode; above that voltage the diode is no longer active and the theory no longer holds. This was confirmed by impedance measurements carried out at these Laboratories by B. B. van Iperen and H. Tjassens [15] on germanium N^+-P and silicon P^+-N diodes and also on Schottky-barrier gallium-arsenide diodes. We shall now briefly consider these measurements.

The measurements are concerned with impedance as a function of signal level, and are carried out in two steps. In the first place the *small-signal impedance* at a single frequency is measured with the aid of a microwave impedance bridge, as a function of the supply current, steps being taken to prevent the diode from oscillating during the measurement. The result for a germanium N^+-P diode is given in *fig. 14*. The impedance was measured before breakdown as a function of the d.c. voltage. The resistance at breakdown, when I_0 is still zero, is R_s, since ω_a and therefore R are zero (see eq. 11). The impedance transformation between the waveguide, where the measurements are made, and the diode terminals is a difficult point [16].

The results of this measurement are needed for the next step, which is to determine the *large-signal impedance* [16]. For this measurement the diode is incorporated in a coaxial resonant cavity. At a given setting of the resonant cavity and of the external load the d.c. cur-

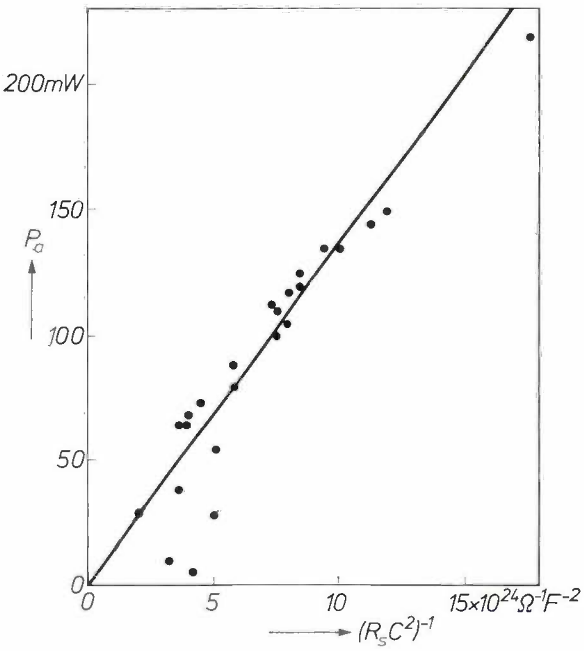


Fig. 13. The available power P_a as a function of $(R_sC^2)^{-1}$ for a large number of diodes with various values of loss resistance R_s and capacitance C; after B. B. van Iperen, H. Tjassens and J. J. Goedbloed [14]. The frequency in all cases was 10 GHz, the supply current 40 mA. Equation (15) indicates a linear relationship between P_a and $(R_sC^2)^{-1}$. The capacitance C_d in equation (15) is equal to $C(l_a + l_d)/l_a$.

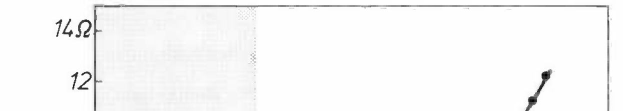


Fig. 14. The impedance (negative resistance $-(R + R_s)$ and reactance X) of a germanium N^+-P diode for small signals (after Van Iperen and Tjassens [15]). The left-hand (shaded) region shows the situation before breakdown; the measured values are plotted here as a function of the d.c. voltage V_0 across the diode. The measured values are plotted in the right-hand region (after breakdown) as a function of the supply current I_0.

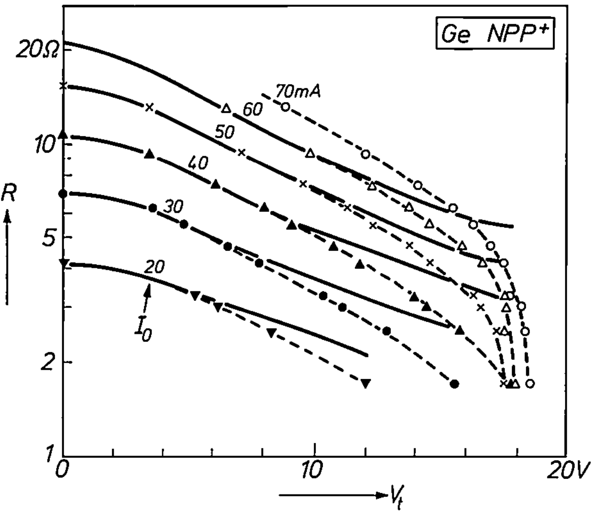
[14] B. B. van Iperen, H. Tjassens and J. J. Goedbloed, Proc. IEEE **57**, 1341, 1969.

[15] B. B. van Iperen and H. Tjassens, Proc. MOGA Conf., Amsterdam 1970, page 7.27.

[16] B. B. van Iperen, IEEE Trans. **MTT-16**, 961, 1968.
B. B. van Iperen and H. Tjassens, Philips Res. Repts. **27**, 38, 1972 (No. 1).

rent is raised until the point is reached at which the diode just starts to oscillate. The impedance of the diode at this instant is known; it is the small-signal impedance measured earlier. The impedance of the loaded cavity is then also known, since $Z_{\text{total}} = 0$. At the same setting of the resonant cavity and external load the output power is now measured at a number of higher currents. When the results are used to make a plot of $-R$ against V_t — there are two examples in *fig. 15* — a series of points on a horizontal line is obtained, since the diode impedance keeps the same value, which is the negative of the unchanged impedance of the loaded cavity, while the output power is proportional to $V_t{}^2$. This measurement is repeated for other cavity settings, with the circuit oscillating at the same frequency in all cases. By joining up the points belonging to the same I_0 the required relationship between impedance and signal level can be found.

Fig. 15 shows the results for a germanium diode and a silicon diode. The theoretical curves in these figures correspond to curves like *1*, *2* and *3* in fig. 12, after duly converting V_a to V_t (see fig. 6). For weak signals the theory is very satisfactory for germanium and moderately so for silicon. As stated above, however, it is clear from these figures that the theory does not hold for larger signals. The "maximum voltage swing", the value of V_t at which the (extrapolated) negative resistance becomes zero, is not strongly dependent on the supply current. This maximum V_t is a particular fraction γ of the breakdown voltage for every type of diode. Values of γ are presented in Table I. It can be seen from this table that the efficiency is particularly high in those diodes (the germanium and galium-arsenide diodes) in which the maximum voltage swing is also relatively high. Apparently the efficiency obtainable depends to a great extent on the maximum voltage swing [17]. The mechanisms to which the occurrence of a voltage-swing limit might be attributable [18] will not be considered here.



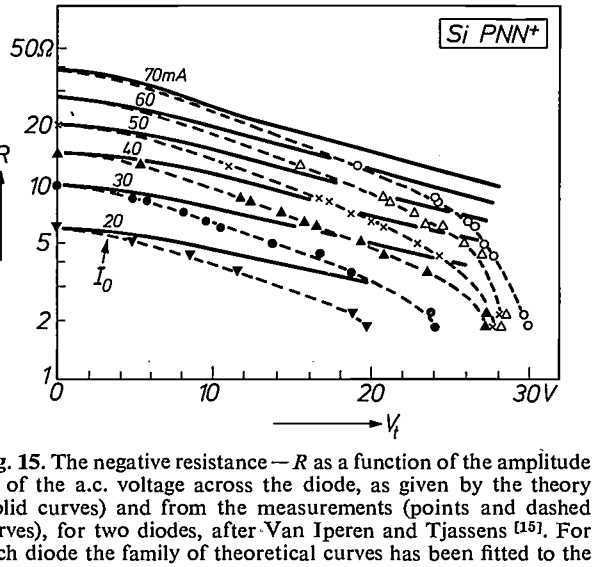


Fig. 15. The negative resistance $-R$ as a function of the amplitude V_t of the a.c. voltage across the diode, as given by the theory (solid curves) and from the measurements (points and dashed curves), for two diodes, after Van Iperen and Tjassens [15]. For each diode the family of theoretical curves has been fitted to the measurements at only two points (e.g. $-R$ at $V_t = 0$, $I_0 = 20$ mA and $-R$ at $V_t = 0$, $I_0 = 70$ mA).

Noise

The usefulness of an oscillator is largely determined by its noise characteristics. Whereas the ideal oscillator would oscillate at constant amplitude at a single frequency, any practical oscillator gives continuous variations in frequency (frequency noise) and amplitude (amplitude noise).

The main source of noise in the IMPATT diode is the "graininess" of the supply current (shot noise) in the avalanche layer, magnified by the avalanche effect. The "primary noise current" [19] is used to describe this noise source; it is defined as the noise current that would flow through the diode if the voltage across the avalanche region were constant.

The primary noise current affects the output signal from the oscillator in two ways. 1) Low-frequency components, which are prevented by the selectivity of the circuit from appearing directly in the output signal, modulate the supply current and thus give rise indirectly to "modulation noise" in the output signal. 2) High-frequency components whose frequency lies close to the oscillator frequency are amplified in the circuit and appear directly in the output signal ("intrinsic noise").

Modulation noise

The low-frequency components of the primary noise current cause slow modulation — slow with respect to the high-frequency signal — of the diode operating current and hence of the diode impedance. If the total impedance is to remain equal to zero — the condition for oscillation — the amplitude and frequency of the oscillation must vary continually. In general, therefore, amplitude noise and frequency noise due to current modulation may be expected [13]. Now in the diode model discussed a change in the supply current does not affect the frequency (see page 335), so that modulation noise here will have an amplitude character only. The frequency noise due to modulation of the supply current that does appear in practical diodes therefore falls outside the scope of the model discussed. On the other hand the model provides a satisfactory means of analysing the amplitude noise [20]. Modulation noise can be reduced by incorporating a high series resistance in the supply circuit.

Intrinsic noise

For small signals the intrinsic noise may be represented as follows. In an ideal oscillator circuit the total resistance is zero, and the oscillation frequency is such that the reactance is also zero. The complete absence of loss in such a circuit gives it an infinite Q (quality factor) or, in other words, it oscillates at only one frequency. If we now include a current source (e.g. the primary noise current) in the circuit, it no longer has to be completely loss-free for it to oscillate. The oscillating loop now has a total resistance different from zero, the Q is no longer infinite, and the signal spectrum therefore has a certain width. The greater the primary noise current the less the loss of the oscillating loop needs to be reduced and the spectrum becomes wider, which means that the noise increases. It seems to us that the intrinsic noise is mainly frequency noise. This can be reduced by using a high-Q resonant cavity in the circuit.

Summarizing in very approximate terms it appears that an IMPATT-diode oscillator has amplitude noise that is essentially modulation noise, and frequency noise that is essentially intrinsic.

With this understanding of the noise behaviour and with the diode theory outlined in the foregoing we can derive an analytical relation between the primary noise current of a diode and the noise spectrum of an oscillator using this diode [13]. This relation can be verified experimentally by measuring the primary noise current of the non-oscillating diode, and by determining the other diode parameters with impedance measurements as described above. Both for modulation noise [20] and for intrinsic noise [21] good agreement is found between theory and experimental results, provided the output level is not too high and the circuit is not tuned to higher harmonics.

Finally, we have collected some of the results of our intrinsic-noise measurements in *Table II* to give an idea of the noise level of the various types of diode. The measurements were carried out with the aid of a very simple oscillator circuit. The results in Table II are expressed in the theoretical half-width Δf_0 of the oscillation spectrum at half height, on the assumption that the broadening is solely due to frequency noise. Details of the way in which Δf_0 is derived from the measurements will be found in *fig. 16*. As can be seen from the table, the noise can be substantially reduced by incorporating a high-Q resonant cavity in the circuit.

Table II. Half-width Δf_0 (in Hz) at half height of the oscillator spectrum, derived from the results of noise measurements on four types of IMPATT diode. LQ and HQ with and without stabilizing resonant cavity in the oscillator circuit.

Diode	LQ	HQ
Si-P^+N	9500	180
Ge	3000	43
GaAs	1800	14
Si-N^+P	380	3.0

Technology

In discussing some of the technological aspects of the fabrication of IMPATT diodes we shall again confine ourselves to the silicon P^+-N diode. We have also made N^+-P silicon diodes, N^+-P germanium diodes and

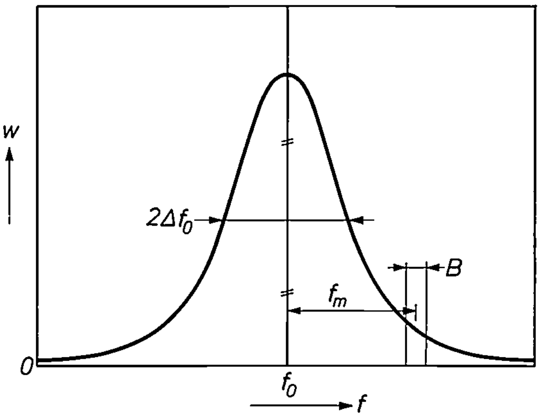


Fig. 16. Spectral-energy distribution w as a function of frequency f in an oscillation line broadened by frequency noise. By definition $w = \mathrm{d}P_0/\mathrm{d}f$, where $\mathrm{d}P_0$ is the power delivered in a frequency band of width $\mathrm{d}f$. In frequency noise measurements it is usual to determine the power, relative to the total output power, in a band B (in our case 100 Hz) displaced by f_m (in our case 100 kHz) from the centre frequency f_0. The usual practice is then to calculate from this the mean amplitude Δf_{rms} of the frequency swing corresponding to the frequency noise in this band. Very generally, Δf_{rms} is independent of f_m, and Δf_{rms} and B are related by $(\Delta f_{\mathrm{rms}})^2 = (1/\pi)\Delta f_0 B$, where Δf_0 is the half-width of the oscillation spectrum at half height [13]. This relation was used for calculating the values of Δf_0 given in Table II.

[17] The relation between an upper limit to the efficiency and the maximum voltage swing is given by equation (14) in the article by Mouthaan in this issue, page 345.

[18] B. B. van Iperen and H. Tjassens, Proc. IEEE 59, 1032, 1971 (No. 6).
E. Allamando, E. Constant, G. Salmer and A. Semichon, Acta Electronica 12, 211, 1969.
See also: Mouthaan [9], pp. 49 and 50.

[19] M. E. Hines, IEEE Trans. ED-13, 158, 1966.

[20] J. J. Goedbloed, Proc. MOGA Conf., Amsterdam 1970, page 12.36.

[21] J. J. Goedbloed, Electronics Letters 7, 445, 1971 (No. 16).

Schottky-barrier diodes of silicon [22] and of gallium arsenide (see Table I, page 336); similar considerations apply to these.

Technologically, a P^+-N diode (*fig. 17*) consists of a *substrate* of N^+ material on which an N-type *epitaxial layer* has been grown, which has subsequently been converted partially by diffusion into a P^+ *diffusion layer*. What is left after diffusion of the N-type layer must have a specific thickness, which, as we shall presently see, is closely related to the frequency. In a silicon diode for 10 GHz, for example, a thickness of about 3 μm is needed. The thickness of the diffusion layer will lie in the range from about 1 to 10 μm, depending on the type of diode. The substrate, acting as the base for various fabrication steps, is in general thicker (e.g. 40 μm). A typical diode has a diameter of between 100 and 200 μm.

We have made two types: *planar diodes* (*fig. 18a*), made by the usual planar techniques, and the somewhat less conventional *mesa diodes* (fig. 18*b*), which have the form of plateaus (mesas) left behind when the surrounding material of a prepared chip is etched away.

The type chosen, the dimensions and other technological details are decided by the following considerations.

To begin with we note that the depletion region at the N side of the P-N junction should preferably be as wide as the N region (*fig. 19a*). If the N region is too wide (fig. 19*b*) the undepleted part of it (the "unswept region") makes an undesirable contribution to the series resistance R_s (see eq. 15). If, on the other hand, the depletion layer extends into the N^+ region (fig. 19*c*) the d.c. voltage across the diode (the area under the $E(x)$ curve) and hence the power supplied will be larger than necessary. Once the required thickness of the depletion layer has been fixed (see below) the situation shown in fig. 19*a* then has to be achieved by appropriate doping of the N region. We gave considerable attention to this point in the series of diodes mentioned earlier with a series resistance of about $\frac{1}{2}$ Ω (page 337).

The *optimum thickness* of the depletion layer (and hence of the N layer) is determined in the first place by the value of θ — the 0.74π mentioned before — for which the factor $(\sin^2 \frac{1}{2}\theta)/\frac{1}{2}\theta$ in the expression for the output power (eq. 15) is a maximum. In terms of the wavelength λ of the required oscillations ($\lambda = 2\,\pi c/\omega$) and the length l_d of the drift region ($l_d = u_s\tau$) setting the "$\omega\tau \approx 0.74\pi$" condition for silicon ($u_s \approx 1\times10^7$ cm/s) is approximately equivalent to specifying $l_d \approx 10^{-4}\lambda$. For silicon we thus have the rule of thumb that we need 1 μm of depletion layer thickness for each centimetre of wavelength; e.g. for a wavelength of 3 cm (10 GHz) we need a thickness of 3 μm. In the second

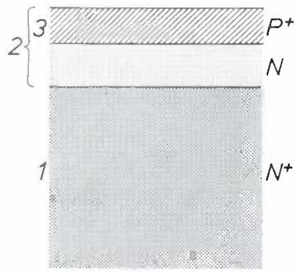


Fig. 17. A P^+-N diode (schematic). *1* substrate, *2* epitaxial layer, and *3* diffusion layer.

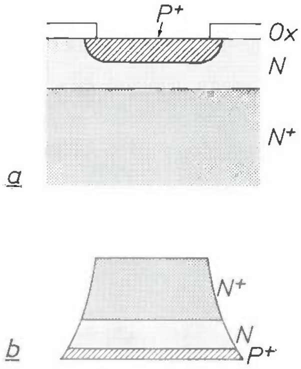


Fig. 18. *a*) Planar P^+-N diode. *Ox* oxide layer. *b*) Mesa P^+-N diode.

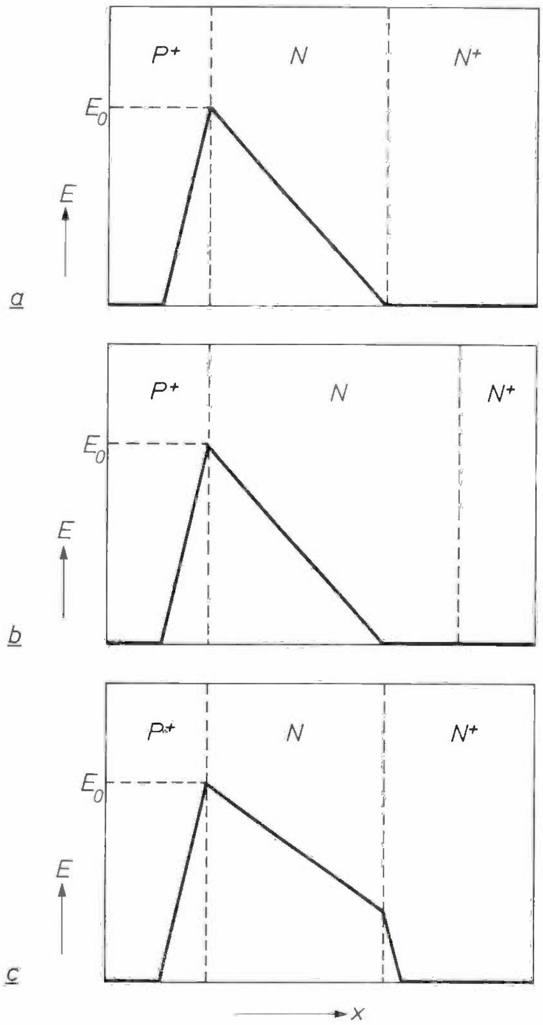


Fig. 19. Field profile $E(x)$ in a P^+-N diode at breakdown in three situatons. *a*) The right-hand boundary of the depletion region coincides with the N-N^+ junction. *b*) It lies in the N region. *c*) It lies in the N^+ region. Situation (*a*) is the ideal one.

place it is necessary to bear in mind that C_d depends on the thickness, but this point will not be dealt with here.

The *diameter* of the diode must remain limited to keep C_d within limits. This is particularly important at high frequencies, where the depletion layer must be thin. On the other hand, too small a diameter will make heat dissipation difficult; we shall have more to say about this presently. At 10 GHz we use diameters of about 140 µm, and at 30 GHz it is desirable to keep the diameter to within 50 µm to ensure satisfactory efficiency.

It is essential that the diode should break down homogeneously over the whole area of the *P-N* junction. Diodes which have received no particular attention to achieve this almost invariably break down at the edges. This is the case in planar diodes, because of the sharp edge of the diffusion region, the field concentrating at the small radius of curvature there (*fig. 20a*). A deeper diffusion (fig. 20*b*) results in a greater radius of curvature, and this alone therefore gives an improvement. If the diffusion layer is made appreciably thicker than the rest of the *N* layer (fig. 20*c*) the field at the edge is even weaker than in the middle. The breakdown voltage is now determined by the thickness of the *N* layer, and this is the method used to avoid edge breakdown in planar diodes.

A disadvantage of this method, and hence a weakness of the planar diode, is that it always gives to some extent the non-ideal situation of fig. 19*c*. Against this is the advantage that, starting from the same epitaxial material, diodes of different *N*-layer thicknesses, suitable for different frequencies, can be made by varying the thickness of the diffusion layer. From the same material we have made diodes for 5 GHz and also for 35 GHz. The thickness of the *N* layer follows directly from the breakdown voltage. It is even possible to apply a subsequent correction if immediately after diffusion the *N* layer is found to be too thick.

To be certain of homogeneous breakdown in mesa diodes it is necessary to pay attention to the slope of the side face. If we compare the P^+-N diodes in *fig. 21a* and *b* we see that the lines of force at the *P-N* junction in fig. 21*a* are bunched together at the edge, indicating an increase in field-strength, whereas in fig. 21*b* they fan out. In the first case breakdown occurs at the edge, and in the second case it does not. The type in fig. 21*b* is therefore the only one suitable for mesa P^+-N diodes.

Non-uniform breakdown can also be caused by imperfections in the *P-N* junction. For this reason it is vital to work in dust-free conditions. This applies in particular to the mesa diodes, where the diffusion layer is usually extremely thin (1 to 2 µm). Impurities at the surface can easily diffuse to the junction at the high temperature required for the P^+ diffusion.

Finally, we come to the subject of *heat removal*. This determines the d.c. current that can be used without the diode becoming too hot, and therefore partly determines the maximum output power obtainable (eq. 15). The heat generated is considerable. In a diode for a wavelength of 5 cm, for example, the depletion layer must be about 5 µm thick, giving a breakdown voltage of about 120 V. At a current of 100 mA this means a dissipation of no less than 12 W. Good thermal contact with a water- or air-cooled heat sink of copper or silver is therefore essential.

In this respect the planar diode is at a disadvantage

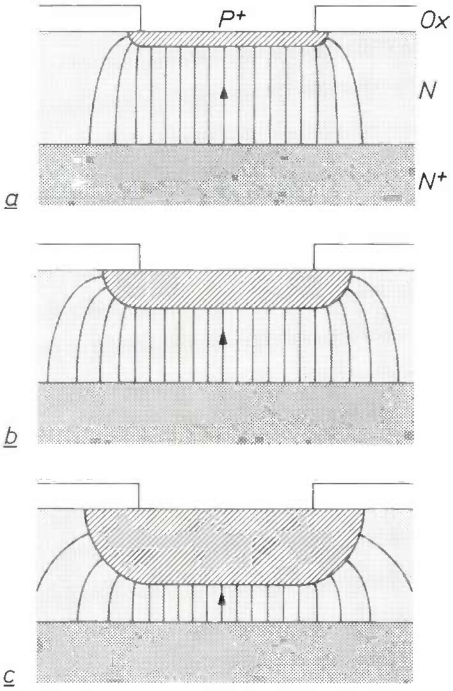


Fig. 20. Field lines in planar diodes with diffusion layers of various thicknesses. In case *a* breakdown definitely occurs at the edges, in *b* this is not so certain, in *c* there is no edge breakdown.

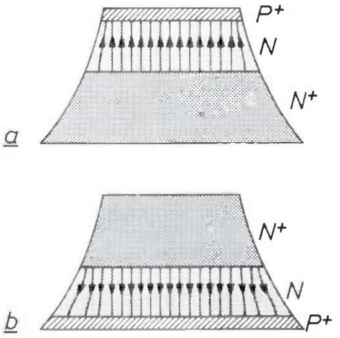


Fig. 21. Field lines in P^+-N mesa diodes with the layers in opposite sequence, so that in going from *N* to P^+ the cross-sectional area varies in opposite sense. In case *a* edge breakdown is to be expected in the *P-N* junction, in c ase*b* it is not.

[22] D. de Nobel and H. G. Kock, Proc. IEEE 57, 2088, 1969.

since the substrate side is mounted on the heat sink, and the heat generated therefore always has to pass through a relatively thick substrate.

The heat generated in the mesa P^+-N diode is directly dissipated through the P^+ diffusion layer, which is made very thin for this reason (1 to 2 μm). To ensure good thermal contact — and also to make the devices more manageable — the diodes are provided with an electrolytically formed layer of copper or silver on the diffusion side during fabrication. This layer is later soldered to the heat sink.

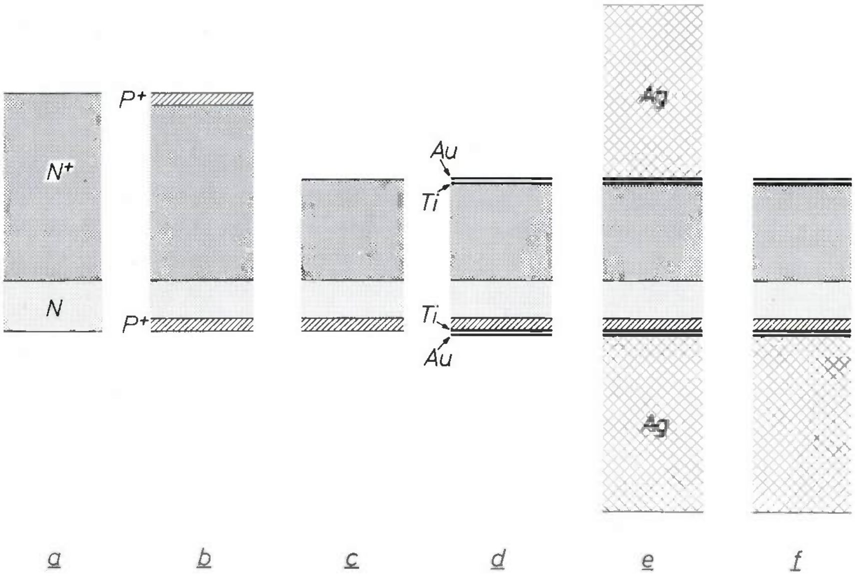


Fig. 22. Successive stages of a silicon slice from which P^+-N mesa diodes are made.

Although there are circumstances in which the planar diode is preferable, the mesa diode seems to us to offer better prospects, particularly since it can generate a reasonably high continuous microwave power. For that reason, and also because the planar techniques are well known, we shall say no more about the planar diode, and conclude this article with a brief description of the fabrication of mesa diodes.

The fabrication of mesa diodes

Fig. 22 shows the various stages that a slice of silicon passes through before mesa diodes are etched out of it. The operations between these stages are the following:
a) A slice of N^+ silicon is provided with an epitaxial N-type layer (e.g. 5 μm thick) and ground on the substrate side into a plane-parallel slice 100 μm thick.
b) A P^+ layer (1 to 2 μm) is formed on both sides by boron diffusion.

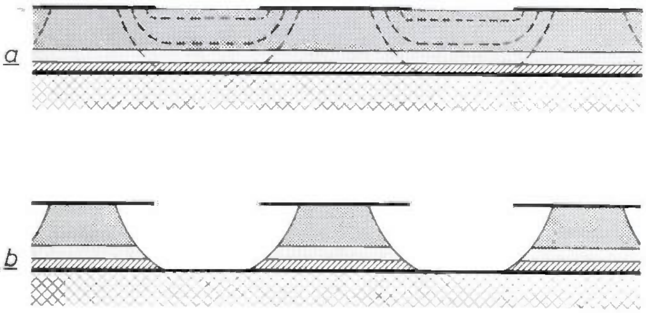


Fig. 23. Etching a slice into mesa diodes. Between the round Ti-Au areas, to which the Ti-Au layer of the last stage in fig. 22 is reduced by photo-etching, the silicon is etched away, leaving the mesas underneath. *a*) Various stages during etching, *b*) the situation immediately after etching.

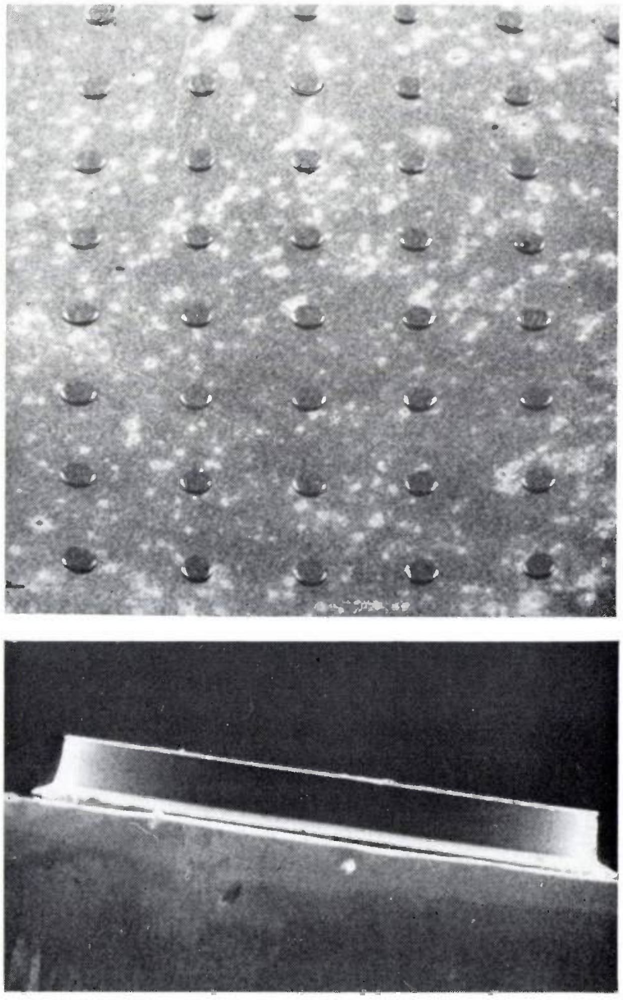


Fig. 24. *Above*: Part of a silicon slice after mesa etching (magnification 20×). *Below*: Side view of a diode (magnification 360×).

c) The slice is etched away on the substrate side to a thickness of about 40 μm. This is necessary to obtain the appropriate shape and proportions for the mesas.

d) A thin layer of titanium is deposited by sputtering or vacuum evaporation, and is followed by a thin layer of gold (0.5 μm). The gold serves as electrical and thermal contact, the titanium bonds the gold firmly to the silicon.

e) Thick layers of silver are electrolytically deposited (100 μm).

f) The upper layer of silver is etched away with dilute nitric acid, which does not attack the Ti-Au layer.

The slice is now ready for etching into mesa diodes (*fig. 23*). By means of a photo-etching technique the upper layer of Ti-Au is etched away except for a lattice of Ti-Au dots with a typical diameter of 160 μm. These act as a mask in etching the mesas. Some of the stages in the etching process are shown in fig. 23*a*. The typical mesa profile is obtained by "underetching" below the gold layer. Fig. 23*b* shows the situation after etching. The gold "overhang" at the tops of the mesas is removed ultrasonically or by blowing with a non-reactive gas. A lattice of mesas after this operation and a profile of a single diode can be seen in *fig. 24*. The silver underlayer is then cut into chips each containing one diode. With the procedure described here some hundreds of diodes are obtained from a single slice of silicon with a diameter of 3 cm.

The silver diode underlayer is then soldered to the base of a varactor package (see fig. 1). A gold contact wire (25 to 50 μm) is bonded to the substrate and to the edge of the varactor package by briefly heating the end of the capillary for manipulating the wire to about 400 °C (see *fig. 25*). Finally, with a short current pulse through the edge, the copper cap is bonded to the package, and the diode is ready to be tested.

We have had very good results with mesa diodes fabricated in this way, and we attribute these in particular to the combination of uniform breakdown and effective heat dissipation. A disadvantage is that the final shape and diameter are obtained by etching, which can lead to variations in diameter. To make a good diode (with the situation shown in fig. 19*a*) for a specific frequency it is necessary to have accurate control of the thickness and resistivity of the epitaxial layer. This requires a well defined resistivity transition from substrate to epitaxial layer ("steep epitaxy"). One final point: we have used Ti-Au contacts, which are very suitable for laboratory purposes. However, for professional diodes it will be necessary to use other contacts (e.g. Ti-Pt-Au, Mo-Au, Pt-Au, Pd-Au), because a Ti-Au contact quickly degenerates at slightly elevated temperatures owing to diffusion.

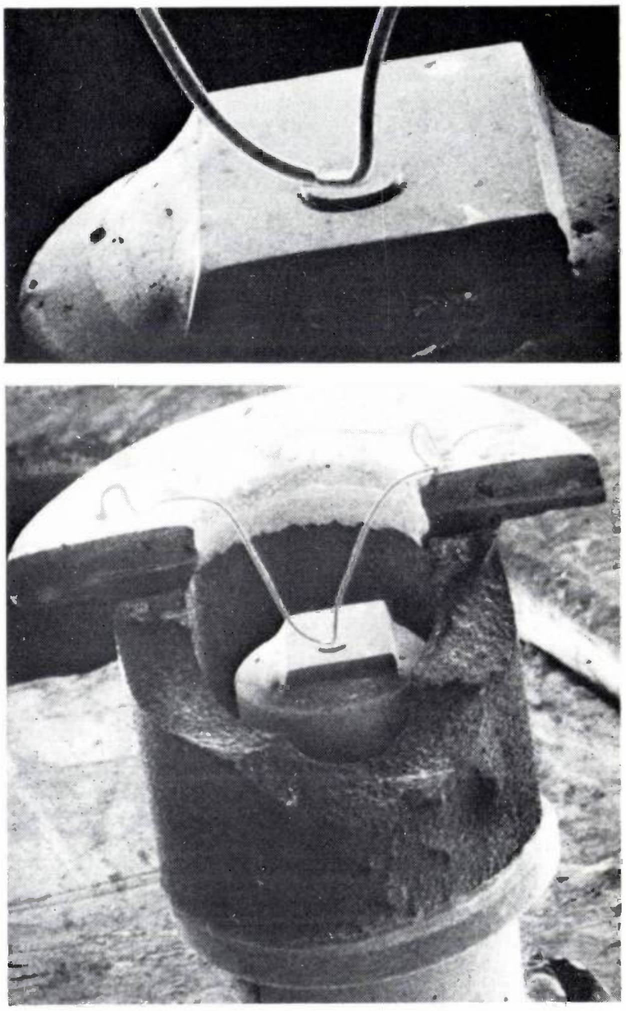


Fig. 25. Varactor package, with part removed to show the mounted diode and the bonded gold wire (magnification above 70 ×, below 20 ×). These photomicrographs, and those in fig. 24, were made with a scanning electron microscope, which gives a much greater depth of focus than optical microscopes. However, it does give misleading shades: here, for example, the almost black wall of the varactor package is in reality white.

Appendix: Theory of the avalanche layer

The exact form of the curves in figs. 8 and 11 can be derived by drawing up the balance of the number of charge carriers in the avalanche layer. For simplicity we assume that the drift velocity and the ionization coefficient for holes are the same as the drift velocity u_s and the ionization coefficient α for electrons. We also assume that the total charge-carrier density N and the electric field — and hence the conduction current density j_c ($= Neu_s$) and the ionization coefficient — are independent of the location in the avalanche layer. The increase per second in the number of charge carriers, both electrons and holes, per square centimetre cross-section of the avalanche layer, $\mathrm{d}(l_aN)/\mathrm{d}t = (l_a/eu_s)\mathrm{d}j_c/\mathrm{d}t$, consists of the following two contributions:

1) Generation by impact ionization. The charge carriers cover between them a total distance of l_aNu_s per second, and in doing so create αl_aNu_s pairs, i.e. $2\alpha l_aNu_s = 2\alpha l_aj_c/e$ charge carriers.

2) A negative contribution due to net transport through the boundaries to the areas outside. Consider first (in fig. 3) the right-hand boundary of the avalanche layer. Electrons (mainly formed by impact ionization) pass outwards through this bound-

ary at the rate of j_n/e per second, and holes thermally generated in the depletion region on the right of the avalanche layer pass inwards at the rate of j_{ps}/e per second, where j_{ps} is the hole contribution to the saturation current density. Since the total conduction-current density j_c is equal to $j_n + j_{ps}$ at this boundary, the net outward transport of charge carriers, $(j_n - j_{ps})/e$, may be written $(j_c - 2j_{ps})/e$. Similarly, $(j_c - 2j_{ns})/e$ charge carriers per second move out through the left-hand boundary. The total net transport outwards is therefore $2(j_c - j_s)/e$, where $j_s = j_{ps} + j_{ns}$ is the saturation current density.

For j_c we thus find the following equation:

$$\frac{l_a}{2u_s}\frac{dj_c}{dt} = (\alpha l_a - 1)j_c + j_s, \qquad (16)$$

which is known as Read's equation.

A steady state is only possible if $\alpha l_a \leqq 1$; if $\alpha l_a > 1$ the current increases exponentially. In the steady state, $dj_c/dt = 0$, and therefore

$$j_c = \frac{j_s}{1 - \alpha l_a} = Mj_s.$$

The factor M, which is equal to $1/(1 - \alpha l_a)$, is called the *current multiplication factor*. If j_c is much greater than j_s, then αl_a approximates to 1 (see eq. 1).

On the basis of such a steady state we shall now consider the effect of an a.c. voltage v_a across the avalanche layer. In equation (16) this appears as variations in α. If α in the steady state has the value α_0 and a derivative $\alpha' = d\alpha/dE$, we can then approximate α by $\alpha_0 + \alpha' v_a/l_a$. If we neglect j_s and put $\alpha_0 l_a = 1$, equation (16) becomes:

$$(l_a/2u_s)\,(dj_c/dt) = \alpha' v_a j_c. \qquad (17$$

Since v_a and j_c occur as a product, we cannot use the complex representation. We therefore equate v_a with the real part of $V_a \exp j(\omega t + \pi/2)$ (see fig. 9): $v_a = -V_a \sin \omega t$, and thus we obtain for j_c the equation:

$$\left(\frac{l_a}{2u_s}\right)\left(\frac{dj_c}{dt}\right) = -\,\alpha' V_a j_c \sin \omega t.$$

This can immediately be integrated. The result is:

$$j_c = j_0 \exp(\xi V_a \cos \omega t), \qquad (18)$$

where $\xi = 2\alpha' u_s/\omega l_a$. Equation (18) gives current pulses for large V_a as in fig. 7.

A special feature of the expression (18) for j_c is that its Fourier expansion with respect to ωt is at the same time an expansion with respect to Bessel functions of ξV_a:

$$j_c = j_0 \{I_{B0}(\xi V_a) + 2 \sum_{n=1}^{\infty} (-1)^{n+1} I_{Bn}(\xi V_a) \cos n\omega t\}.$$

The functions I_{Bn} are "modified Bessel functions". The coefficient j_0 is determined by the d.c. current density J_0, which is equal to $j_0 I_{B0}(\xi V_a)$. If we let V_a increase indefinitely, I_{B0} will tend to infinity and hence j_0 will tend to zero. The amplitude J_{c1} of the first Fourier component of j_c is:

$$J_{c1} = 2\,j_0 I_{B1}(\xi V_a) = \frac{2\,I_{B1}(\xi V_a)}{I_{B0}\,(\xi V_a)}\,J_0. \qquad (19)$$

As there is a constant fixed ratio between the currents (i,I) and the current densities (j,J), we can substitute I_{ca} for J_{c1} and I_0 for J_0 in equation (19), which gives us the equation for the curve in fig. 8. (The relevant units are given below.)

An expression for ω_a which has general validity within the terms of this model is found by taking the absolute value of the terms of equation (6) and substituting the expression we have just found for I_{ca} in the resultant equation, and also substituting $\omega C_a V_a$ for $I_{\mathrm{cap}\,a}$. The result is:

$$\omega_a^2 = \frac{2\omega I_0}{C_a V_a}\,\frac{I_{B1}(\xi V_a)}{I_{B0}(\xi V_a)}. \qquad (20)$$

This is the equation of the curves in fig. 11. If V_a becomes infinitely large, $I_{B1}(\xi V_a)/I_{B0}(\xi V_a)$ acquires the value 1, and therefore I_{ca} becomes equal to $2I_0$, while ω_a^2 becomes $2\omega I_0/C_a V_a$. For small values of ξV_a the value of $I_{B1}(\xi V_a)/I_{B0}(\xi V_a)$ is approximately equal to $\frac{1}{2}\xi V_a$. In particular this gives a relation between ω_a for small signals and α':

$$\omega_a^2 = \frac{\xi \omega I_0}{C_a} = \frac{2\alpha' u_s I_0}{l_a C_a} \quad (V_a \to 0). \qquad (21)$$

The small-signal behaviour is easier to find by directly linearizing equation (17), which enables the complex representation to be used, and v_a thus becomes once again a small complex quantity. We then write $j_c = J_0 + j_{c1} \exp(j\omega t)$, and neglect the produce $v_a j_{c1}$.

It only remains to mention the units we have used in figures 8, 10, 11 and 12. The quantity on the horizontal axes of figs. 8, 11 and 12 is really the dimensionless quantity ξV_a; in other words, V_a is expressed in terms of the unit $\xi^{-1} = \omega l_a/2\alpha' u_s$. In figs. 11 and 12 the praameter I_0 is expressed in terms of an arbitrary unit. If we call this I_{00}, then the avalanche frequency ω_a is expressed in figs. 10 and 11 in terms of the "unit" $\sqrt{(2\alpha' u_s I_{00}/l_a C_a)}$, which is in effect the small-signal value of $\omega_a(V_a \to 0)$ for the "unit of current" I_{00} (see fig. 11). Fig. 10 is valid for an angular signal frequency ω equal to $\sqrt{3.5}$ times the unit of ω_a. In figs. 10 and 12, $-R$ is expressed in terms of the unit $(2 \sin^2 \frac{1}{2}\theta)/\theta\omega C_d$.

Summary. The IMPATT diode, potentially the most powerful solid-state microwave generator, is a reverse d.c.-biased avalanching *P-N* junction, or Schottky barrier. The depletion layer consists of an avalanche region in which hole-electron pairs are formed by impact ionization and avalanching, and a drift region for the electrons or holes. When an a.c. voltage is superimposed on the d.c. voltage the pair formation becomes a periodic process and "packets" of charge carriers travel through the drift region. An analysis of the phase relations shows that the diode is active, i.e. acquires an impedance with a negative real part, when the drift region is sufficiently long and the "avalanche frequency" — a quantity that depends on supply current and signal level — is lower than the operating frequency. When the diode is used in an oscillator circuit the frequency is determined by the terminating impedance, and the level of oscillation by the supply current. Measurements on experimental silicon, germanium and gallium-arsenide diodes confirm the theory, except for large-signal operation. In the frequency range from 5 to 10 GHz output powers of 1 to 1.5 W have been obtained, and efficiencies of about 7%, and even greater than 15% in some cases. Deviations from the theory occur, and the diode loses its characteristics as an active device, when the amplitude of the a.c. voltage exceeds a third of the breakdown voltage (for silicon) or half the breakdown voltage (for germanium and gallium arsenide). Noise in the diode is attributed to the shot noise of the supply current modified by the avalanche effect; it appears in the output signal as modulation amplitude noise and as intrinsic frequency noise. The optimum thickness of the depletion layer of a silicon diode for centimetre wavelengths is a few microns, the diameter in practice is between 100 and 200 μm. Uniform breakdown, which is an essential prerequisite, is ensured in planar diodes by providing for deep diffusion, and in mesa diodes by giving the side face the proper slope. The heat generated in mesa diodes is effectively dissipated by direct contact between the thin diffusion layer and a heat sink. The various diffusion, etching and other processes used in the fabrication of mesa diodes are briefly described.

IMPATT-diode oscillators

K. Mouthaan

Introduction

Microwave energy can be produced with various kinds of semiconductor devices, and of the various practical possibilities the IMPATT-diode oscillator is potentially the most powerful. The physics and technology of the IMPATT diode, the active element in this oscillator, have been dealt with in detail in the preceding article [1]. In the present article we shall consider the IMPATT-diode oscillator as a whole, confining ourselves to oscillators that deliver c.w. microwave energy.

A simple oscillator construction is shown in *fig. 1*. It consists of a coaxial line with the diode mounted in the varactor package *A* at the end of the line. The coaxial output, which has a given characteristic impedance Z_0, is matched to the diode with a quarter-wave transformer *T*. The frequency of the microwave energy produced is determined by the distance from *T* to *A*. An experimental model of an oscillator designed along these lines is shown in fig. 11. *Fig. 2* shows the operating point of the diode on the current-voltage characteristic: the diode is reverse-biased into the breakdown region. The d.c. current through the diode is kept at a constant value I_0 by using a current source. The d.c. voltage V_0 across the diode is usually not very different from the breakdown voltage V_b. The microwave output power can be controlled by varying I_0. *Fig. 3* shows the output power and efficiency of an IMPATT-diode oscillator for 5 GHz.

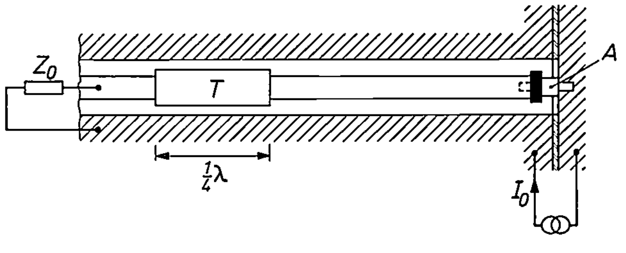


Fig. 1. A coaxial IMPATT-diode oscillator. *A* varactor package containing the active element, the IMPATT diode. The diode is supplied by a d.c. current I_0. The oscillator output is at the left. *T* quarter-wave transformer for matching the coaxial line to the diode. Z_0 characteristic impedance of the coaxial line. The oscillation frequency is determined by the distance from *T* to *A*. See also fig. 11.

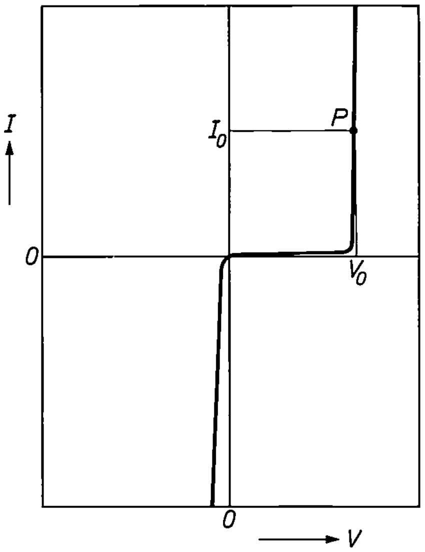


Fig. 2. Current-voltage characteristic of a diode (schematic). The diode is reverse-biased by a voltage high enough to cause breakdown. The diode is set to the operating point $P(V_0,I_0)$ by means of a d.c. current source (a source with a high internal resistance). The current and voltage in the reverse direction are shown here as positive, which is the only direction of importance here, unlike the situation with ordinary semiconductor diodes, in which the forward direction is also important.

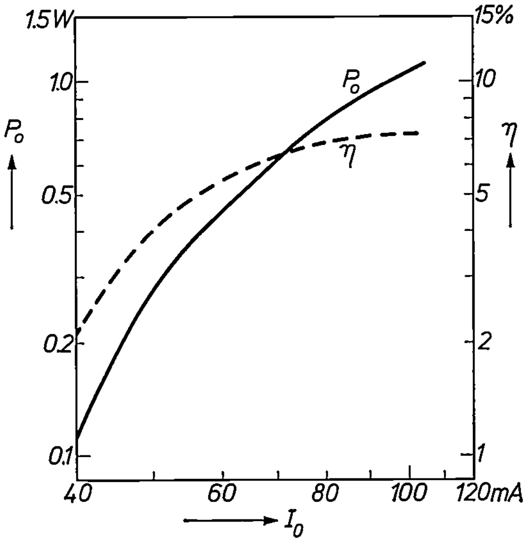


Fig. 3. Microwave output power P_0 and efficiency η as a function of the d.c. current I_0 of a c.w. oscillator for 5 GHz, of the type shown in fig. 1.

Dr. Ir. K. Mouthaan is with Philips Research Laboratories, Eindhoven.

[1] D. de Nobel and M. T. Vlaardingerbroek, IMPATT diodes; this issue, page 328.

In this article a few general performance characteristics of the oscillator will be discussed first: the level of oscillation, the output power, the efficiency and the starting current. Equivalent circuits for the diode and the oscillator will be produced which give a convenient survey of the microwave behaviour. These equivalent circuits are important aids in oscillator design. Two of the oscillators investigated at these Laboratories will then be discussed [2], and we shall examine some of the factors that determine the stability of an oscillator. This will be followed by a brief discussion of noise, and in conclusion we shall look at the possibility of improving the characteristics of an oscillator by tuning to higher harmonics.

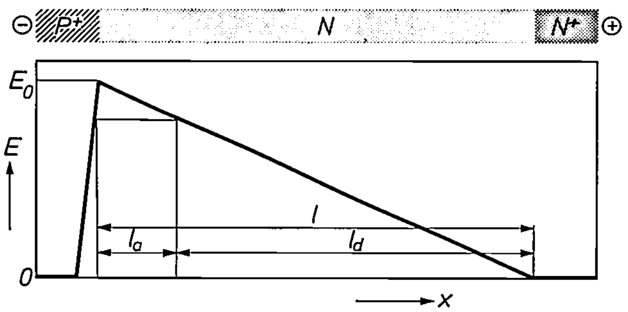


Fig. 4. Internal field profile of a reverse-biased avalanching P^+-N-N^+ diode. In the avalanche region (length $l_{\rm a}$) breakdown is caused by impact ionization and avalanche formation. The electrons produced travel at a constant velocity $u_{\rm s}$ in a time $\tau_{\rm d}$ through the drift region (length $l_{\rm d}$) to the N^+ contact region. Transit-time effects in the avalanche region and in the drift region for holes (left) are neglected.

Equivalent circuit of the IMPATT diode

To arrive at an equivalent circuit for the diode, we shall first summarize the behaviour of the diode for d.c. and a.c. applied voltages.

Let us consider a reverse-biased P^+-N-N^+ diode in the avalanching condition. The profile of the internal field in this situation is shown in *fig. 4*. Under the influence of the field the electrons travel to the right and the holes to the left. In the "avalanche region", of length $l_{\rm a}$, impact ionization and avalanching occur. If α is the ionization coefficient of the charge carriers (i.e. the average number of ionizing collisions per unit distance travelled) then the product $\alpha l_{\rm a}$ must be equal to unity in the steady state. This statement indicates that as many charge carriers leave the avalanche region as are produced by impact ionization. The electrons produced travel through the "drift region" (of length $l_{\rm d}$) to the N^+ contact region on the right of the depletion layer; they do this at a *constant* drift velocity $u_{\rm s}$ (see fig. 4 of the preceding article). The transit time $\tau_{\rm d}$ is thus equal to $l_{\rm d}u_{\rm s}$ [3]. The drift region for the holes (on the left in fig. 4) and the avalanche region are assumed to be so thin that transit-time effects in them are negligible. The holes produced are thus "immediately" taken up in the P^+ contact region on the left of the depletion layer. The peak field in the breakdown region of silicon is about 360 kV/cm. (The effective average value of the field in the breakdown region is somewhat lower than the peak value, and is about 330 kV/cm.) For the "saturated" drift velocity $u_{\rm s}$ in silicon we shall take a value of 0.85×10^7 cm/s.

To investigate the high-frequency behaviour of the diode we imagine an a.c. voltage $v_{\rm a}(t)$ to be superimposed on the d.c. voltage $V_{0\rm a}$ across the avalanche region (*fig. 5a*):

$$v_{\rm a}(t) = V_{\rm a} \sin \omega t. \qquad (1)$$

Since the ionization coefficient α increases with the field, the product $\alpha l_{\rm a}$ is greater than unity during the half-cycles of positive $v_{\rm a}(t)$ and the number of charge carriers increases. Therefore the number of charge carriers, and hence the current, reaches a maximum at the end of the half-cycles of positive $v_{\rm a}(t)$. Similarly a current minimum occurs after a half-cycle of negative $v_{\rm a}(t)$. If the amplitude $V_{\rm a}$ is sufficient, current pulses arise (fig. 5*b*); these are always a quarter of a cycle behind $v_{\rm a}(t)$. If the amplitude $V_{\rm a}$ were to rise without limit, these pulses would be infinitely sharp. The average current of the pulses is the d.c. current which the current

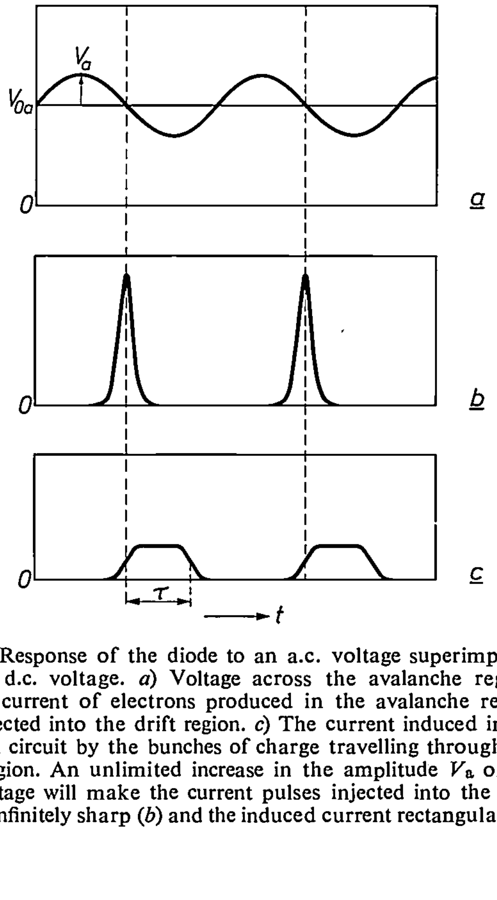


Fig. 5. Response of the diode to an a.c. voltage superimposed on the d.c. voltage. *a*) Voltage across the avalanche region. *b*) The current of electrons produced in the avalanche region and injected into the drift region. *c*) The current induced in the external circuit by the bunches of charge travelling through the drift region. An unlimited increase in the amplitude $V_{\rm a}$ of the a.c. voltage will make the current pulses injected into the drift region infinitely sharp (*b*) and the induced current rectangular (*c*).

source maintains constant at the value I_0. The fundamental (angular frequency ω), like all the harmonics, has an amplitude $2I_0$ in the limiting case of infinitely sharp current pulses. Since the fundamental is also 90° in phase behind $v_a(t)$, it takes the form:

$$i_{ca}(t) = -2I_0 \cos \omega t. \tag{2}$$

A bunch of electrons generated in the avalanche layer, and corresponding to one current pulse, travels through the drift region at the velocity u_s in the time τ_d. During its transit time this bunch of electrons, travelling at constant velocity, induces a constant current in the circuit connected to the diode. The successive bunches of charge thus give rise to a series of rectangular pulses of current in the external circuit (fig. 5*c*). Calculation of the fundamental of this induced current gives:

$$i_i(t) = -2I_0(\Phi_1 \cos \omega t + \Phi_2 \sin \omega t), \tag{3}$$

where

$$\Phi_1 = \frac{\sin \theta}{\theta}, \tag{4a}$$

$$\Phi_2 = \frac{\sin^2 \frac{1}{2}\theta}{\frac{1}{2}\theta} \tag{4b}$$

and

$$\theta = \omega\tau_d = \omega l_d/u_s. \tag{4c}$$

The quantity θ is the transit angle. An essential feature in the operation of the IMPATT diode is the phase difference between $i_i(t)$ and $i_{ca}(t)$, expressed in equation (3) by the term $\Phi_2 \sin \omega t$. This phase difference is a *transit-time effect*: Φ_2 only differs from 0 when θ, and hence τ_d, differs from 0.

The currents mentioned so far are conduction currents, carried by electrons and holes. In addition, capacitive currents occur both in the avalanche region and in the drift region.

The equivalent circuit that can be drawn on the basis of these data [4] is shown in *fig. 6*. The conduction currents $i_{ca}(t)$ and $i_i(t)$ (eqs. 2 and 3) are represented by current sources, the capacitive character of the avalanche region and drift region by capacitors C_a and C_d. The resistance R_s represents losses in the diode.

The two terms that form $i_i(t)$ in (3) are each represented by a current source. There are therefore three current sources in all; two of them ($i_{ca}(t)$ and the first term of $i_i(t)$) are 90° behind $v_a(t)$ in phase. This is expressed by the factor +j in the complex notation we shall use here. The third source (the second term of $i_i(t)$), which represents the transit-time effect, is in antiphase with $v_a(t)$.

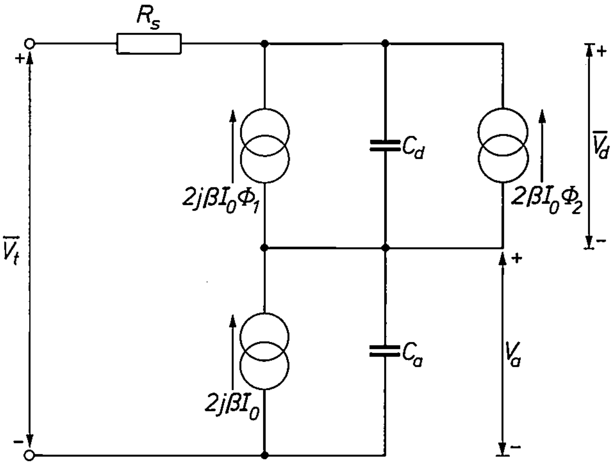


Fig. 6. Equivalent circuit of the IMPATT diode for signals of angular frequency ω. C_a and C_d are the capacitances of the avalanche and drift regions. R_s is the loss resistance of the diode. The complex amplitudes are given for the voltages and currents. V_a, $\overline{V}_d$, $\overline{V}_t$ are the amplitudes of the voltages across the avalanche region, the drift region and the complete diode. The current source $2j\beta I_0$ represents the first Fourier component $i_{ca}(t)$ (see eq. 2) of the series of current pulses in fig. 5*b*. The two other sources represent the two terms of $i_i(t)$ (see eq. 3), which are the in-phase and 90° out-of-phase components of the current of rectangular pulses in fig. 5*c*. These terms are dependent through Φ_1 and Φ_2 on the transit angle θ, i.e. on the frequency (see eq. 4c). The factor β is equal to 1 in the limiting case of infinitely large V_a; β for finite V_a is given in fig. 7.

(The complex exponential notation used here also expresses the phase relationships. The voltage polarities and current directions indicated are the positive directions for the calculation. The bars above V_d and V_t indicate that these amplitudes are complex, i.e. the a.c. voltages $v_d(t)$ and $v_t(t)$ over the drift region and the complete diode are not in phase with $v_a(t)$. The factor β is real. A factor j implies a 90° phase lead with respect to a real current in the direction indicated by the arrows.)

At finite values of the amplitude V_a the current pulses are not infinitely sharp. Although the phase relations remain the same, the conduction current in the avalanche region is smaller than that given by equation (2). This is allowed for in the equivalent circuit by the factor β, which is real and cannot exceed the value 1. Since the relation between the current in the drift region and the current in the avalanche region is linear, we can use the same factor β in the three current sources. *Fig. 7* shows β, obtained from a more detailed analysis, as a function of V_a [5]. The value v_1 plotted horizontally in fig. 7 is a normalized value of V_a. The relation between v_1 and V_a is:

$$v_1 = \alpha_c' V_a/\omega\tau_a. \tag{5}$$

In this expression α_c' is the derivative of the ionization coefficient α with respect to the field-strength at the operating point, and τ_a is the transit time in the ava-

[2] The diodes for these oscillators were made at these laboratories by Dr. D. de Nobel and H. G. Kock.

[3] In the preceding article this transit time in the drift region was indicated by τ.

[4] K. Mouthaan, Characterization of nonlinear interactions in avalanche transit-time oscillators, frequency multipliers, and frequency dividers, IEEE Trans. MTT-18, 853-862, 1970.

[5] Fig. 7 is equivalent to fig. 8 in the preceding article. The variable v_1 is equal to $\frac{1}{2}\xi V_a$ in the Appendix of that article.

lanche region. If V_{0a} is the d.c. voltage across the avalanche region, then in silicon α_c' can in practice be equated with $5/V_{0a}$.

The equivalent circuit of fig. 6 gives a good description of the high-frequency behaviour of the diode, provided that V_t, the amplitude of the r.f. voltage across the diode, is not too large. In general, rectifying effects start to cause considerable deviations, and the performance of the diode starts to deteriorate severely if V_t becomes greater than a particular fraction γ of the d.c. voltage V_0 across the diode. In practice the fraction γ is usually about 30%. In this article we shall assume that V_t is not larger than γV_0. A "maximum permissible voltage swing" of this type was noted in the preceding article (see page 337). The deviations that occur when V_t approaches the value γV_0 can be taken into account by using a somewhat smaller value for β than the value given by fig. 7.

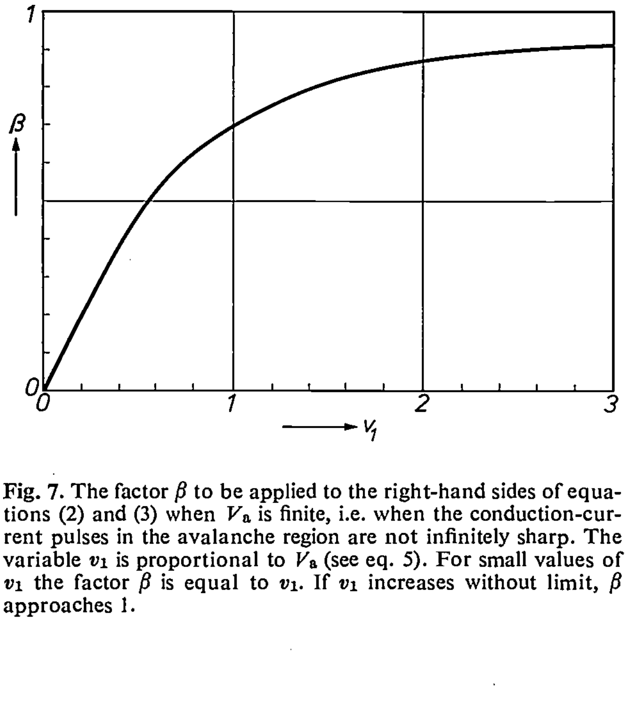


Fig. 7. The factor β to be applied to the right-hand sides of equations (2) and (3) when V_a is finite, i.e. when the conduction-current pulses in the avalanche region are not infinitely sharp. The variable v_1 is proportional to V_a (see eq. 5). For small values of v_1 the factor β is equal to v_1. If v_1 increases without limit, β approaches 1.

Characteristics of the IMPATT-diode oscillator

It will make it easier to discuss the characteristics of the IMPATT-diode oscillator if we simplify the equivalent circuit in fig. 6. We note that the amplitude of the current sources can never be greater than $2I_0$, while the capacitive currents are proportional to the frequency and independent of I_0. At a given value of I_0 and at sufficiently high frequencies, the capacitive currents will therefore predominate. We shall assume that this is the case. If R_s is also sufficiently small, the diode will behave approximately like a capacitive voltage divider, consisting of two capacitors C_a and C_d; in this case $v_a(t)$, $v_d(t)$ and $v_t(t)$ are approximately in phase, and the relation between the amplitudes is given by:

$$CV_t = C_aV_a = C_dV_d, \qquad (6)$$

where C is given by:

$$1/C = 1/C_a + 1/C_d. \qquad (7)$$

The two current sources $2j\beta I_0$ and $2j\beta I_0\Phi_1$, like the capacitive currents, are reactive (i.e. they are 90° out of phase with the voltage), and may therefore be neglected in comparison with the capacitive currents, which are assumed to be much greater. The simplified equivalent circuit obtained in this way is shown in *fig. 8.*

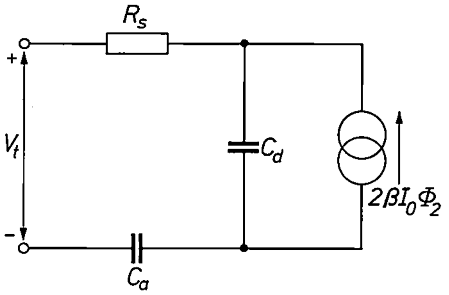


Fig. 8. Simplified equivalent circuit of the IMPATT diode, obtained by neglecting the reactive current sources in fig. 6. This is a good approximation if the frequency is sufficiently high.

The remaining current source, $2\beta I_0\Phi_2$, is in antiphase with the voltages, and therefore describes the active characteristics of the diode. For large values of V_a it is independent of $V_a (\beta \rightarrow 1)$. The function Φ_2, which represents the transit-time effect, has a maximum of 0.72 for $\theta = 0.74\,\pi$. This optimum value of θ determines the optimum length of the drift region at a given frequency f ($l_d = u_s\tau_d = u_s\theta/\omega$; $\omega = 2\pi f$). Since, with a triangular field profile as in fig. 4, the breakdown voltage V_b is approximately equal to $\frac{1}{2}E_0 l$, there is also an optimum value for V_b at a given ratio l/l_d. In this way we find that in diodes made from silicon ($E_0 \approx 360$ kV/cm, $u_s \approx 0.85 \times 10^7$ cm/s), in which l/l_d is equal to 1.2, the product fV_b must have a value of 600 to 700 VGHz. The function Φ_2, however, does not vary much with frequency near its maximum, so that a diode that gives optimum performance at a particular frequency will also perform well in a wide band on either side of that frequency. With a diode for 6 GHz, for example, the power that can be obtained at a given supply current will usually vary by no more than a factor of 2 in a band from 4,5 to 8 GHz.

An oscillator is obtained by terminating the diode with a suitable impedance. The equivalent circuit of an oscillator is shown in *fig. 9*. The diode is represented here by the circuit in fig. 8, and the terminating impedance Z_L is taken to be a resistance R_L and a reactance jX_L in series. It is assumed in the following that R_L is a useful load; stray resistances in the oscillator circuit are included in R_s.

The oscillation level

To find the tuning condition and oscillation level of the oscillator, we consider the part of the circuit to the right of PQ in fig. 9 to be a "source", which is terminated by an impedance Z_{PQ}, the part to the left of PQ. The source thus defined represents the drift region; the impedance Z_{PQ} represents the avalanche region, the losses and the external load.

In steady-state oscillation the current in Z_{PQ} must be equal but opposite to the current in the source. Since the capacitive currents predominate, the terminating reactance is determined by:

$$X_L - 1/\omega C_a = 1/\omega C_d$$

or (see eq. 7):

$$X_L = 1/\omega C. \qquad (8)$$

The frequency with a given diode can thus be adjusted by means of the terminating reactance X_L.

Next, the ratio of the reactive to the resistive components of the current in the termination must be equal to that in the source. This condition determines the oscillation level. This ratio is equal to the ratio of the imaginary part and the real part of the impedance. For the termination Z_{PQ} this is equal to the "loaded Q" of the drift region:

$$\frac{X_L - 1/\omega C_a}{R_L + R_s} = \frac{1}{\omega C_d(R_L + R_s)} \equiv Q. \qquad (9)$$

For the source this ratio is equal to $\omega C_d V_d / 2\beta I_0 \Phi_2$. With $C_d V_d = C V_t$ we thus find:

$$\omega C V_t = 2\beta Q I_0 \Phi_2. \qquad (10)$$

The oscillation level for large values of V_a ($\beta = 1$) is directly given by equation (10). For smaller values of V_a the oscillation level can be found by plotting both sides of equation (10) as functions of V_a (*fig. 10*). The left-hand side is proportional to V_a (because $CV_t = C_a V_a$), and the right-hand side (as a function of V_a) is proportional to β. The oscillator adjusts itself to the point of intersection, thus setting the level. Whether there is a point of intersection, and if so where it is located, depends on the proportionality factors, in particular on I_0.

We note that the assumption that the capacitive currents predominate is equivalent to the statement $Q \gg 1$.

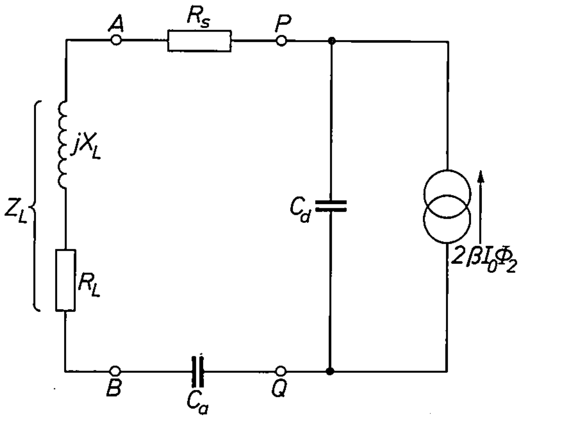


Fig. 9. Oscillator circuit obtained by connecting an impedance $Z_L = R_L + jX_L$ to the terminals (AB) of a diode circuit as in fig. 8. For calculating the oscillation level the part of the circuit to the right of PQ is regarded as a source, the other part as a load. R_L is a useful load; losses in the microwave circuit are allowed for in R_s.

Output power and efficiency

The quantity $\omega C V_t$ in equation (10) is the amplitude of the total r.f. current in the diode and in the load. The power developed in the load resistance R_L, i.e. the output power P_o of the oscillator, is thus given by:

$$P_o = \tfrac{1}{2}(2\beta Q I_0 \Phi_2)^2 R_L. \qquad (11)$$

To obtain a high output power at a given d.c. current I_0, we first have to optimize the length of the drift region, thus making the factor Φ_2 as large as possible; this point has already been mentioned. We also have to maximize the factor $Q^2 R_L$, which is proportional to $R_L/(R_L + R_s)^2$. This is done in the first place by minimizing the loss resistance R_s. We assume that R_s can be kept so small that $R_s \ll R_L$ for every useful load encountered in practice. Next, we should make R_L as small as possible. Now, as R_L becomes smaller (and hence Q greater) the amplitude V_t increases. We had stipulated, however, that V_t should not exceed a specific fraction γ of the d.c. voltage V_0. From equa-

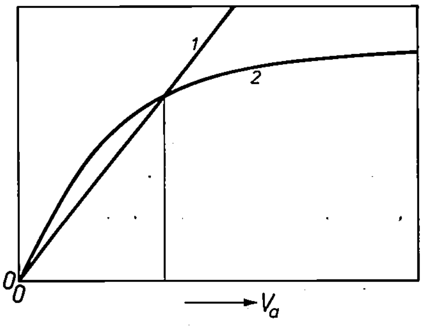


Fig. 10. The two sides of eq. (10) as a function of V_a; *1* is the left-hand side, *2* the right-hand side. The oscillation level is given by the point of intersection. As I_0 increases, the right-hand side increases in proportion. The oscillation level then rises as well.

tion (10) it then follows that Q should not be greater than a maximum value Q_m, given by:

$$Q_m = \frac{\gamma\omega CV_0}{2\beta I_0\Phi_2}. \tag{12}$$

Putting $Q = Q_m$ in equation (11) we then find an upper limit P_m for the output power which, because of the equality $R_L = (R_s + R_L) - R_s = 1/Q\omega C_d - R_s$, can be written in the form:

$$P_m = \gamma\beta(C/C_d)\Phi_2 I_0 V_0 - \tfrac{1}{2}(\gamma\omega CV_0)^2 R_s. \tag{13}$$

The second term is the power used up in the loss resistance R_s. If we neglect this loss, we find from (13) an upper limit η_m for the efficiency:

$$\eta_m = P_m/I_0V_0 = \gamma\beta(C/C_d)\Phi_2. \tag{14}$$

In *Table I* we present an estimate of η_m for silicon diodes with a triangular internal field profile, as in fig. 4. Appreciably higher efficiencies than the upper limit of 15% found here could only be expected if diodes could be made that worked well at voltage amplitudes substantially greater than 30% of the breakdown voltage.

The starting current

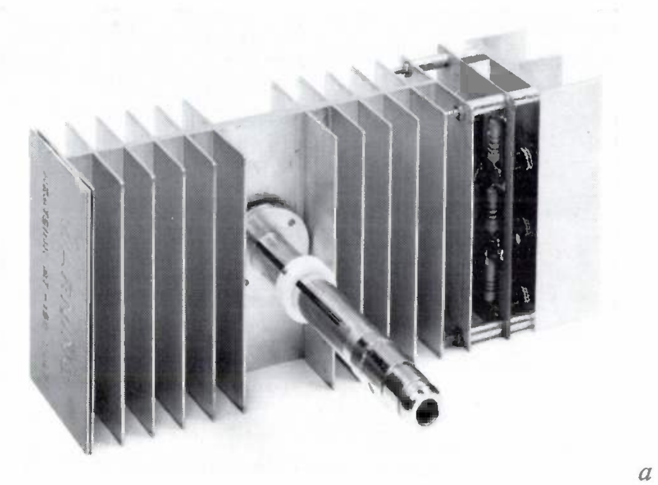
The circuit cannot oscillate unless the d.c. current I_0 exceeds a critical minimum value I_{start}. This starting current is the value of I_0 at which the curve in fig. 10 is tangential to the straight line at the origin, or in other words the value at which equation (10) is satisfied for small values of V_t. In this case β is equal to v_1. Combining this with equations (5), (6) and (10), we find that the starting current is given by:

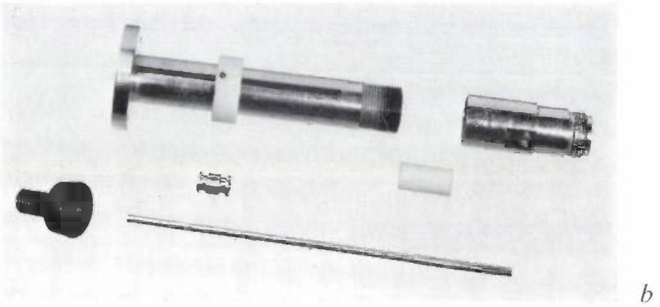
$$I_{start} = C_a\tau_a\omega^2/2\alpha_e' Q\Phi_2. \tag{15}$$

Table I. Estimate of an upper limit η_m for the efficiency, and of the maximum permissible current I_{0m}, expressed in the starting current I_{start}, for silicon diodes.

Values used	$\gamma = 0.3$ $\theta = 0.74\,\pi$ $\Phi_2 = 0.72$ $C/C_d = l_d/l = 0.85$ $V_{0a}/V_0 = 0.3$
Calculated $v_1 = 5\gamma CV_0/\theta C_d V_{0a}$	$v_1 = 1.8$
Read from fig. 7	$\beta = 0.85$
Calculated $\eta_m = \gamma\beta(C/C_d)\Phi_2$	$\eta_m = 0.15$
Calculated $I_{0m} =$ $= (5\gamma CV_0/\beta\theta C_d V_{0a})I_{start}$ (with $\beta = 0.7$)	$I_{0m} = 2.5\ I_{start}$

Explanatory note. It is assumed that the voltage amplitude has its maximum permissible value at 30% of the d.c. voltage ($\gamma = 0.3$) and that the diode is of optimum design ($\theta = 0.74\pi$, hence $\Phi_2 = 0.72$). The value for C/C_d ($= l_d/l$) of 0.85 is about the highest that can be expected from diodes with a triangular field profile. The value for V_{0a}/V_0 follows from fig. 4 with $l_d/l = 0.85$. The expression for v_1 follows from eq. (5) with the aid of the relations $\alpha_e' = 5/V_{0a}$, $\omega\tau_d = \theta$, $\tau_a/\tau_d = l_a/l_d = C_d/C_a$, $C_aV_a = CV_t$ and $V_t = \gamma V_0$. The value of β corresponds to the value found for v_1 from fig. 7. This was used for calculating the upper limit η_m of the efficiency. In calculating the permissible current, using eq. (16), with $\alpha_e' = 5/V_{0a}$, the deviations from the theory (see page 348) were taken into account by assuming a rather smaller value for β.





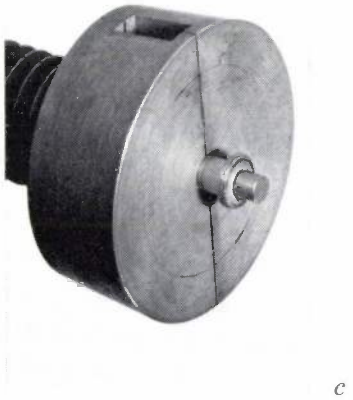


Fig. 11. Coaxial IMPATT-diode oscillator. *a*) The complete oscillator, *b*) components of the coaxial line with diode, *c*) enlarged heat sink with diode (on the left in *b*). The sliding sleeve in (*b*) is the quarter-wave transformer; its position determines the oscillation frequency. The heat sink consists of a slightly tapered collet that fits into the part with the fins. This provides good heat contact between the diode and the heat sink when the collet is clamped tight into the part with the fins. The current-stabilizing circuit can be seen on the right in (*a*).

The starting current (in the approximation with $Q \gg 1$) is thus inversely proportional to Q, and therefore directly proportional to $R_L + R_s$, the sum of the load resistance and the loss resistance. (Departing from the simplified equivalent circuit and using the complete circuit given in fig. 6, we find that the factor $Q\Phi_2$ in (15) should be replaced by $1 + Q\Phi_2$; the expression is then also valid for small Q.) The lowest possible starting current is the current at which the load resistance is equal to 0 (and Q thus equal to $1/\omega C_d R_s$). This no-load starting current, $I_{\text{start s}}$, is a direct measure of the losses, and consequently an important parameter of the diode and the associated oscillator circuit. In a circuit like that in fig. 1 the useful load resistance can be made equal to zero — and the no-load starting current thus measured — by substituting a short-circuit piston for the quarter-wave transformer. A weakly coupled probe is used for determining the start of the oscillation. It should be pointed out, however, that the loss resistance R_s measured in this way is not always equal to the value of R_s in a circuit oscillating at a higher level, since the losses in the diode often decrease when the d.c. current increases. This effect is related to the resistance of the "unswept region", i.e. of the non-depleted part of the N layer (see fig. 19b of the previous article) which is present at low currents but not at the currents for which the diode was designed.

The requirement that the a.c. voltage should not be greater than a fraction γ of the d.c. voltage is equivalent to saying that the d.c. current in the oscillator must not be greater than a certain factor times the starting current. Given $V_t < \gamma V_0$, $C_a\tau_a = C_d\tau_d$ and $\omega\tau_d = \theta$, we find from equation (10) and (15):

$$I_0/I_{\text{start}} < C\alpha_c'\gamma V_0/C_d\beta\theta. \qquad (16)$$

Table I gives an estimate of the maximum permissible d.c. current calculated from equation (16). In this estimate the deviations from theory mentioned earlier (page 348) have been taken into account by taking a rather smaller value for β than fig. 7 would indicate. The result is that I_0 must not be greater than about 2.5 times I_{start}. Or, the other way round, if we want to obtain a high output power for a given d.c. current I_0, we have to ensure that I_{start} is about 2.5 times smaller than I_0. The requirement mentioned earlier that R_s should be much less than R_L can be equated with the requirement that $I_{\text{start s}}$ should be much less than I_{start}.

Two experimental oscillators

The coaxial oscillator of *fig. 11*, already mentioned in the introduction, will now be discussed in somewhat more detail, and another type of oscillator, the microstrip oscillator of *fig. 12*, will also be discussed. Both were designed for a frequency of 5 GHz and based on the same type of diode. Some data and results of measurements are presented in *Table II*.

The breakdown voltage of the diode used here is 120 V. The diode is therefore suitable for the design frequency of 5 GHz: the product fV_b has a value of 600 V GHz, which is near the optimum value. In Table II a value of 0.75 has been assigned to C/C_d;

Table II. Data for the coaxial oscillator in fig. 11 and the microstrip oscillator in fig. 12. Both oscillators were designed for a frequency of 5 GHz and based on the same type of diode.

Diode	
Breakdown voltage $V_b = 120$ V Capacitance at breakdown $C = 0.8$ pF	} measured
$C/C_d = 0.75$	assumed
Coaxial oscillator	
Design values: $f = 5$ GHz, $X_L = 1/\omega C = 40\ \Omega$	
Measured no-load starting current $I_{\text{start s}} = 5$ mA	
Design value for maximum d.c. current $I_{0m} = 100$ mA	
Design value for $I_{\text{start}} = I_{0m}/2.5 = 40$ mA ($I_{\text{start}} \gg I_{\text{start s}}$)	
Value of R_L at which $I_{\text{start}} = 40$ mA, determined experimentally: $R_L = 5\ \Omega$.	
Calculated $Q = 1/\omega C_d R_L = 6$	
at $I_0 = 100$ mA: measured $P_o = 1.1$ W	
calculated $P_o = 1.5$ W	
calculated $V_t = 30$ V	
Microstrip oscillator	
Design values	$f = 5$ GHz, $X_L = 1/\omega C = 40\ \Omega$
Unloaded	$I_{\text{start}} = 20$ mA found, by comparative measurement on coaxial oscillator, $R_s = 2\ \Omega$
Loaded	$I_{\text{start}} = 40$ mA (by adjusting coupling gap S) $R_L + R_s = 5\ \Omega$ (found by comparison with coaxial oscillator) $R_L = 3\ \Omega$ at $I_0 = 100$ mA: measured $P_o = 0.75$ W calculated, by comparison with coaxial oscillator, $P_o = \frac{3}{5} \times 1.1 = 0.66$ W

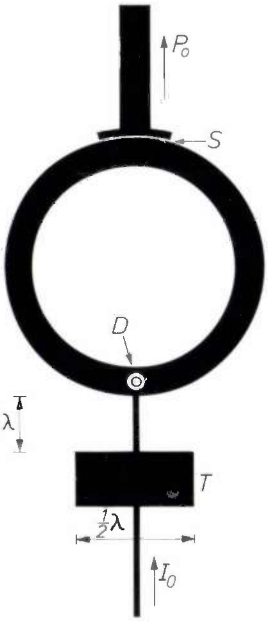


Fig. 12. Oscillator circuit in microstrip line. The dielectric sheet carrying the microstrip [6] and the varactor package with the diode D are mounted on a single heat sink; the varactor package projects through a hole in the dielectric. The d.c. supply (below) is decoupled to r.f. energy by the symmetrical quarter-wave tee junction T. The microwave power P_o is coupled to the output port by the gap S.

this value is used for estimating the coaxial-oscillator Q defined in eq. (9).

In a coaxial oscillator of the type shown in fig. 1 the load reactance X_L is determined by the spacing between the quarter-wave transformer and the diode, and the load resistance R_L is determined by the diameter of the inner conductor of the transformer. The complete coaxial inner conductor can be made in one piece, but in the circuit shown in fig. 11 the transformer is in the form of a sliding sleeve, which can be used to tune the oscillator.

As can be seen in fig. 11, the cooling fins account for most of the bulk of the oscillator. They are necessary for the rapid removal of the heat generated in the diode, which would become far too hot with no cooling; the heat developed is as much as 12 W at the voltage and current values of 120 V and 100 mA quoted in Table II. The cooling fins are designed so that the heat sink in which the diode is mounted rises no more than 15° above the temperature of the ambient air at a dissipated power of 15 W. To ensure good heat contact between diode and heat sink, the varactor package is clamped to it firmly.

From eq. (8) it follows that X_L must be equal to 40 Ω if a frequency of 5 GHz is to be obtained with the diode capacitance of 0.8 pF. The no-load starting current $I_{\text{start s}}$ is measured in the way indicated above, i.e. by substituting a short-circuit piston for the quarter-wave transformer and monitoring the start of oscillation by means of a weakly coupled probe. In this way a value of 5 mA was found for $I_{\text{start s}}$. To keep the heat generated within acceptable limits the maximum d.c. current was set to 100 mA. If the voltage excursion at this value is to be kept just inside the maximum permissible value, the starting current must be 100/2.5, i.e. 40 mA. We then have $I_{\text{start s}} \ll I_{\text{start}}$, and hence $R_s \ll R_L$, so that only a small part of the microwave power generated is dissipated in the loss resistance. By trying out quarter-wave transformers of different diameters, it was found that the starting current had the required value of 40 mA when R_L was equal to 5 Ω. This value of R_L — at which the maximum permissible voltage excursion is obtained for $I_0 = 100$ mA — could also be calculated from eq. (12) and eq. (9). The procedure described above for determining R_L is however more direct and therefore more attractive. The table also gives the Q associated with $R_L = 5$ Ω.

At $I_0 = 100$ mA the factor β should now theoretically be about 0.85 (see Table I). From eq. (10) we then find a voltage amplitude V_t of about 30 V, i.e. 25% of the breakdown voltage. This is in fact practically equal to the maximum permissible voltage amplitude, as appears from the saturation of the power for any further increase in the d.c. current. The output power at 100 mA is 1.1 W (see fig. 3). Calculating the output power from eq. (11) with $\beta = 0.85$, we find 1.5 W. The difference gives some idea of the deviations from the theory for $V_t = \gamma V_0$. With a value of 0.73 for β the measured and calculated output power would be in agreement.

If we assume that equation (15) is valid, i.e. that the starting current is proportional to the resistance, then the loss resistance R_s follows from the measured no-load starting current. With $I_{\text{start s}} = 5$ mA, and $I_{\text{start}} = 40$ mA at $R_L = 5$ Ω, we find $R_s = (5/40) \times \times 5 \approx 0.6$ Ω. Since the coaxial circuit is practically lossless, this loss resistance must be attributed almost entirely to the diode. Other direct measurements of the loss resistance have shown that at $I_0 = 40$ mA R_s is less than 0.2 Ω. This illustrates the point made earlier that R_s often decreases as the bias current increases.

In the second oscillator to be discussed the diode is connected to a microstrip circuit, seen from above in fig. 12. The circuit consists of a thin layer of metal on a dielectric substrate [6]. The varactor package containing the diode D projects through a hole in the dielectric and is fixed in a metal heat sink, which also supports the dielectric. The actual oscillator consists of the diode and the annular part of the circuit. Compared with other shapes, a ring gives low radiation losses. The d.c. is supplied from below and the microwave power is extracted at the top. The d.c. supply is decoupled from the microwave circuit by the symmetrical quarter-wave stub T. The oscillator is coupled to the output port via the gap S, which also isolates the output circuit from the d.c. supply. Another advantage of this arrangement is that varying the coupling by varying the gap width has little effect on the frequency. To obtain reproducible results, however, the circuit should be screened. The entire circuit is mounted in a metal box, fitted with connectors for the d.c. supply and the output power.

In designing the oscillator the output coupling is not initially taken into account. For reasons of symmetry the point diametrically opposite the diode is then an open-circuit to r.f. signals ($i = 0$). The left- and right-hand halves of the ring form two pieces of microstrip line in parallel, which together form the termination of the diode. After the characteristic impedance of the microstrip line has been selected (in practice this means choosing the width of the strip for a given dielectric material and thickness) the line lengths that will give a frequency of 5 GHz are calculated; these are the line lengths for which the reactance X_L of the termination is 40 Ω (for a diode capacitance of 0.8 pF, as quoted in Table II). In the practical version an integral number of half wavelengths are added to these lengths. This gave a ring circuit with an outside diameter of about 4 cm for the particular dielectric used.

The output strip was also omitted for the initial measurements on the oscillator. The starting current was found to be 20 mA. Comparative measurements on the coaxial oscillator showed that this corresponds to a total load resistance of 2 Ω, which in the present case is a pure loss resistance, due mainly to the circuit. These relatively high losses are attributed to the use of microstrip line.

The oscillator was then coupled to the useful load via the output strip and the gap was adjusted to give a starting current of 40 mA. The gap width was determined empirically, since the theoretical treatment of this problem is extremely complicated. The measurements on the coaxial oscillator indicated that the total load resistance was then 5 Ω. With the measured loss resistance of 2 Ω, about 40% of the power generated by the diode is lost in the microstrip circuit, leaving about 60% as useful power. With this coupling the "maximum permissible d.c. current" is again 2.5×40 mA, i.e. 100 mA. Considering that the diode in the coaxial oscillator develops a power of 1.1 W at 100 mA, we therefore expect the microstrip oscillator to produce a useful output of about 0.66 W. The measured output was 0.75 W. This result agrees well with the expected value, since with the unloaded oscillator the radiated power at the open end was treated as a loss, whereas with the loaded oscillator part of it appears as useful power.

Stability

The frequency and amplitude of the output from an oscillator should be constant (except for tuning and modulation, of course). There should be no variations, either spontaneous or as a result of changes in the conditions. In an IMPATT-diode oscillator, however, low-frequency instabilities easily occur if no special attention is paid to the d.c. current supply. The frequency is also fairly sensitive to external load and temperature variations. These two points will now be considered. The spontaneous fluctuations that can be considered as "noise" will be discussed in the next section.

D.C. current supply

Unsatisfactory impedance characteristics of the d.c. current supply circuit can cause instabilities in the oscillator, since the diode can behave as a device with a *negative* differential resistance at low frequencies as well as at high frequencies. This is related to the way in which the ionization coefficient α depends on the field E (*fig. 13*). The steady state at breakdown is determined by the condition $\langle\alpha\rangle\, l_a = 1$, where $\langle\alpha\rangle$ is the time average of α. Near the point (α_0, E_0) on the α,E curve where this state is reached at zero r.f. amplitude, the second derivative of α with respect to E is positive. If an a.c. field is now superimposed on a constant field E_0' in the vicinity of E_0, then $\langle\alpha\rangle$ is greater than $\alpha(E_0')$, and to satisfy the relation $\langle\alpha\rangle\, l_a = 1$ the value of E_0' must be less than E_0. This gives a decrease in the d.c. voltage across the diode with increasing amplitude of the r.f. voltage. Since the r.f. amplitude in an oscillator generally increases with rising d.c. current, this effect gives rise to a negative differential low-frequency resistance which is "induced" by the r.f. voltage. We refer to this resistance as R_{ind}.

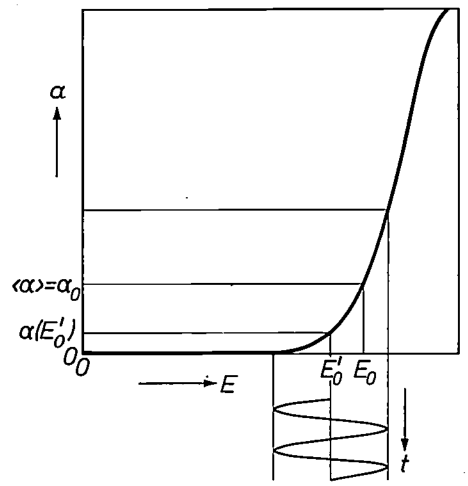


Fig. 13. The ionization coefficient α as a function of the field E. At high r.f. amplitudes there is a rectifying effect: the constant field E_0' on which the r.f. field is superimposed must be smaller than the breakdown field E_0 to ensure that the relation $\langle\alpha\rangle l_a = 1$ remains satisfied.

This effect is opposed by two other d.c. current effects, which are both equivalent to a *positive* differential resistance. The first is a thermal effect. A variation in the d.c. current causes a variation in the heat generated in the diode, and hence a variation in temperature. Now, in general, the breakdown voltage increases with temperature. In silicon diodes with a triangular internal field profile as in fig. 4, and at diode temperatures obtained in practice, this increase in V_b is about 0.12% per degree. Consequently the voltage rises when the d.c. current rises. The corresponding differential resistance due to this thermal effect is referred to as R_{th}. Since the temperature can only follow the heat dissipation variations if they are sufficiently slow, this effect is only found at frequencies below a value known as the "thermal cut-off frequency".

The second positive-resistance effect is that of the *space charge* of the current-carrying electrons in the

[6] More information about microstrip line appears in the article "Microwave integrated circuits" by J. H. C. van Heuven and A. G. van Nie, this issue, page 292.

drift region. This charge partly compensates the space charge of the donors, thus making the field gradient less steep and consequently, since the breakdown field is fixed, causing an increase in the voltage across the drift region (*fig. 14*). If the current rises, the electron space charge also increases; and with it the voltage. The corresponding resistance is referred to as $R_{\rm sc}$.

At low frequencies the diode can therefore be represented by the diagram in *fig. 15*. The inertia of the thermal effect is represented by the capacitance $C_{\rm th}$, which short-circuits $R_{\rm th}$ at high frequencies.

Further analysis shows that for large signals ($\beta \rightarrow 1$) the quantity $R_{\rm ind}$ is about $(V_{\rm t}/V_0)Q$ times greater in absolute value than $R_{\rm sc}$. For $V_{\rm t}/V_0 \approx 0.3$ the resistance $R_{\rm ind}$ is therefore dominant when Q is greater than 3. Above the thermal cut-off frequency the net resistance is then negative. If Q is large enough, this will also be the case below the thermal cut-off frequency.

The instabilities caused by the induced negative resistance can be countered by including in the supply circuit a resistance of about 100 Ω in series with the diode. (This resistance must be bypassed by a capacitor in the r.f. circuit.) Insufficient stabilization can lead to amplification of the low-frequency noise caused by avalanche multiplication, and even low-frequency oscillations (e.g. in the MHz region). The result can be a serious parasitic modulation of the r.f. signal [7].

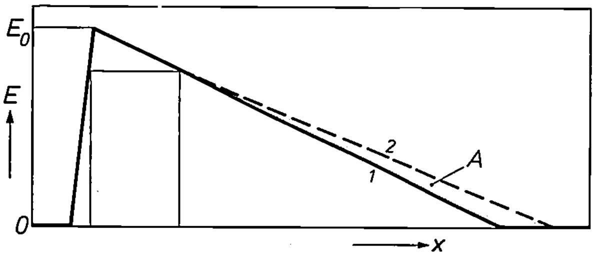


Fig. 14. The field profile in the diode, *1* for low current, *2* for higher current. In case *2* the space charge of the donors is partly compensated by that of the electrons in the drift region. The field gradient is thus smaller, and the voltage is greater than in case *1* by an amount equal to the area *A*. At low frequencies this effect corresponds to that of a positive resistance $R_{\rm sc}$.

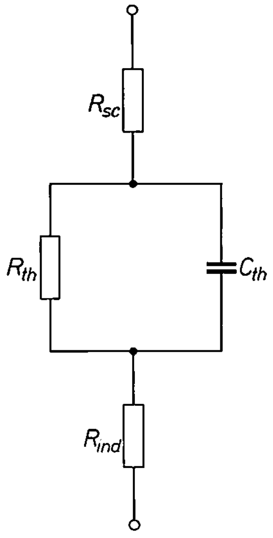


Fig. 15. Low-frequency equivalent circuit for the diode. If the negative resistance $R_{\rm ind}$, induced by rectification of the r.f. signal, is dominant, the circuit can become unstable at low frequencies. The resistances $R_{\rm sc}$ and $R_{\rm th}$ represent the space-charge effect of fig. 14 and a thermal effect, and are positive. The inertia of the thermal effect is accounted for by the capacitance $C_{\rm th}$, which short-circuits $R_{\rm th}$ above the "thermal cut-off frequency".

Frequency stability; tuning

Load variations

Variations in the load lead to variations in both the output power and the frequency of the oscillator. In a circuit like the one shown in fig. 1 the oscillator should ideally be terminated with a matched load, i.e. a load whose complex amplitude-reflection coefficient ϱ is zero. If ϱ is not equal to zero, power and frequency will both differ from the ideal values. The maximum relative deviations of power and frequency at a given ϱ are given by:

$$(\varDelta P)_{\rm max}/P_0 = 2|\varrho| \tag{17}$$

and

$$(\varDelta\omega)_{\rm max}/\omega = 2|\varrho|/Q_{\rm L}. \tag{18}$$

It depends on the argument of ϱ whether the deviations have the maximum value or are smaller. Equations (17) and (18) are valid if $|\varrho| \ll 1$, $Q_{\rm L} \gg 1$ and $1/\omega C \lessgtr Z_0$. ($Q_{\rm L} = 1/\omega C R_{\rm L}$ and Z_0 is the characteristic impedance of the coaxial line.)

The relative frequency variations are a factor of $Q_{\rm L}$ smaller than the relative power variations, and decrease with increasing $Q_{\rm L}$. A simple oscillator of the type in fig. 1 has a $Q_{\rm L}$ of 10 to a maximum of 100. Greater frequency stability can be obtained by coupling the oscillator to a high-Q resonant circuit. In such a case the $Q_{\rm L}$ in (18) must be replaced by an effective Q for the whole circuit.

Conversely the oscillator can be *tuned* by varying the load. This principle is applied in the *electronically tuned* oscillator of *fig. 16*. This contains a second diode, and the voltage across this diode is kept so low that no breakdown occurs. Variation of this voltage changes the capacitance of the diode and hence the load on the oscillator. The oscillator can be tuned in this way over more than 150 MHz near 9.4 GHz, with no more than 10% change in the power output (*fig. 17*).

Temperature variations

To a first approximation the frequency is determined by eq. (8). When the temperature is varied the frequency therefore changes if $X_{\rm L}$ and C are temperature-dependent. It follows from eq. (8) that:

$$\frac{1}{\omega}\frac{{\rm d}\omega}{{\rm d}T} = -\frac{1}{C}\frac{{\rm d}C}{{\rm d}T} - \frac{1}{X_{\rm L}}\frac{{\rm d}X_{\rm L}}{{\rm d}T}. \tag{19}$$

The variation in the tuning reactance $X_{\rm L}$ with temperature is a question of the expansion of the material from which the microwave circuit is made. In a circuit like that shown in fig. 1, made from brass or aluminium, the second term in (19) gives a contribution of about -2×10^{-5} K^{-1} when the shortest possible distance is chosen between diode and transformer (approximately one-eighth of a wavelength); the frequency variation is greater if this distance is made a whole number of half-wavelengths greater.

The capacitance variation depends to a great extent on whether or not the depletion layer at breakdown extends to the substrate (the N^+ layer in fig. 4). If it does, the capacitance is virtually independent of temperature. If it does not, the thickness of the depletion layer, and hence the capacitance C, varies with the breakdown voltage V_b. In diodes with a triangular internal field profile the relative variation of C is then half that of V_b, and in the other direction. With the increase for V_b of 0.12% per degree mentioned earlier, we thus find a contribution of $+6\times10^{-4}$ K^{-1} for the first term in eq. (19). To reduce the temperature-dependence of the frequency diodes would thus have to be made in which the depletion layer extended to the substrate. However, this is only one of the various considerations in choosing the thickness of the epitaxial layer (see also the preceding article, page 340).

An important source of temperature sensitivity of the frequency does not appear from eq. (19). This can be seen by going back from the diagram of fig. 8 to that of fig. 6. The reactive current sources in fig. 6, which were later neglected, contribute to the reactive component of the total current, and thus affect the frequency. One of these sources, the one corresponding to the drift region, contains the factor $(\sin\theta)/\theta$. Near the optimum value of θ this factor varies strongly with θ and hence with u_s (see eq. 4c). It can be shown that this makes a contribution to $(1/\omega)\mathrm{d}\omega/\mathrm{d}T$ of $+(1/Q_L\Phi_2)(1/u_s)\mathrm{d}u_s/\mathrm{d}T$. Even though this is only a variation of a correction (as is evident from the presence of the factor $1/Q_L$), the effect nevertheless makes a significant contribution. Measurements show that u_s varies considerably with temperature: a value for $(1/u_s)\mathrm{d}u_s/\mathrm{d}T$ of -10^{-3} K^{-1} has been reported [8]. With a Q_L of 10 to 100 for a circuit like that of fig. 1 we thus arrive at a contribution of -10^{-4} K^{-1} to -10^{-5} K^{-1}. Further reduction in temperature sensitivity can be obtained by increasing the effective value of Q_L by coupling to a high-Q resonant circuit.

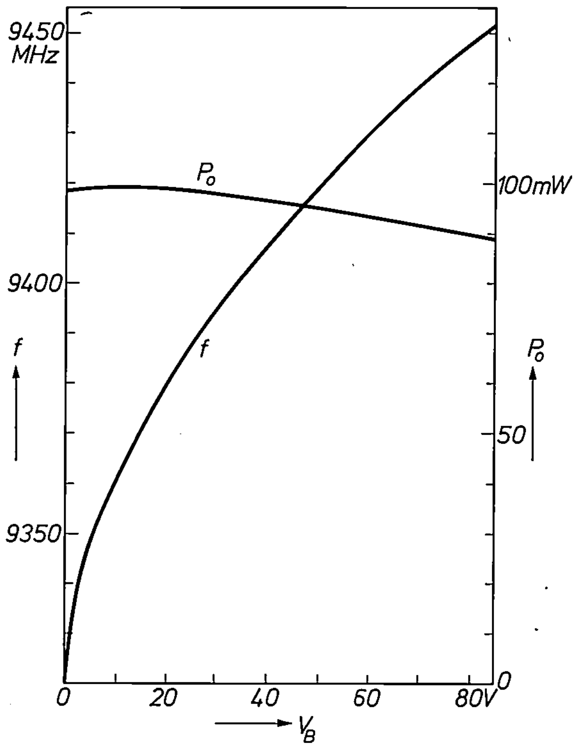


Fig. 17. The frequency and the output power P_o of the electronically tuned oscillator of fig. 16 as a function of the control voltage V_B. In this measurement the d.c. supply I_0 (see fig. 16) was kept at 70 mA. The dissipation in the oscillator diode A is about one-third of the maximum permissible dissipation. (The diode, operated here at about 9.4 GHz, had a breakdown voltage of 88 V and was thus optimized for about 7.5 GHz — illustrating the wideband nature of the diodes, mentioned on page 348.)

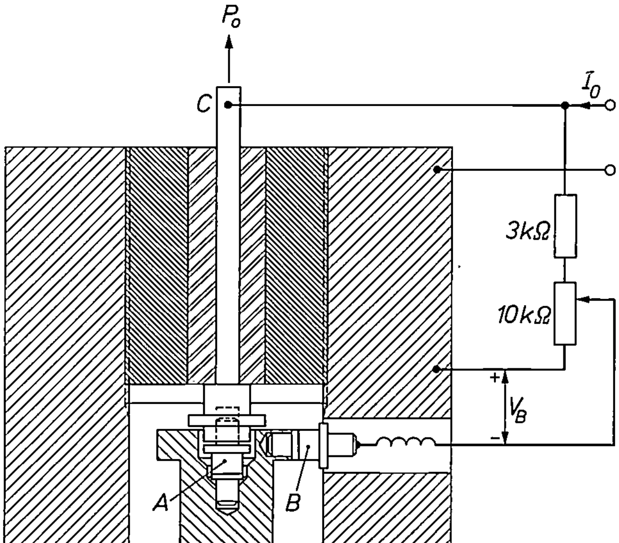


Fig. 16. An electronically tuned oscillator c rcuit. A oscillator diode. The auxiliary diode B is reverse-biased by a voltage V_B below the breakdown value. The oscillator is tuned by varying V_B. This has the effect of varying the capacitance of B, and consequently the terminating reactance of A and therefore the frequency. The oscillator can be connected to a coaxial line or mounted on the broad side of a rectangular waveguide; the inner conductor C then functions as an aerial in the waveguide.

[7] The effect of the rectification and of the induced negative resistance on the noise characteristics of the oscillator are analysed in more detail in: K. Mouthaan and H. P. M. Rijpert, Nonlinearity and noise in the avalanche transit-time oscillator, Philips Res. Repts. 26, 391-413, 1971 (No. 5).

[8] C. Y. Duh and J. L. Moll, Electron drift velocity in avalanching silicon diodes, IEEE Trans. ED-14, 46-49, 1967.

Noise

The usefulness of an oscillator depends to a very great extent on its noise characteristics. In an IMPATT-diode oscillator operating at about the maximum permissible amplitude these characteristics are determined by a complicated interaction between the diode, the supply circuit and the microwave circuit [7]. This will not be dealt with here, but the practical significance of noise will be indicated briefly, and a few experimental results for our coaxial oscillator will be quoted.

The main applications of the oscillator will be in telecommunications, for example in microwave radio links. Here the oscillator has to generate a carrier (of frequency f_0) on which information can be superimposed by amplitude or frequency modulation. If the modulation is effected by frequencies in a band B about a central modulation frequency f_m ($B \ll f_m \ll f_0$), the spectrum of the modulated carrier — neglecting noise — will consist of a principal component at the frequency f_0, and of two components in two sidebands of bandwidth B at the frequencies $f_0 + f_m$ and $f_0 - f_m$. These two components contain the information, and we are therefore concerned with the noise in the two sidebands. In telephony, for instance, a widely used value for the bandwidth B is 3 kHz.

In amplitude modulation it is *amplitude noise* (random, spontaneous fluctuations in amplitude) that is of interest, and in particular the ratio P_n/P_c. Here P_n is the amplitude-noise power in the two sidebands together, and P_c is the power in the carrier. With the coaxial oscillator discussed here, loaded to give a Q of about 7, a noise power of -115 dB with respect to the carrier power was found in sidebands of 3 kHz bandwidth. Given an oscillator power of 1 W the noise power in 3 kHz sidebands is therefore 10^{-12} to 10^{-11} W, allowing AM signals with a modulation level of 100 pW to be handled. For modulation frequencies f_m in the range from a few kHz to several MHz, the noise level found is not greatly dependent on f_m. This implies that an r.m.s. amplitude assigned to this modulation would be virtually independent of f_m; in other words, the noise is virtually "white".

In frequency modulation it is *frequency noise* that is of interest. The level of this noise is usually expressed in an r.m.s. value Δf_{rms} ($= \sqrt{\langle \Delta f^2 \rangle}$) of the instantaneous frequency deviations Δf. This value is $1/\sqrt{2}$ times the amplitude of a sinusoidal frequency modulation giving the same power (P_n) as the frequency noise in sidebands of bandwidth B. From frequency-modulation theory the relation between Δf_{rms} and P_n is given by:

$$(\Delta f_{rms}/f_m)^2 = P_n/P_c.$$

The noise power of frequency noise, unlike that of amplitude noise, decreases with increasing f_m if Δf_{rms} is constant, i.e. if the noise is "white". With the coaxial oscillator discussed here, it was found that Δf_{rms} was between 400 and 800 Hz for a Q of 6 to 8 and 3 kHz sidebands, and that this value was practically independent of f_m from very low values of f_m to a few MHz. This means that the oscillator would be able to handle FM signals with modulation amplitudes of a few kHz.

Improving the oscillator performance by tuning to higher harmonics

The conduction current through the diode is rich in higher harmonics if the a.c. component of the voltage across the avalanche layer is not too small (see fig. 5). It is possible to turn this feature to advantage [9]. In particular the performance of the oscillator can be improved by tuning the circuit to one of the higher harmonics. We shall shortly discuss some of the essential features of this technique. Let us confine ourselves to a circuit which, in addition to being tuned to a particular fundamental frequency, is also tuned to the second harmonic of that frequency. This case is relatively easy to realize in practice.

First we shall recapitulate the situation where the oscillator is tuned to only one frequency. The purely sinusoidal a.c. voltage across the diode then gives rise to a conduction current in the avalanche region. The conduction current is rich in higher harmonics, and so too, therefore, is the total current, i.e. the sum of conduction current and capacitive current. The external load also determines the relation between current and voltage. Its impedance is tuned to the correct value at the required frequency for oscillations to occur. At all higher harmonics, however, the impedance must be very low, because only then can the a.c. voltage be sinusoidal with a current rich in higher harmonics.

We now change the external circuit in such a way that the oscillator circuit is also tuned to the second harmonic, while its impedance remains unchanged for the fundamental and all other harmonics. To determine the resultant waveform of the voltage we must know the relationship between the conduction current through the avalanche region and the voltage across it, for the case where the voltage is no longer purely sinusoidal. If the voltage across the avalanche region differs by a relatively small amount $v_a(t)$ from the average value — the value at the operating point — the conduction current $i_a(t)$ through the avalanche region is given by:

$$i_a(t) = i_a(0) \exp \{k \int_0^t v_a(\tau) d\tau\}, \qquad (20)$$

where $k = 2\alpha_c'/\tau_a$ [10]. This expression reveals the exponential nature of the current, which is inherent

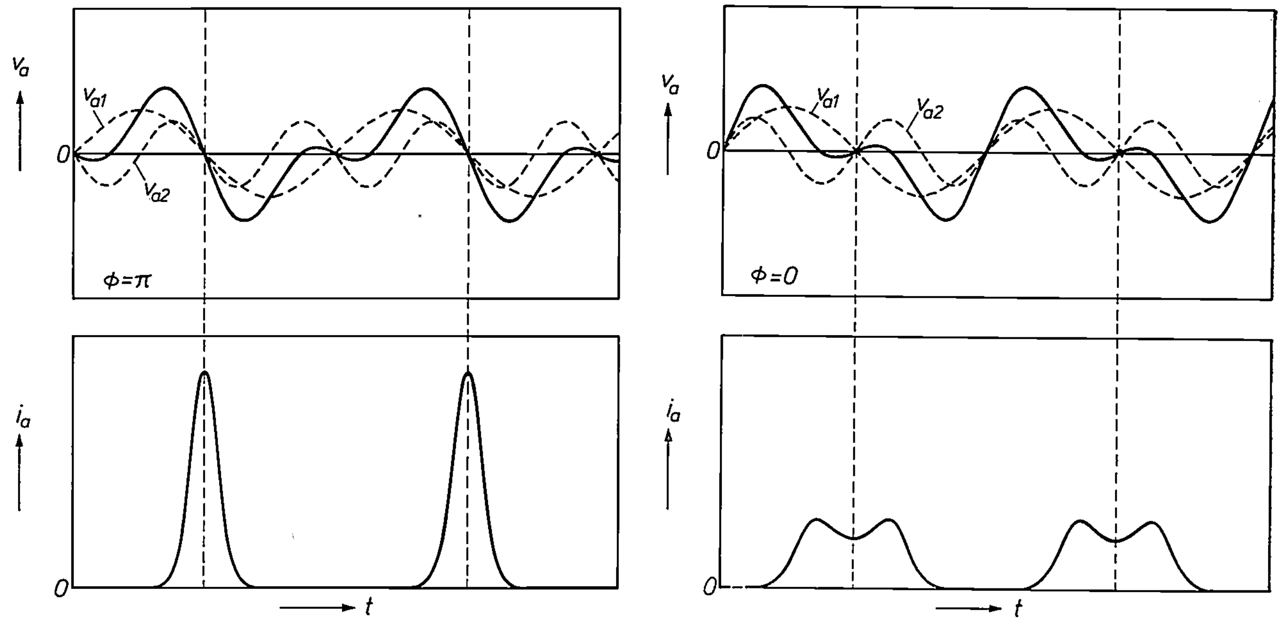


Fig. 18. The conduction current $i_a(t)$ from the avalanche region (below) as given by eq. (20) for two voltage waveforms. The variation $v_a(t)$ of the voltage across the avalanche region (the solid curve in the upper figure) is a superposition of the fundamental mode (v_{a1}) and the second harmonic (v_{a2}), as given by eq. (21). The two cases differ in phase ϕ in (21). The voltage amplitudes V_1 and V_2 and the average value of the current $i_a(t)$ are the same in both cases. The phase difference between the fundamental Fourier components (first harmonics) of current and voltage, ψ_1, and the phase difference between the second-harmonic Fourier components $\psi_2 - \phi$, are in both cases independent of V_1 and V_2. At $\phi = \pi$ the current pulses are sharper than at $\phi = 0$.

in avalanche multiplication. It can also be seen that the current increases as long as v_a is positive, and decreases when v_a is negative. One conclusion in particular that can be drawn from eq. (20) is of interest here: if the voltage variation is an odd function of time, $v_a(-t) = -v_a(t)$, then the current is an even function: $i_a(-t) = i_a(t)$. This has a direct bearing on the possible voltage waveform when the oscillator is tuned to a higher harmonic.

Suppose that the voltage variation and the current are given by:

$$v_a(t) = V_1 \sin \omega t + V_2 \sin (2\omega t + \phi),$$
$$i_a(t) = I_0 + I_1 \cos (\omega t + \psi_1) + I_2 \cos (2\omega t + \psi_2) + \dots \quad (21)$$

For oscillation to occur the phase relations between the Fourier components of $v_a(t)$ and $i_a(t)$ that follow from eq. (20), the "internal phase relations", must correspond to the "external phase relations", i.e. those determined by the load. Before the circuit is tuned to the second harmonic, V_2 is equal to zero. In this case $v_a(t)$ is odd, so that $i_a(t)$ is even, and thus contains only cosine terms, which means that ψ_1 (like ψ_2) is equal to zero. If now, as a result of tuning to 2ω, a term $V_2 \sin (2\omega t + \phi)$ of arbitrary phase ϕ were to be added to $v_a(t)$, then $v_a(t)$ would no longer be an odd function, and from eq. (20) ψ_1 would in general be dependent on V_1 and V_2. On the other hand, ψ_1 must remain equal to zero, since the external impedance for the fundamental frequency has not changed.

This difficulty is not encountered when $\phi = 0$ or $\phi = \pi$, since in these two cases $v_a(t)$ remains an odd function. Stable simultaneous oscillations at the frequencies ω and 2ω can therefore occur when ϕ has the values 0 or π, and it can be proved that these are in fact the only possible values [9]. These two cases are illustrated in *fig. 18*.

The improvement we are concerned with is obtained at $\phi = \pi$. In fig. 18 it can be seen that in this case the intervals during which the avalanche of charge carriers is built up and then broken down — i.e. the intervals in which v_a has successively a high positive and high negative value — are shorter than when the second harmonic is not present. Consequently the current pulse is sharper. For a given voltage amplitude of the fundamental, this implies a greater value of β than fig. 7 would indicate, and therefore, for a given d.c. current, a greater current amplitude and hence greater

[9] K. Mouthaan, Nonlinear analysis of the avalanche transit-time oscillator, IEEE Trans. ED-16, 935-945, 1969; Two-frequency operation of the avalanche transit-time oscillator, Proc. IEEE 58, 510-512, 1970; Nonlinear characteristics and two-frequency operation of the avalanche transit-time oscillator, Philips Res. Repts. 25, 33-67, 1970. See also note [4].

[10] Equation (20) follows directly from equation (16) in the preceding article on putting $j_s = 0$, $\alpha l_a = \alpha_c l_a + \alpha_c' v_a$, $\alpha_c l_a = 1$, $l_a = u_s \tau_a$, and making $i_a(t)$ proportional to j_c.

power at the fundamental frequency. In the case $\phi = 0$ the current pulse is flatter and the power at the fundamental frequency lower.

To calculate the voltage waveform that occurs at a particular supply current, it is necessary to set up an equivalent circuit like the one in fig. 9 for both ω and 2ω (for an accurate analysis diagrams based on fig. 6 should be used). The quantities X_{L1}, R_{L1}, ... in the ω diagram are different from X_{L2}, R_{L2}, ... in the 2ω diagram. In particular we must have (see eq. 8):

$$X_{L1} = 1/\omega C,$$
$$X_{L2} = 1/2\omega C.$$

The diode is optimized to give the maximum value at the fundamental frequency for the factor Φ_2 in the current source in fig. 9, so that $\Phi_{21} > \Phi_{22}$. For each of the frequencies the circuit has a separate Q (see eq. 9).

The current-peaking effect demonstrated in fig. 18 is now taken into account by postulating that β_1 no longer depends on V_1 alone, as in fig. 7, but also on V_2 and ϕ. In particular, β_1 is greater than the value given by fig. 7 if $\phi = \pi$ (and $V_2 \neq 0$). Similarly β_2 is affected by the fundamental frequency, but this is of less importance here. It follows immediately from equations (10) and (11) that the increase in β_1 corresponds to a greater current excursion and a higher power at the fundamental frequency.

In normal single-frequency operation the oscillation starts, on increasing the supply current I_0, when the initial slope of $\beta_1 I_0$ as a function of V_a reaches a critical value (see fig. 10). This is how we found the starting current previously (eq. 15). When β_1 increases the initial slope of β_1 also increases, so that the required initial slope of $\beta_1 I_0$ is reached earlier and the starting current is smaller. Conversely, for $\phi = 0$, the starting current is greater than with a single-frequency oscillation. If we increase the supply current from low values, the first signal we obtain is therefore a two-frequency oscillation with the phase relationship $\phi = \pi$.

A significant value of V_2 is only obtained provided Q_2 is sufficiently large (see eq. 9), i.e. if the starting current for oscillations at 2ω alone, $I_{\text{start }2}$, is sufficiently low (see eq. 15). In this context "sufficiently" means in practice that $I_{\text{start }2}$ should be of the same order of magnitude as I_0. If R_{S2} is very small, it is then possible to give Q_2 a high enough value to make $I_{\text{start }2}$ even smaller than $I_{\text{start }1}$. When I_0 is increased the oscillator then starts at the higher harmonic, and the fundamental frequency does not come in until later. Such an oscillator can be regarded as a 2ω oscillator that generates subharmonics. However, for this to happen Q_2 must be substantially greater than Q_1 (e.g. $Q_2 > 100$ for $Q_1 = 7$). This can be seen from eq. (15): not only is $\omega_2{}^2$ four times greater than $\omega_1{}^2$, but Φ_{22} is smaller than Φ_{21}.

Apart from increasing the power as described above, coupling the fundamental mode to an upper harmonic mode of higher Q also gives an appreciable improvement in the noise characteristics of the oscillator and also greater frequency stability.

The method has been tested experimentally with an oscillator of the design shown in fig. 1, with four sliding-sleeve tuners. With the same type of diode for 5 GHz that was used in the oscillators described earlier, the output power and efficiency shown in *fig. 19* were obtained [11]. On comparing this with fig. 3 the following improvements can immediately be seen: a) greater power for the same d.c. current; b) good operation still possible at a supply current considerably higher than 100 mA; although the efficiency does not increase further when I_0 rises above 90 mA, the power goes on increasing; c) higher maximum efficiency (9% as against 7%).

An oscillator designed for two-frequency operation is shown in *fig. 20*. Here again the diode is of the same type and the fundamental frequency is again 5 GHz. As in fig. 1, the quarter-wave transformer is used for tuning to the fundamental frequency. Resonance at 10 GHz is obtained with the short-circuited tuning stub, which is $\frac{1}{2}\lambda$ long for 10 GHz. A standing wave is produced at 10 GHz with a node at the junction, so that no power goes to the output. The tuning stub is $\frac{1}{4}\lambda$ long at 5 GHz and therefore has very little effect on the 5 GHz energy. The two oscillations can

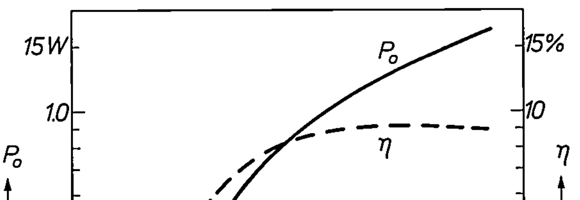


Fig. 19. Output power and efficiency for 5 GHz oscillations of a two-frequency oscillator (5 and 10 GHz), a modified version of the oscillator in fig. 1, with four sliding-sleeve tuners. The advantage of the second-harmonic tuning can be seen from a comparison with fig. 3. Both figures relate to the same type of diode.

be accurately matched (frequency ratio 1 : 2) by means of the capacitive tuning screw, which, for 5 GHz, is situated half a wavelength from the diode, i.e. electrically in parallel with it.

In this oscillator the starting current for 10 GHz oscillations varied between 35 and 45 mA from one diode to another. The starting current for 5 GHz was made larger than 45 mA (45-65 mA) by choosing appropriate values for the quarter-wave transformer. The low starting current obtained here for 10 GHz (the high value of Q_2) can easily lead to an excessive voltage excursion at the upper harmonic. This was found to give rise to unwanted modulation by amplified low-frequency noise at output powers greater than 1 W. For output powers less than 1 W, however, the frequency noise was 20 dB less than in single-mode operation.

Other possible applications of circuits operating in two modes include amplifying frequency multipliers and frequency dividers [4], but these will not be dealt with here.

Injection of a stabilized signal at the second harmonic

We have seen above that it is possible, by means of a simple modification to the microwave circuit, to make a two-frequency oscillator that has better characteristics than a single-frequency oscillator using the same diode. Achieving this in practice, however, sets a difficult specification for the diode. The diode is optimized for the fundamental frequency f_0. This means that the negative resistance at the frequency $2f_0$ will be comparatively small, so that even with relatively low internal losses the diode will behave passively instead of actively at the frequency $2f_0$. This disadvantage is not found with the modified circuit shown in *fig. 21*, which is therefore of practical interest.

The circuit contains two IMPATT-diode oscillators (A and B), one for the frequency f_0 and one for the frequency $2f_0$, each with a diode optimized for its own frequency. Oscillator B is stabilized by a transmission cavity resonator C of high Q, which suppresses frequency noise. The signal obtained at frequency $2f_0$ is injected into oscillator A. This gives the same coupling in oscillator A between the signals of frequency f_0 and $2f_0$ as described above. The transmission cavity resonator is designed so that signals of frequency f_0 are not transmitted, thus preventing oscillator A from reacting on oscillator B. The output signal from oscillator A at the frequency f_0 now has the good noise characteristics of the stabilized signal with frequency $2f_0$.

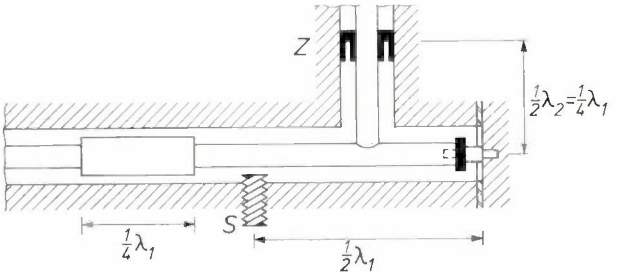


Fig. 20. Coaxial oscillator circuit designed for operation at two frequencies (5 and 10 GHz). At 5 GHz the oscillator is equivalent to the design of fig. 1. A high-Q resonance at 10 GHz is obtained by means of the short-circuited stub Z which is $\frac{1}{2}\lambda$ long at 10 GHz. The capacitive tuning screw S is used to match the two oscillations accurately in the frequency ratio 1 : 2.

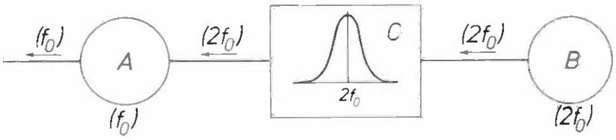


Fig. 21. Diagram for injection of a signal with double frequency. A main oscillator (frequency f_0), B auxiliary oscillator (frequency $2f_0$), C high-Q transmission cavity resonator for the frequency $2f_0$. The signal from B, stabilized by the cavity resonator, is injected into A, thus improving the Q of the output signal from A. The transmission cavity resonator only passes signals of frequency $2f_0$, not those of frequency f_0.

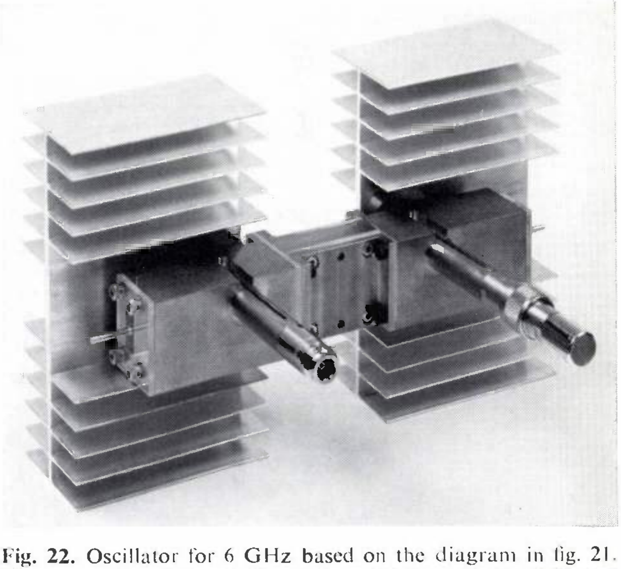


Fig. 22. Oscillator for 6 GHz based on the diagram in fig. 21. The connection between the main coaxial oscillator (left) and the auxiliary coaxial oscillator (right), including the cavity resonator in the middle, consists of a waveguide with a cut-off frequency of 8 GHz, so that 12 GHz energy is passed, but not 6 GHz energy. The 6 GHz output signal contains about 30 dB less noise than that of the 6 GHz oscillator with no injected signal.

An oscillator designed on these principles has been built to operate at 6 GHz (*fig. 22*). The main coaxial oscillator for 6 GHz and the auxiliary coaxial oscillator for 12 GHz are connected by a piece of rectangular waveguide. Two irises in the waveguide are arranged to form a $\frac{3}{2}\lambda$ cavity resonator at 12 GHz, with a

[11] K. Mouthaan and H. P. M. Rijpert, Second-harmonic tuning of the avalanche transit-time oscillator, Proc. IEEE 57, 1449-1450, 1969.

loaded Q of about 200. The cut-off frequency of the waveguide is about 8 GHz, so that 12 GHz energy is transmitted, but not 6 GHz energy. The coaxial terminals of the 12 GHz oscillator are terminated with a matched load to avoid parasitic oscillations at low frequencies. The output power is taken from the coaxial terminals of the 6 GHz oscillator.

With this circuit an r.m.s. frequency fluctuation of only 10 to 20 Hz in an output signal of 1 W was measured in two symmetrical sidebands of 3 kHz bandwidth. This is an improvement of more than 30 dB on the 6 GHz oscillator with no injected signal.

Summary. The characteristics of IMPATT-diode oscillators are analysed with the aid of equivalent circuits. The active element, the IMPATT diode, is reverse-biased into breakdown and supplied from a d.c. current source. Characteristic of the high-frequency behaviour are the creation of bunches of charge carriers in the avalanche region and their transit-time effect in the drift region. Analysis gives an equivalent circuit which, at sufficiently high frequencies, can be simplified to two capacitances and an r.f. current source in antiphase with the r.f. voltage. The frequency of the oscillator (diode plus termination) is primarily determined by the terminating reactance, and the amplitude is determined by the supply current and terminating resistance. The r.f. voltage must be limited to about 30% of the breakdown voltage. For the diodes normally used this gives an upper limit of about 15% for the efficiency, and a maximum permissible supply current of about 2.5 times the starting current.

The analysis is applied to two experimental oscillators with silicon diodes, both for 5 GHz, one in coaxial form and the other in microstrip. The output power of the coaxial oscillator is more than 1 W, the efficiency about 7%.

Rectification of the r.f. signal in the diode can cause low-frequency instabilities, which must be allowed for in the design of the supply circuit. Load and temperature variations can cause frequency variations. Frequency stability is improved by coupling to a high-Q resonant circuit. An oscillator can be tuned by means of the load. As an example an oscillator for about 9.4 GHz is discussed, which can be electronically tuned over 125 MHz.

The output power, efficiency, frequency and noise performance of an IMPATT-diode oscillator can be considerably improved by tuning the circuit to a higher harmonic. This is explained theoretically and demonstrated with experimental examples. A coaxial oscillator for 5 GHz operating on these principles, and using a silicon diode, gave an output power of 1.75 W, an efficiency of more than 9%, and a noise level 20 dB lower than with single-frequency tuning. A modification of this circuit that probably offers greater practical prospects is an oscillator coupled to a stabilized auxiliary oscillator for the second harmonic.

Anomalous oscillations with an IMPATT diode

P. J. de Waard

Introduction

In 1967 it was discovered that an IMPATT-diode oscillator can have a mode of oscillation entirely different from the normal transit-time oscillations dealt with in the two preceding articles [1]. This "anomalous mode" [2] is remarkable in the first place for its high efficiency. Recently anomalous oscillations were produced with an efficiency of 75% [3], whereas not much more than 15% can be expected for normal oscillations [4]. Moreover the power output (more than 10 W) and current density (a few thousands of A/cm^2) are much higher than are found with normal oscillations.

A further characteristic of the anomalous mode is the frequency, which is lower than that of a transit-time oscillator and far less affected by the diode used in the circuit. For example, a diode suitable for transit-time oscillations of a few GHz can be made to give anomalous oscillations that can be tuned to any frequency in the range from about 100 MHz to 1 GHz by adjusting the circuit.

These anomalous oscillations are easy to produce in circuits that can be tuned to different frequencies at the same time. Usually, therefore, they are produced and investigated in circuits in which the diode is connected to a coaxial line containing various tuning slugs. In anomalous oscillation a section of the coaxial line functions as a *delay line*, whose length determines the oscillation frequency. It is generally assumed that normal transit-time oscillations are essential to the excitation of the anomalous oscillation. The circuit must therefore be tuned not only to the required anomalous mode but also to a normal mode which appears in the circuit as a higher harmonic of the anomalous one.

Apart from the transit-time mode and the high-efficiency anomalous mode, other kinds of oscillation may occur in IMPATT-diode oscillators, for example relaxation oscillations. Because of their low efficiency and low frequency, however, they are of little interest and are regarded more as unwanted side-effects. Although they all really come under the term "anomalous oscillations", this term will be confined to the high-efficiency mode in this article.

Oscillations can only occur when there is an element in the circuit that has a negative resistance at the oscillation frequency. The d.c. or low-frequency current-voltage characteristic of an IMPATT diode alone shows a positive differential resistance at all current values. At Philips Research Laboratories we have found however that if the d.c. or low-frequency characteristic is measured while the diode is actually oscillating in the anomalous mode in a circuit of the type indicated above, then quite a different shape is obtained. This curve includes a large section with negative slope. A characteristic measured in this way is described as "quasi-static". The diodes investigated were all "punch-through" diodes, i.e. diodes whose middle section is so lightly doped that at breakdown the depletion layer extends to *both* contact regions, and the field gradient in the middle section is small (see fig. 2*a*). It has been possible to generate anomalous oscillations only with diodes of this type in our experimental arrangement [5].

As a working hypothesis it was assumed that the measured quasi-static characteristic forms the basis for the anomalous oscillations, in other words that it can be identified with a "true" characteristic valid for instantaneous values of current and voltage. Starting from this working hypothesis it was possible to make computer simulations of the oscillations; these simulations consisted of calculations of the current and voltage variations for an equivalent circuit in which the diode was simply defined by its measured characteristic. This gives useful information about the operation of the oscillator as a whole, and in particular made it possible to optimize the circuit by comparison of the computed results for various values of the circuit parameters. Computer simulations of this type are widely used in microwave work, owing to the difficulty of varying the parameters in the actual circuit.

In these investigations of the anomalous oscillator as a whole no account has been taken of what happens in the diode itself. Nevertheless, in the next section we

Ir. P. J. de Waard is with Philips Research Laboratories, Eindhoven.

[1] D. de Nobel and M. T. Vlaardingerbroek, IMPATT diodes; this issue, page 328.
K. Mouthaan, IMPATT-diode oscillators; this issue, page 345.

[2] H. J. Prager, K. K. N. Chang and S. Weisbrod, Proc. IEEE **55**, 586, 1967.

[3] D. F. Kostishack, Proc. IEEE **58**, 1282, 1970.

[4] See the tables on page 336 and page 350 of the two preceding articles in this issue [1].

[5] Anomalous oscillations have been obtained elsewhere with diodes of a different type.

shall consider some ideas concerning the processes in the diode, and a qualitative explanation for the shape of the measured characteristic will be given, since these ideas strongly support the working hypothesis. However, the picture of the effects in the diode given there is no more than speculative, and differs to some extent from computer simulations of these effects that have been made elsewhere on the basis of the equations describing the fundamental processes in the diode [6].

The other sections will present in turn a discussion of the equivalent circuit used for the computer simulations and its use in understanding the operation of the anomalous oscillator, the efficiency predicted by the theory, the experimental arrangement and some results of the measurements and calculations.

The quasi-static diode characteristic

An example of a quasi-static characteristic, measured on one of our diodes, is given by the solid curve in *fig. 1*. It has already been mentioned that curves of this type are obtained while the circuit is oscillating. More accurately, the circuit is first adjusted to oscillate in the anomalous mode, and from then on, without making any other adjustment, the average current I_0 through the diode (the supply current) is changed and the average voltage V_0 across the diode is measured as a function of I_0; this gives the solid curve. If, however, oscillation is prevented by damping the circuit in some way, the dashed curve is obtained. Of course, for a given I_0 that is not too small, there are various settings at which anomalous oscillations occur. There is not much difference, however, in the characteristics measured. Where differences of any significance occur, the characteristic with the deepest minimum is taken as representative of the diode, for reasons that will later become clear.

By combining some ideas [7] [8] about high-current effects in the diode, a qualitative explanation of this behaviour can be given. Let us divide the characteristic into two parts. The first part can be explained as an ordinary rectifying effect [9], as follows. Suppose that I_0 is slowly increased from zero. At first the voltage rises rapidly until it reaches the breakdown value V_b. When I_0 exceeds the starting current, transit-time oscillations begin, so that in addition to the d.c. voltage there is an a.c. voltage across the diode, whose amplitude increases with the supply current. If the peak voltage in each period were to become much greater than V_b, a very large current pulse would be expected in each period, owing to the exponential nature of the avalanche effect. The limited average current, however, rules this out. The average voltage therefore adjusts itself differently, so that the peak value is not so far above V_b. The average voltage thus falls, and this explains the initial part of the falling section of the characteristic.

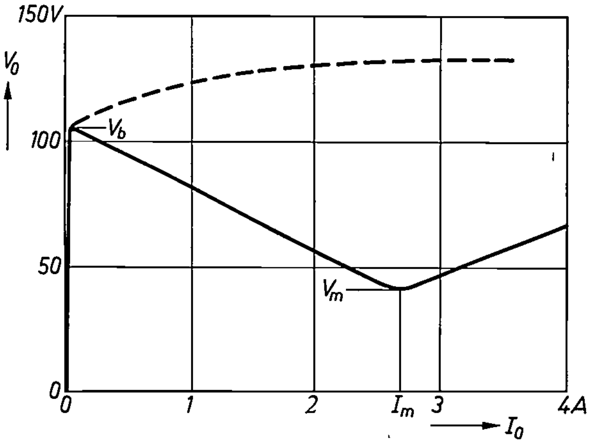


Fig. 1. The quasi-static characteristic (solid curve) of a planar silicon diode, i.e. the d.c. voltage V_0 across the diode as a function of the d.c. current I_0 through the diode, measured while the diode is in a circuit adjusted to give anomalous oscillation (unless I_0 is too low). If oscillation is prevented, the dashed curve is measured. V_b breakdown voltage. V_m and I_m are referred to as valley voltage and valley current.

This explanation does not apply to the whole of the falling section, because beyond a certain value of the supply current the amplitude of the transit-time oscillation rises no further [10]. The following explanation of the second section seems to be the most likely one.

Fig. 2a shows a schematic breakdown field profile typical of the diodes used in these experiments. As discussed in the previous articles, when the current is raised slightly an avalanche region is formed near the position of maximum field-strength, and to the right of it, in fig. 2*a*, a drift region for electrons. On increasing the current further two effects arise that can result in a profile of the type shown in fig. 2*b* [7]. Since figs. 2*a* and *b* show the magnitude of the field directed towards the *left* a positive charge will result in a negative slope (decreasing from left to right) and a negative charge will give a positive slope. Consequently, the effect of the increasing number of electrons is in the first place to make the field profile in the drift region more and more horizontal, causing the avalanche region to spread out. Secondly, in the avalanche region there are large numbers of holes on the left and large numbers of electrons on the right; the slope is thus negative on the left and positive on the right, resulting finally in a dip or "valley" in the field profile. In fig. 2*b* avalanching takes place on both sides of the depletion layer. With a field profile of this type it is reasonable to assume that the characteristic will have a negative slope, for as the current increases, the sides of the valley become steeper. Since the depletion layer cannot become much wider (if the P^+ and the N^+ regions are doped heavily enough) and the peaks cannot rise far above the breakdown field, the valley must become deeper. The area underneath the curve therefore decreases, i.e. the voltage drops.

Results of this kind for diodes of this type are obtained quantitatively if the equations for the field and the hole and electron current are solved numerically [7], starting from analytical expressions for the doping profile and also known expressions [11] for the ionization coefficients and drift velocities of electrons and holes in silicon, as functions of the field. Calculations of this type have been performed at Philips Research Laboratories for one of the experimental diodes. Results are presented in *fig. 3*. The valley obtained in the field profile when the current reaches a certain value can be seen in fig. 3*a*. In fig. 3*b* it can be seen that when the current is high the hole and electron currents have about the same magnitude over a wide region in the middle of the diode. In this region there is an almost neutral cloud of charge carriers. This is the plasma implicit in the acronym TRAPATT (trapped plasma avalanche triggered transit), the name often used for diodes oscillating in the anomalous mode. This name, which will not be discussed further here, is based on a somewhat different, more dynamic picture of the processes in the diode, deduced from the computer simulations of these effects mentioned earlier [6]. In this picture a wave passes through the field profile during each period. In the picture outlined here the field profile simply varies up and down between different curves like those illustrated in fig. 3*a*.

The non-appearance of the negative slope when the diode is not oscillating (fig. 1, dashed curve) can be explained as follows. If a random fluctuation arises in the current distribution in a plane at right angles to the direction of the current, causing a local increase in current density, the voltage drop across the depletion layer will be lower at this position, because of the falling characteristic. Consequently more charge carriers are drawn from the supply regions, and the current density rises still further. Fluctuations of this kind thus reinforce themselves, and *current channels* are formed [8]. At the same time, however, the current in the supply region is also compressed into narrow channels, whose effective resistance is thus increased. It is assumed that this resistance effect exceeds the negative resistance effect in the depletion layer, so that the characteristic of the whole diode nevertheless has a positive slope.

The final link in the explanation is the assumption that it takes so much time for these current channels to

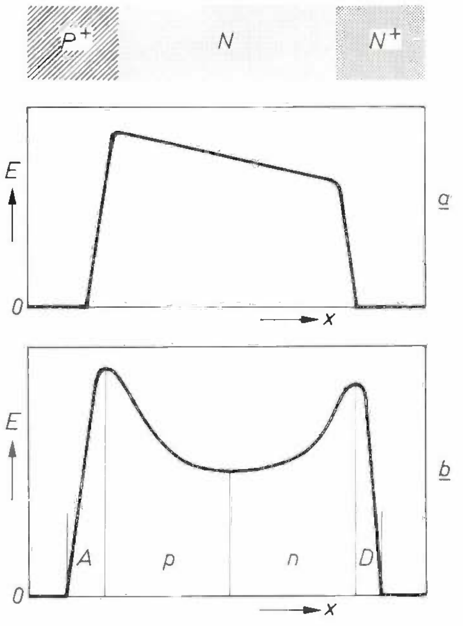


Fig. 2. The field profile assumed in a diode at breakdown (*a*) and at high current (*b*). In the diodes used here the depletion layer extends to the N^+ region at breakdown. At low current, avalanching releases holes and electrons in the high-field region (on the left in *a*); the electrons move through the *N* region to the right. At high current, the charge of the holes and electrons gives rise to a profile with a dip or "valley"; *A*, *p*, *n* and *D* indicate whether the prevailing charge is due to acceptors, holes, electrons or donors, respectively. Avalanching now occurs at both peaks.

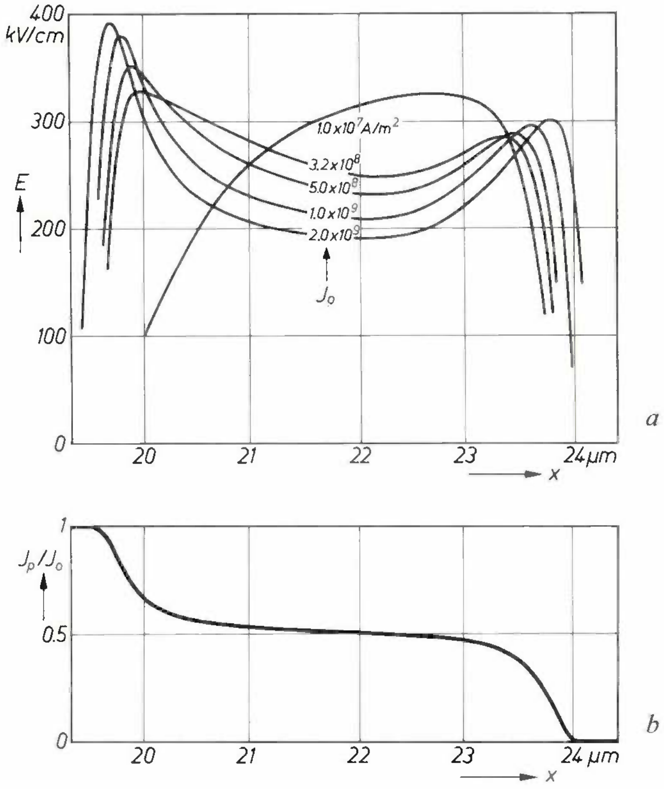


Fig. 3. *a*) Results of computer calculations of the field *E* as a function of *x*, the distance from the original silicon surface, for various values of current density J_0. *b*) The same for J_p/J_0, the fraction of the current carried by holes, for the case where $J_0 = 2 \times 10^9$ A/m². The calculations were based on the known variation [11] of the ionization coefficients and the drift velocities of holes and electrons in silicon as a function of the field, and on the measured doping profile.

[6] B. C. DeLoach, Jr., and D. L. Scharfetter, IEEE Trans. ED-17, 9, 1970.

[7] See H. C. Bowers, IEEE Trans. ED-15, 343, 1968.

[8] A. M. Barnett, IBM J. Res. Devel. 13, 522, 1969.

[9] This rectifying effect is discussed in another way in the article by K. Mouthaan [1]; see fig. 13 (page 353).

[10] See the article by D. de Nobel and M. T. Vlaardingerbroek [1], page 338.

[11] C. A. Lee, R. A. Logan, R. L. Batdorf, J. J. Kleimack and W. Wiegmann, Phys. Rev. 134, A 761, 1964.

form that the *oscillating* diode has a falling characteristic: the current then varies so rapidly that the current channels have no chance to develop. The tendency for channels to form increases, however, as the average current increases, so that above a certain current value the voltage increase associated with channel formation becomes dominant after all. This provides an explanation for the rising part of the characteristic (fig. 1).

In many respects the IMPATT diode is the dual of the Gunn diode [12]. The current channels, for instance, correspond to the high-field regions in the Gunn effect. Also, the characteristic of the anomalously oscillating diode, if current and voltage are interchanged, is much like that of the Gunn diode oscillating in the LSA mode [12].

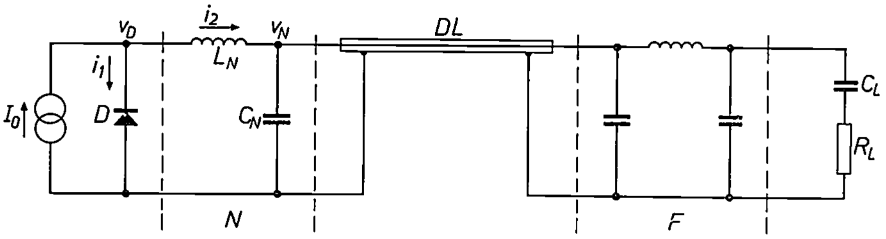


Fig. 4. The equivalent circuit for the anomalous oscillator used as the basis for the computer simulations. The diode is defined here as an element with the characteristic of fig. 5. The current I_0 supplied by the current source is divided between the diode D and the rest of the circuit; $I_0 = i_1 + i_2$. The network N, in particular the inductance L_N, causes the diode to switch easily from the "switched on" state ($i_1 = I_m$) to the "shut-off" state ($i_1 = 0$), and *vice versa*. When the diode shuts off, a voltage and a current front appear at the input of the delay line DL (characteristic impedance Z_0 and delay time τ). Part of this signal is reflected by the low-pass filter F and causes the diode, which had meanwhile switched on again, to shut off once more after a time 2τ. Similarly the switching on of the diode causes it to switch on again after a period 2τ. Only the fundamental frequency of the oscillations is passed by F and arrives via the large bypass capacitor C_L in the load R_L. The diode and the delay line should preferably be matched. The coupling network N has an important corrective function if there is a mismatch.

Equivalent circuit; operation of the oscillator

The equivalent circuit on which the computer simulations were based is shown in *fig. 4*. It is assumed that the diode has the characteristic shown in *fig. 5*, which is a schematic piecewise-linear version of the characteristic in fig. 1. In fig. 4 a reverse-biased diode D is connected to a current source, and there is a delay line DL connected in parallel with D via a coupling network N. The delay line is connected to a matched load R_L through a low-pass filter F. A large bypass capacitance C_L ensures that no d.c. current can flow through R_L.

This diagram will now be used to explain the operation of the oscillator, assuming for convenience that the current i_1 at which v_D is equal to V_b (see fig. 5) can be neglected. It is also assumed that the valley voltage V_m is zero, and that I_0, the value of the source current, is half that of the valley current I_m.

In the steady state the delay line draws no current. The whole of the supply current I_0 then flows through the diode, which is therefore "half on": $v_D = \frac{1}{2}V_b$; $i_1 = I_0$; $i_2 = 0$. Owing to the interaction of the inductance L_N with the diode, however, this state is easily perturbed. Assume, for instance, that a fluctuation occurs in the current distribution at the junction point above D, causing i_1 to drop and thus v_D to rise accordingly; the diode "shuts off" a little. Since $i_1 + i_2$ is constant, i_2 must rise at the same time. This rise induces an emf in L_N that opposes i_2 in the circuit external to L_N. Since v_N remains constant at first, because the impedance of L_N for rapid variations is much greater than that of C_N and DL, there is consequently a further rise of v_D. The diode is thus still further shut off by L_N. This process continues until the diode is completely shut off. It follows analogously that a fluctuation of the appropriate sign can easily cause the diode to switch from the shut-off to the switched-on state, and *vice versa*.

Once the diode is shut off, i_2 remains temporarily constant, the induced emf thus disappears, and v_N follows v_D. At the same time charge begins to flow into the delay line. The voltage and current front which thus form at the line input propagate along the line and

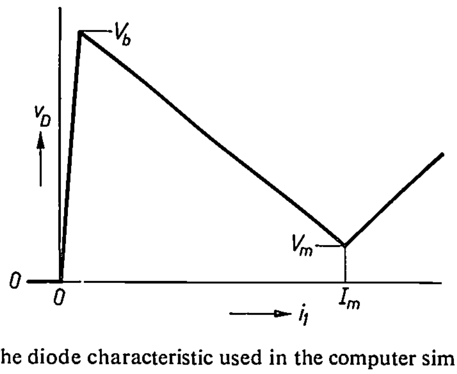


Fig. 5. The diode characteristic used in the computer simulations.

reach the output after a time τ, the delay time. If v and i are the voltage and current at an arbitrary point of the line, and $v - \frac{1}{2}V_{\rm b}$ and i are therefore their respective *deviations* from the average steady state, then the ratio $(v - \frac{1}{2}V_{\rm b})/i$ is equal to Z_0, the characteristic impedance of the line, at any rate as long as there are no reflected signals.

It will now be shown that oscillations in which the diode switches from the switched-on to the shut-off state and back to the switched-on state again, are self-sustaining. For the time being we shall not be concerned with how the oscillation starts, in other words how the first two reversals take place; we shall return to this point later. We assume that $t_2 - t_1$ is smaller than 2τ, where t_1 and t_2 are the times at which the diode first shuts off and then switches on again (*fig. 6* above). For simplicity let us make t_2 roughly equal to $t_1 + \tau$, though this is not essential to the argument. Voltage and current distributions in the delay line following these reversals are given in fig. 6. Shortly after the first reversal ($t = t_1 + \Delta$) there is a front at the beginning of the line. This reaches the output at the time $t_1 + \tau$. The slow component of the voltage and current variation passes the low-pass filter F and vanishes in the matched load. For the fast components, however, F constitutes a short-circuit; these components are therefore reflected, resulting in a negative voltage and positive current pulse travelling to the left. The situations just before and just after this reflection — and hence before and after the second reversal, which took place at approximately the same time — can be seen in fig. 6 ($t = t_2 - \Delta$ and $t = t_2 + \Delta$). The reflected signal reaches the input at the time $t = t_1 + 2\tau$. The current pulse arriving there is a "fluctuation" whose sign is such as to shut off the diode again, which had been on since the second reversal. This is because the current pulse arriving at the input is a momentary increase of i_2 which must be associated with a decrease of i_1, i.e. a variation in the direction of the shut-off state of the diode. Analogously the signal appearing at the output as a result of reflection at time $t_2 + \tau$ causes the diode to switch on again

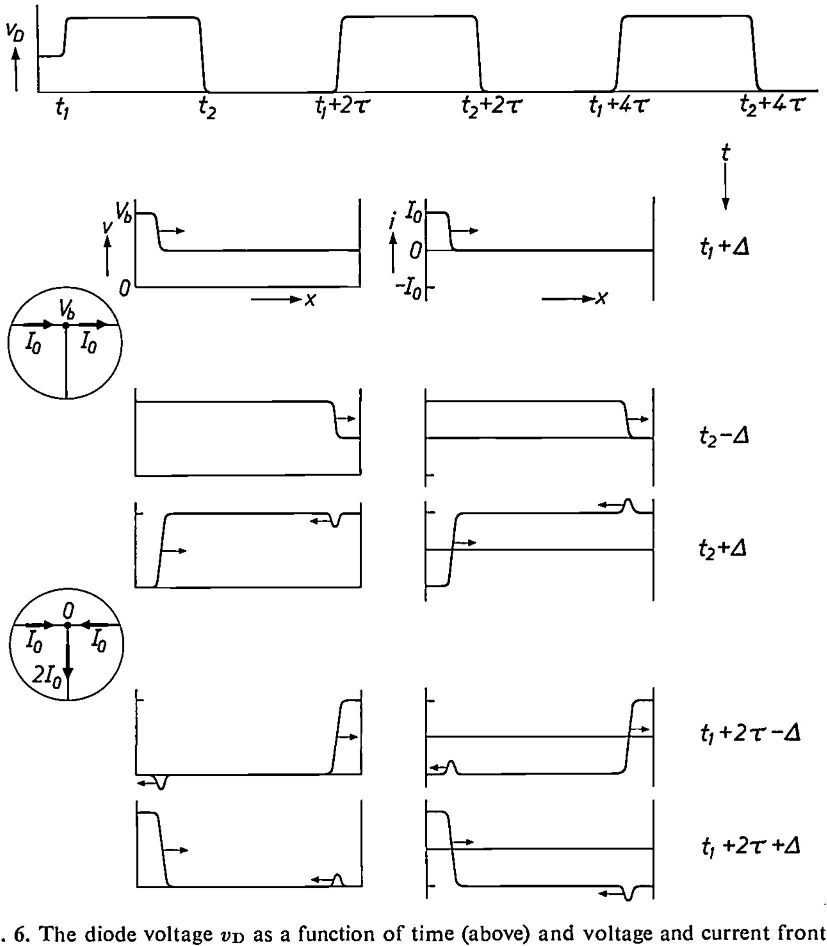


Fig. 6. The diode voltage $v_{\rm D}$ as a function of time (above) and voltage and current fronts in the delay line, each shown a time Δ after a diode reversal and a time Δ before the next reversal. The current distribution at the junction point above the diode is shown in the circular insets.

[12] See the article by G. A. Acket, R. Tijburg and P. J. de Waard in this issue, page 370.

at time $t_2 + 2\tau$. Each event is repeated after a period 2τ, which is the period of the oscillations.

The question that remains to be answered is how the first two reversals come about. For the first reversal this presents no problem: as we have seen, any arbitrary fluctuation on the initial steady state can cause the diode to switch on or off. The second reversal, after a time of approximately τ, can be attributed to the operation of the coupling network. This forms an *LC* circuit which, excited by the first reversal, reacts like a resonator. Here it will be sufficient to note that, if suitable values are chosen for the parameters, the computer simulation also shows that oscillations are initiated by a single excitation of the oscillator.

To summarize, the coupling network, driven by the diode plus current source and working into the load Z_0, has an inherent tendency to oscillate. The frequency of the oscillations, however, is determined by the delay line through the mechanism described above of reflected signals that initiate diode reversals.

In the ideal case the diode and the delay line are matched ($Z_0 = (V_b - V_m)/I_m$). In the absence of the coupling network, the diode reversals can then in principle take place in any arbitrary way, including instantaneously. This is so because at each change in v_D (identical with v_N if there is no L_N) changes can be found in i_1 and i_2 that correspond to the diode characteristic and to the line impedance as well as to the conditions that $i_1 + i_2$ must be constant. In this case oscillations can be imagined in which the variation of voltage and current is that of a square wave, the voltage varying between V_m and V_b, the current between I_m and approximately zero. If moreover "on" and "off" periods are equally long, then these are the oscillations with the maximum efficiency. In the event of a mismatch — which cannot always be avoided in practice — currents can arise in the diode that may be larger than I_m. The characteristic then appears to be flatter. The average voltage is greater and the oscillations are no longer "square-wave", two causes of a drop in efficiency.

The coupling network *N* has an important corrective function if there is a mismatch. If Z_0 is greater than $(V_b - V_m)/I_m$, the delay line responds to a voltage variation $V_b - V_m$ with a current variation smaller than I_m. This means that with no coupling network present the diode would not switch over completely; the capacitor C_N, however, now supplies the missing current variation. Conversely, L_N supplies the missing voltage variation if Z_0 is too small. Because of these effects the waveform of the oscillations is however no longer purely rectangular.

As noted in the introduction, the practical circuit must also resonate in a normal transit-time mode, since the initial part of the characteristic in fig. 1 has a negative slope only because of oscillations in this mode, and anomalous oscillations can only be initiated because of this initial part. This feature of the circuit does not appear from the diagram in fig. 4, but is implicit in the diode characteristic.

Efficiency

The efficiency of the diode is greatest when it is producing square-wave oscillations between the peak ($I_0 = 0$, $V_0 = V_b$) and valley ($I_0 = I_m$, $V_0 = V_m$) of the characteristic, and when the durations of these states are equally long. The a.c. current and voltage are in antiphase. The output power of the diode is then the product of the square-wave current amplitude $\frac{1}{2}I_m$ and the square-wave voltage amplitude $\frac{1}{2}(V_b - V_m)$, i.e. $\frac{1}{4}I_m(V_b - V_m)$. The d.c. current required is $\frac{1}{2}I_m$, at a d.c. voltage of $\frac{1}{2}(V_m + V_b)$, so that the d.c. power supplied is $\frac{1}{4}I_m(V_m + V_b)$. Under these conditions the efficiency is therefore $(V_b - V_m)/(V_b + V_m)$.

Now only the power W_L delivered to the matched load R_L in fig. 4 is useful power; this is the power in the fundamental component. If I_a and V_a are the current and voltage amplitudes of this fundamental component the useful power is $\frac{1}{2}I_aV_a$. From Fourier analysis we know that the amplitude of the fundamental component of a square wave is $4/\pi$ times the square-wave amplitude. The efficiency is thus $\frac{1}{2}\times(4/\pi)^2$ or 0.81 times the value calculated above. The characteristic in fig. 1, with $V_b = 105$ V and $V_m = 42$ V, gives an efficiency of 35% calculated in this way. The theoretical maximum is reached at $V_m = 0$ and amounts to 81%. As noted in the introduction, oscillations have recently been obtained with an efficiency of 75% [3]. If V_m is to have a low value, the two peaks in the field which occur when the current in the depletion layer is high (see fig. 2*b*) must be narrow, which implies a steep doping profile.

The experimental arrangement; results

The experimental oscillator is shown in *fig. 7*. The largest item is the coaxial line, which has a characteristic impedance of 50 Ω. The oscillation frequency is determined by the distance between the diode *D* and the filter *F*; the frequency can be adjusted by moving the filter. For a frequency of 1 GHz — requiring a delay time τ of 0.5 ns — the diode/filter spacing must be 15 cm (the dielectric is air). The diode is connected to the power source *S* through the inner conductor; the source can supply either d.c. current, or pulses, to prevent thermal overloading of the diode. Since the resistance R_S in the supply line is high (500 Ω), the source is a good approximation to a constant-current source. The inductance L_S prevents high-frequency energy from reaching the supply line. The load R_L is a 50 Ω microwave matched load and is made of attenuating material. The bypass capacitance C_L is a gap in the inner conductor. The low-pass filter F consists of two metal slugs, which do not make contact with the inner conductor. Each of these slugs behaves like a capacitor, and the part between them as an inductance. The coupling network N consists of two PTFE slugs, which are also used to resonate the circuit in the transit-time mode.

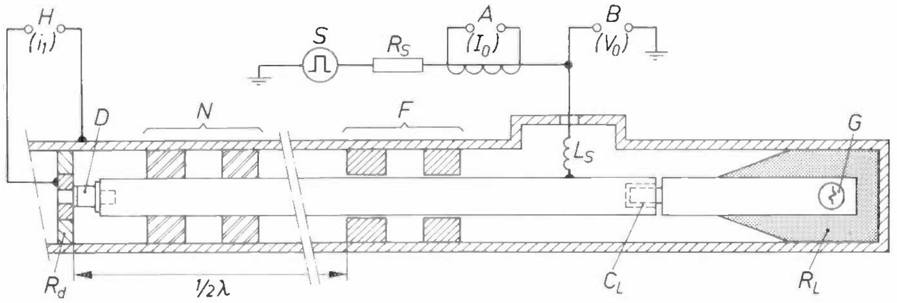


Fig. 7. The experimental arrangement (schematic). The diode D is mounted at the end of a coaxial line, the length between the diode and the filter F acting as a delay line. The coupling network N and the filter F of fig. 4 are formed by inserting slugs in the line: PTFE slugs for the coupling network and capacitive metal slugs for the filter. The diode is connected to the power source S through the inner conductor. Because of the high resistance R_S (500 Ω) the supply is a good approximation to a constant-current source. L_S inductance, acting as high-frequency separation between coaxial line and supply. C_L gap in the inner conductor, acting as a d.c. isolating capacitance. R_L load resistance. G bolometer for power measurement. The quasi-static characteristic is determined by measuring I_0 at A (using a current transformer if the supply current is pulsed) and V_0 at B. The high-frequency diode current i_1 is determined by measuring the voltage across the disc resistor R_d (1 Ω), using a sampling oscilloscope at H.

Various types of diode made at Philips Research Laboratories were used for exciting oscillations in this arrangement. The quasi-static I_0-V_0 characteristic of each diode was plotted (I_0 was measured at A and V_0 at B), and attempts were made to set the slugs, in particular N, to the position that gave the lowest possible value of V_m, the valley voltage. According to the theory, this should yield the highest efficiency. A typical characteristic of this type was shown in fig. 1. The power generated in the anomalous mode was measured with a bolometer in place of the load (at G in fig. 7). Some results relating to power and efficiency are given in *Table I*.

The measured efficiency of a planar diode is sometimes greater than would be expected from the characteristic. The explanation for this is probably as follows. According to our working hypothesis, the quasi-static characteristic measured during oscillations corresponds to the "true" characteristic valid for instantaneous values of currents and voltages. This would be correct if the oscillations all occurred on the negative slope of the linear part of the true characteristic. In reality, however, both current and voltage go beyond the valley during the oscillations, and this means that the average values give an unfavourable picture of the real situation. In the case of the Gunn effect, *known* a.c. voltages are applied (to avoid the analogue of "current channel formation") and the true characteristic can be calculated from the measured one [12]. In our arrangement, however, the diode is left to oscillate freely; the waveform and amplitude of the oscillations are unknown, and there is consequently some uncertainty about the relation between the measured and the true characteristic.

Table I. Power and efficiency of the anomalous oscillations obtained with various diodes in the arrangement shown in fig. 7.

Diode	Supply	Frequency	Power	Efficiency
Si, planar	pulsed	0.8-1.5 GHz	30-50 W	40-50%
Ge, mesa	d.c.	1 GHz	1.5 W	32%
Ge, mesa	d.c.	2 GHz	2 W	22%

In addition to transit-time oscillations and anomalous oscillations of the type described, unwanted relaxation oscillations can occur in our oscillators. Their frequency is in general much lower than $1/2\tau$, and they are mainly determined by C_L, R_L, the supply voltage and the diode characteristic; the coaxial line has scarcely any effect since it is only a small fraction of the wavelength. These oscillations can cause the diode to burn out if C_L discharges for too long a time, which can happen if C_L is large.

Calculated and measured current waveforms

In the computer simulation [13] based on the circuit diagram of fig. 4 the parameter values of the filter F were assumed to be those of a low-pass filter for 50 Ω with a cut-off frequency midway between the fundamental frequency and the second harmonic. Calculations were carried out for various values of L_N and C_N, and these were optimized to give the highest calculated efficiency.

Results of such calculations [14] include the waveform of the current through the diode and the voltage across the diode as a function of time. If these waveforms can also be measured a comparison of the calculated and measured results provides a useful check on the method. To measure the high-frequency current through the diode and its waveform, the diode was placed between inner and outer conductor in series with a one-ohm disc resistor R_d (fig. 7). The voltage measured across the disc resistor gave the desired current. It is difficult to measure the high-frequency voltage across the diode; this would have to be done with a probe near the inner conductor, but the stray capaci-

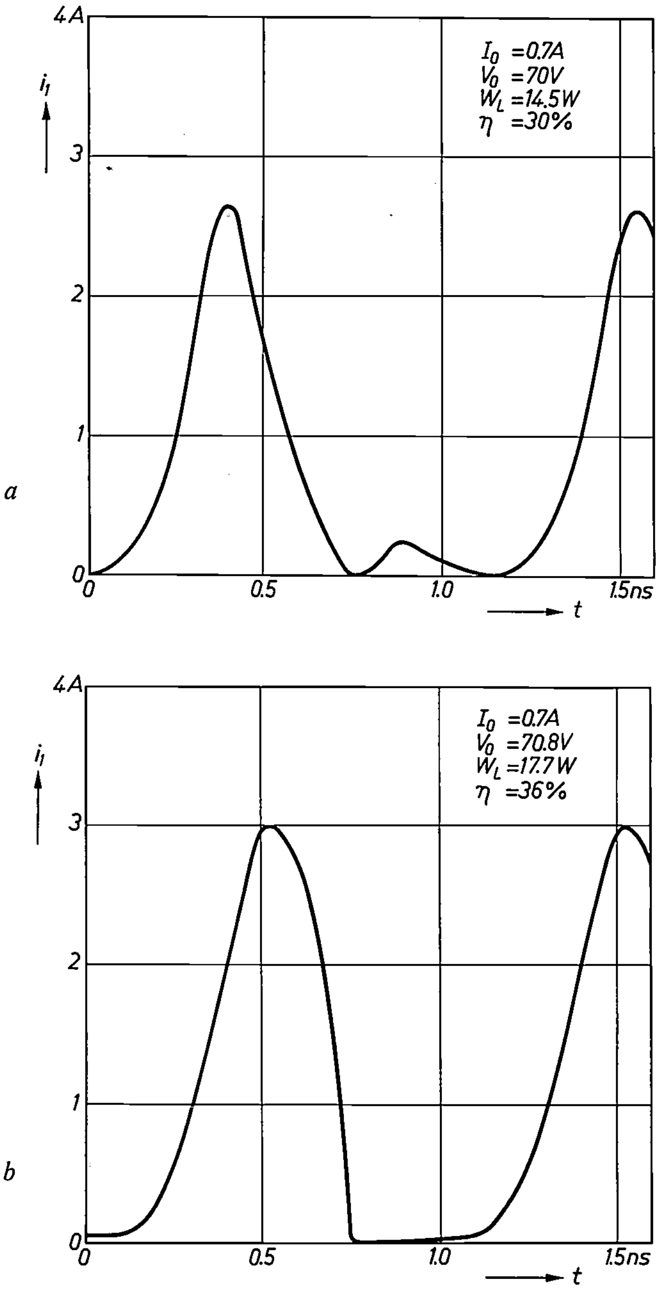


Fig. 8. The current i_l through the diode as a function of time t, for the case in which the filter F passes the fundamental frequency and the oscillator is thus normally loaded (matched load).
a) Result of measurements on the arrangement in fig. 7.
b) Result of calculations using the equivalent circuit in fig. 4.

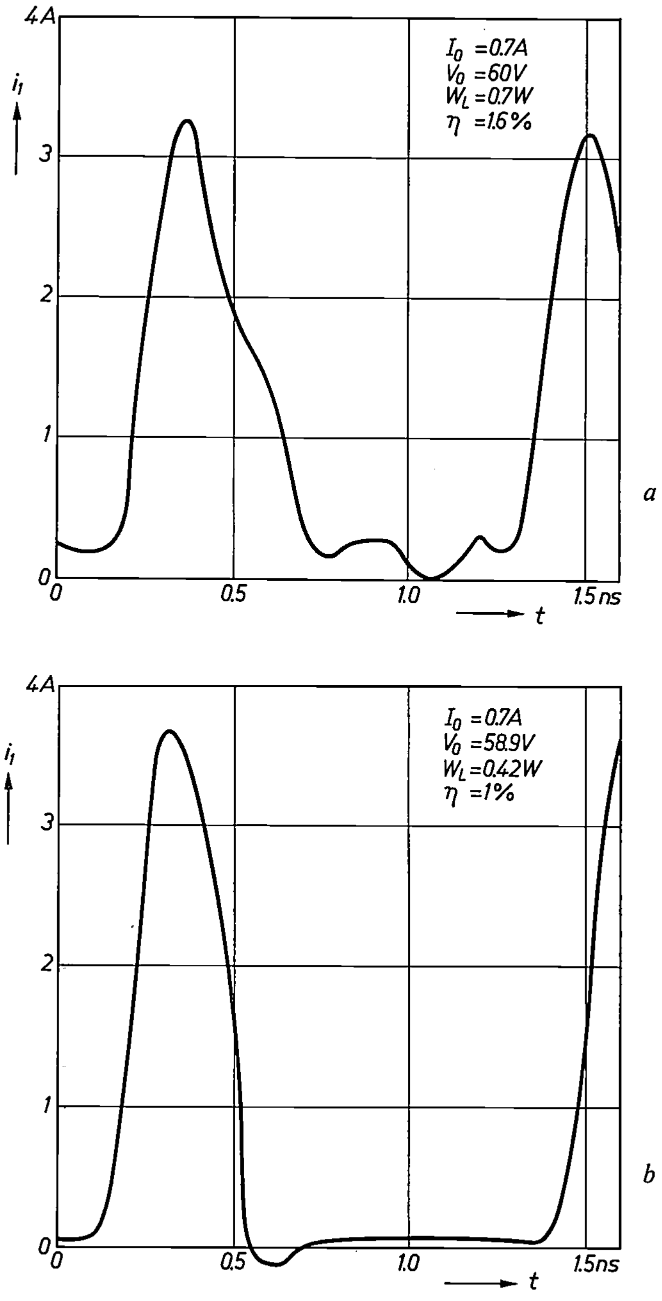


Fig. 9. As in fig. 8, but now for an oscillator with an almost zero load, obtained by lowering the cut-off frequency of the filter F below that of the oscillation frequency. The peaks are now sharper because of the mismatch.

tance introduced would then make it difficult to tune the diode to the transit-time frequency. The high-frequency measurement was performed with a sampling oscilloscope.

To conclude, two examples of the results of our measurements and calculations will now be given. *Fig. 8* shows the r.f. current i_1 as a function of time for a case where the slugs in F were positioned to pass the fundamental frequency but reflect all higher harmonics. The diode current i_1 is by no means a square wave; this is because, for practical reasons, the coaxial line and the diode were not matched ($Z_0 = 50\ \Omega$, $(V_b - V_m)/I_m \approx 20\ \Omega$). *Fig. 9* gives the results for an oscillator with an almost *zero load*; this was realized by setting the slugs in F (and by selecting the parameters of F in the calculations) in such a way that the fundamental frequency was also almost entirely reflected. This gave a severe mismatch of the coaxial line and its load, and consequently the current peaks through the diode were larger and sharper, as can be seen in the figure. In such a case, of course, the power W_L delivered to the load and the efficiency η are very small.

Since the measurements and calculations were almost independent of one another, the agreement between the two is highly satisfactory in a qualitative sense (the sharpness of the peaks) as well as in the quantitative sense (the magnitude of the current). Since the model is really a rather crude one, in particular because of its identification of the quasi-static with the true characteristic, it is not surprising that there are still distinct discrepancies.

Summary. In 1967 it was found that an IMPATT-diode oscillator could operate in a mode that differed from the normal transit-time mode. This "anomalous mode" is characterized by a much lower frequency and a much higher efficiency. The quasi-static characteristic of the IMPATT diodes used here (the relation between d.c. voltage and current while the diode is oscillating) has a region with negative slope, and the anomalous oscillations are attributed to this. At low currents the negative slope is explained as a rectifying effect associated with transit-time oscillations. At higher currents it is explained by the occurrence of a valley in the internal field profile, as a result of the charge of the holes and electrons produced. As the current rises, this valley becomes deeper and the voltage decreases. Because of the formation of "current channels" no negative slope is measured in the static state; during oscillation, however, the current channels have no time to develop. Owing to the negative slope the diode is unstable and easily switches from the shut-off to the switched-on state. In an anomalous oscillator each reversal is initiated by signals reflected from the end of a delay line to which the diode is connected. In the oscillator used here the delay line is a coaxial line with several slugs for tuning and matching. A computer simulation of the oscillations has been made, based on an equivalent circuit in which the diode is defined by the measured quasi-static characteristic. The current and voltage waveforms calculated from this model give a satisfactory agreement with the results of the measurements.

[13] These were carried out on the Philips Electrologica ELX-8 computer at Philips Research Laboratories.

[14] P. J. de Waard, Proc. 1971 European Microwave Conf., Stockholm, Vol. 1, paper A 9/1.

The Gunn diode

G. A. Acket, R. Tijburg and P. J. de Waard

Introduction

The Gunn effect — or, to give it its full name, the Watkins-Gunn effect — is a high-frequency instability that occurs in some semiconductors subjected to high electric fields. When a bar of such a semiconductor (usually N-type) is provided with non-rectifying ("ohmic") contacts and then subjected to a high constant voltage, it is found that the current in the bar is not constant but oscillates at a microwave frequency. This oscillation was first observed by J. B. Gunn [1] in N-type gallium arsenide and N-type indium phosphide. Later the effect was also observed in other semiconductors [2] [3]. A typical example of the current variation through the sample as a function of time can be seen in *fig. 1*. The current waveform consists of a succession of peaks, or spikes, with fairly broad valleys in between. The period of the oscillation is approximately proportional to the distance between the contacts. Using a capacitive probe, Gunn found that the oscillations were accompanied by the periodic movement of domains of very high field through the sample [4]. A high-field domain forms at the cathode, travels at constant velocity through the sample and disappears at the anode, where upon a new domain forms at the cathode, and so on. Before the actual discovery of such domains, their existence was predicted by B. K. Ridley and T. B. Watkins [5]. In fig. 1 the leading edge of each spike corresponds to the disappearance of a domain at the anode, and the trailing edge to the formation of a new domain at the cathode.

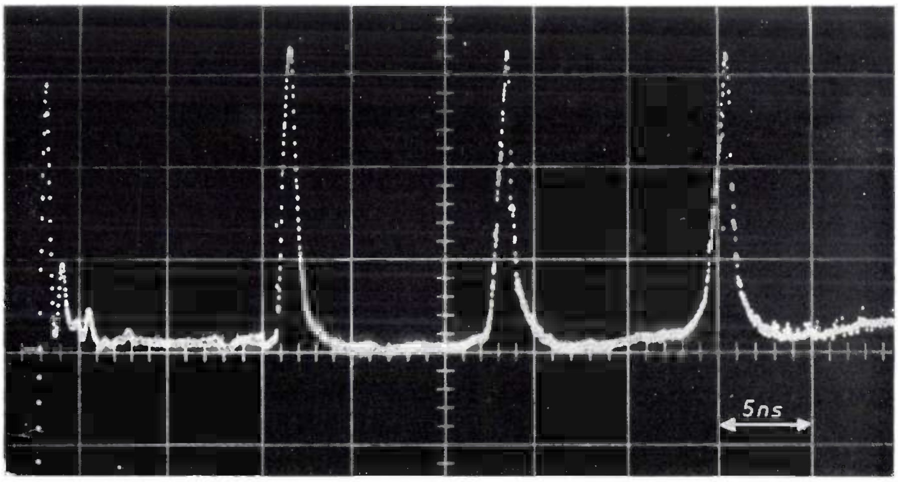


Fig. 1. Characteristic curve of the current I through a sample of N-type gallium arsenide as a function of time t during the occurrence of domain oscillations. The current shows equidistant spikes separated by flat valleys.

Before going deeper into the conditions postulated by Ridley and Watkins for the occurrence of a domain, we must first touch briefly on the high-field behaviour of semiconductor devices in general. At high fields the average kinetic energy of the charge carriers may be substantially higher than the average thermal energy at room temperature; the charge carriers are described as "hot". The current density is then no longer proportional to the electric field, as it is at low fields, and consequently Ohm's law no longer applies. In N-type germanium, for example, the drift velocity at room temperature acquires a constant value as the field is raised. In a certain field region an increase of field-strength may even cause the average drift velocity of the charge carriers to decrease (*fig. 2*): the differential conductivity of the semiconductor — or the differential mobility of the carriers — is then negative. (The differential mobility μ_d is the derivative of the velocity v with respect to the electric field-strength E.) Ridley and Watkins showed that this negative differential conductivity could cause the formation of the domains of extremely high field-strength mentioned above. The experimental observation of the domains suggests that such a negative differential conductivity must be present in all materials that give the Gunn effect.

Dr. G. A. Acket, R. Tijburg and Ir. P. J. de Waard are with Philips Research Laboratories, Eindhoven.

It is easy to understand qualitatively how a sufficiently high field can give rise to space-charge instabilities in such a material [6]. If there is a slight surplus of electrons present anywhere in a Gunn diode, the initially homogeneous field will take the form illustrated in *fig. 3a*, in accordance with Poisson's equation. If now the surrounding field is everywhere higher than E_{th} (see fig. 2), then the high field at the anode end will give the conduction electrons a low average drift velocity, but the low field at the cathode end will result in a high drift velocity. Consequently more electrons will be supplied from the cathode end than disappear at the other end in the same time; the space-charge surplus in the region considered thus increases and an accumulation layer is formed. This layer moves in the direction of the anode, and will go on growing until diffusion prevents further accumulation.

An accumulation layer will always form at the cathode when the field is sufficiently high. Fig. 3*b* shows how, upon the transition from the strongly doped cathode region to the weakly doped diode material, the field-strength increases as a result of the decrease of the electron concentration *n*. Wherever the field exceeds the value E_{th}, the average drift velocity of the conduction electrons there will be higher than anywhere in the immediate vicinity. An accumulation of negative charge thus occurs, which will grow in the manner described above into an accumulation layer.

A local *deficit* of electrons also continuously grows on its way towards the anode as long as the field in the neighbourhood of the deficit is higher than E_{th} (fig. 3*c*). At the anode side of the deficit the field is lower and therefore the average velocity of the conduction electrons is higher than at the cathode side; the electron deficit therefore increases, resulting in a depletion layer. Since this layer can become no more depleted than to an electron density of zero, any further growth from then on will only be in its thickness.

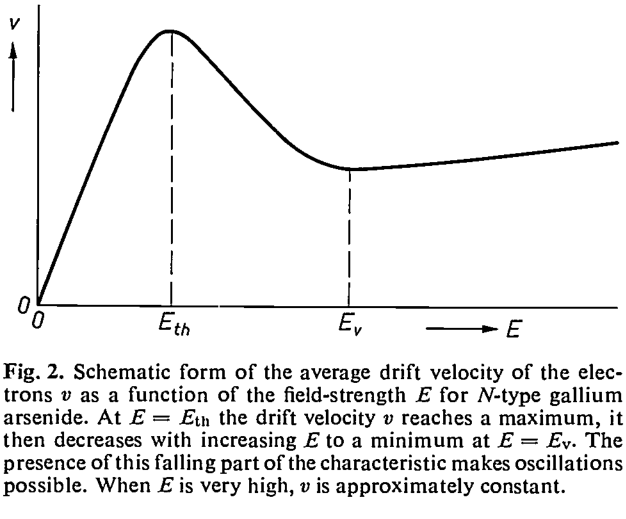


Fig. 2. Schematic form of the average drift velocity of the electrons v as a function of the field-strength E for *N*-type gallium arsenide. At $E = E_{\text{th}}$ the drift velocity v reaches a maximum, it then decreases with increasing E to a minimum at $E = E_{\text{v}}$. The presence of this falling part of the characteristic makes oscillations possible. When E is very high, v is approximately constant.

When an accumulation layer and a depletion layer become close neighbours, they will attract one another and then travel through the diode together. A combination like this, in which the field is higher than in the rest of the diode, constitutes a domain (fig. 3*d*). Here again, because the velocity of the carriers outside the domain is higher than inside it, the charge density inside the domain will differ increasingly from the charge density outside it. The field inside the domain will therefore increase, and may become so high that the field outside the domain drops to below E_{th} so that no new domains can form before those already in existence have disappeared at the anode contact. Domains form at the cathode when there is a local deficit of electrons present there that can grow into a depletion layer, which in practice is always the case. The required accumulation layer is brought about by the mechanism in fig. 3*b*.

The growth in the space-charge layer of all the inhomogeneities described can be characterized by a time constant τ_{d}, given by:

$$\tau_{\text{d}} = \frac{\varepsilon}{|\sigma_{\text{d}}|}, \tag{1}$$

where σ_{d} is the differential conductivity and ε the dielectric constant of the semiconductor. Considering that in an *N*-type semiconductor σ_{d} is equal to $ne\mu_{\text{d}}$, where e is the elementary charge, n the electron concentration and μ_{d} the differential mobility of the carriers, it is evident that a particular doping concentration is required if the fluctuations are to grow fast enough. Material of 1 Ωcm has a time constant τ_{d} of roughly 5×10^{-12} s. In weakly doped material τ_{d} is so great that no domains can form at all.

It has been found that if the instability underlying the Gunn effect is to occur, a charge fluctuation must increase by at least an order of magnitude during the transit from cathode to anode, which means that the factor exp t/τ_{d} must have a value between 10 and 100. Now $t/\tau_{\text{d}} = l\, n\, e\, \mu_{\text{d}}/\varepsilon v$, where l is the distance from cathode to anode. At the values of ε, μ_{d} and v in gallium arsenide (the dielectric constant is 12.5, μ_{d} is -1500 cm^2/Vs, and v is about 1.5×10^7 cm/s) this means that the product nl in this material must be greater than about 10^{12} cm^{-2}.

[1] J. B. Gunn, Solid State Comm. **1**, 88, 1963.

[2] B. J. Elliott, J. B. Gunn and J. C. McGroddy, Appl. Phys. Letters **11**, 253, 1967.

[3] G. W. Ludwig, IEEE Trans. **ED-14**, 547, 1967.

[4] J. B. Gunn, in: Plasma effects in solids, Proc. Symp. Paris 1964, p. 199. See also P. Guétin, IEEE Trans. **ED-14**, 552, 1967.

[5] B. K. Ridley and T. B. Watkins, Proc. Phys. Soc. **78**, 293, 1961.

[6] A mathematical treatment of the space-charge instabilities in Gunn diodes is given by J. de Groot and A. Mircea in this issue, p. 385.

In this article we shall be mainly concerned with the Gunn effect in *N*-type gallium arsenide, which is the material that has hitherto been most widely studied in this respect. After discussing the conduction properties of this material, we shall deal with the oscillations that can occur in a Gunn diode incorporated in a microwave circuit. We shall then look at the technology of the Gunn diode of gallium arsenide, and finally we shall give some idea of the performance of the diodes that we have made.

The *v*-*E* characteristic of *N*-type gallium arsenide

In fig. 2 we saw the remarkable way in which the drift velocity v of the electrons in *N*-type GaAs depends on the field-strength E. The mechanism responsible for this characteristic can only be explained in terms of a wave-mechanical treatment of the motion of a charge carrier in a periodic potential field, i.e. in a crystal lattice. It will suffice here to say that situations may occur in some substances where the majority carriers do not form a single group of particles of the same kind, but form more than one group consisting of particles of different kinds, each characterized by their own mobility, etc. These groups may also have a different minimum energy. The distribution of the electrons among the groups — in our case there are two — depends on the strength of the electric field. At low fields all electrons belong to the group with the lower minimum energy, where they have a high mobility (about 8000 cm^2/Vs.). When the electric field is raised to about 3 kV/cm the average energy of the electrons increases to a value at which the electrons are able to go over to the other group, where their mobility, however, is much smaller (about 200 cm^2/Vs.). When the field is increased still further, more and more electrons go over to the low-mobility group, so that in this field region an increase of field-strength is accompanied by a decrease in the average velocity of the carriers. This decrease is responsible for the region of negative slope in the *v*-*E* characteristic. At still higher fields nearly all electrons have transferred to the second group and the characteristic becomes more or less horizontal.

The difference in minimum energy is such that at room temperature transition to the higher group due to thermal excitation will occur only sporadically. At about 400 °C, on the other hand, such transitions are frequent, and the characteristic falls less steeply.

Because the occurrence of the Gunn effect thus depends ultimately on the possibility of electrons transferring from one group to the other, the Gunn effect is sometimes referred to as the "transferred-electron effect".

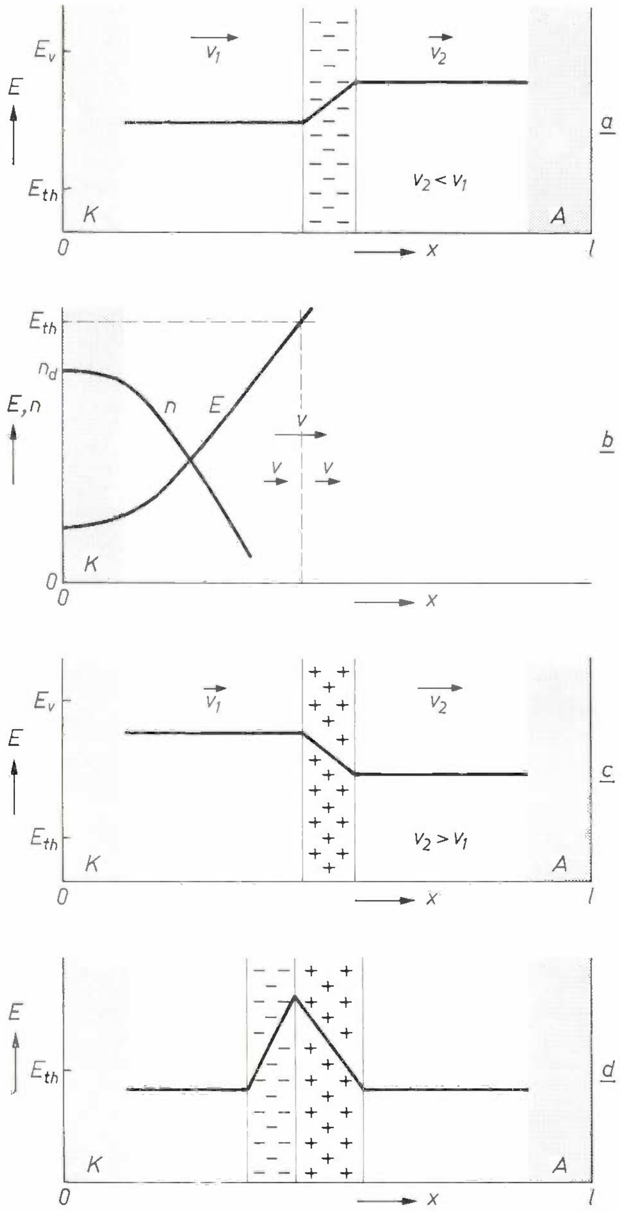


Fig. 3. Illustrating the occurrence of growing space-charge fluctuations and high-field domains in material with a falling *v*-*E* characteristic. *a*) The field-strength E as a function of the space coordinate x in a Gunn diode in which there is a local surplus of electrons. *A* anode. *K* cathode (both heavily doped regions). $E_{\rm th}$ and $E_{\rm v}$ are the values of E between which the *v*-*E* characteristic falls (see fig. 2). When E at both sides of the charge fluctuation has a value between $E_{\rm th}$ and $E_{\rm v}$, the electrons at the cathode side (velocity v_1) move faster than those in the fluctuation, while those at the other side (velocity v_2) move more slowly, so that the charge surplus grows, forming what is called an accumulation layer. *b*) A plot of the field-strength E and the electron concentration n as a function of x, illustrating the formation of an accumulation layer near the cathode. Just beyond the location where the field exceeds the value $E_{\rm th}$ an accumulation layer forms because the electrons in the high-field region have a lower velocity than in the neighbouring region of lower field-strength. *c*) As in (*a*), but now for a local deficit of electrons (depletion layer). This deficit grows as well. A difference compared with (*a*) is that a surplus of electrons can in principle go on accumulating unlimitedly, but a deficit can grow no further when the electron concentration has reached zero. *d*) Like (*a*) and (*c*) for a domain, i.e. a double layer formed by an accumulation layer and a depletion layer. The field inside the domain can become so high that the field outside it drops to a value lower than $E_{\rm th}$. No new domain can then form in the diode until the existing domain has reached the anode.

The difference between the two groups of electrons found in N-type GaAs, and the conditions under which electrons will transfer from one group to the other, become clear if we plot in a graph the dependence of the energy ϵ of the electrons on the magnitude of the wave vector k occurring in the wave function ϕ of the electron [7]. For an electron in a field-free space, ϕ is proportional to $e^{i(k.r)}$ and k is equal to $p/\hbar$, where p is the momentum, or: $\epsilon = \hbar^2k^2/2m$. Here the relationship between ϵ and k thus has the form of a parabola.

For an electron moving in a periodic potential field, like that in a crystal, the wave function has the form:

$$\phi = \mathrm{u}(k,r)e^{i(k.r)},$$

where u is a function which has the same periodicity in space (r) as the crystal lattice. In this case the relationship between ϵ and k is more complicated. In the first place there are k values at which ϵ shows a discontinuity; the intermediate region of ϵ-values is called a "forbidden zone". In semiconductors like N-type GaAs the ϵ-k curve shows a satellite minimum (*fig. 4*) in addition to a minimum at $k = 0$.

It should be noted here that the relation between ϵ and k depends on the direction of the vector $\boldsymbol{k}$: in the [100] directions the satellite valley in N-type GaAs is 0.36 eV higher than that at $k = 0$, in the other directions this energy difference is greater.

It is perhaps also useful to point out that in the movement of an electron in a crystal lattice the direction in which the electron moves — i.e. the direction of its velocity vector $\boldsymbol{v}$ — does not in general coincide with the direction of the corresponding wave vector $\boldsymbol{k}$. This is because the surfaces of constant energy in the $\boldsymbol{k}$ space, to which $\boldsymbol{v}$ is always perpendicular, are not spheres in a periodic potential field, as they are in a field-free space.

We shall now consider the case of an electron in a crystal which is exposed to an electric field E. The proportionality factor between the force $-eE$ acting on the electron and the acceleration which it undergoes, called the effective mass m^*, is inversely proportional to $\partial^2\epsilon/\partial k^2$ in any given direction. Where ϵ varies quadratically with k, m^* is constant: elsewhere m^* will depend on k. In the case illustrated in fig. 4 it can immediately be seen from the curve that m^* is much greater in the satellite valley than in the central valley.

The difference in the height of the valleys makes it clear that at room temperature (0.36 eV corresponds to 4175 K) the satellite valleys cannot be occupied by virtue of thermal excitation. When the crystal is subjected to weak electric fields there is little change in the velocity distribution of the electrons and no transfers will take place. It is only at high fields (about 3 kV/cm) that some electrons will acquire an energy of more than 0.36 eV and be able to transfer to the satellite valley of the ϵ-k curve. It should be noted here that it is not necessary for the electron to acquire an energy sufficient to surmount the "energy hill" between the two valleys: when ϵ is greater than 0.36 eV, an interaction with lattice vibrations can cause the rise in the value of k required for the transfer from the central to the satellite valley.

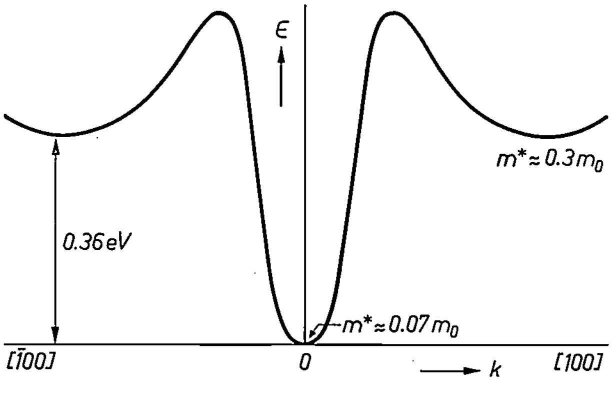


Fig. 4. The energy ϵ of the conduction electrons in N-type gallium arsenide as a function of the wave vector k for the [100] direction and equivalent directions (schematic). Apart from the central valley at $k = 0$ there is a satellite valley lying 0.36 eV higher. The effective mass m^* of the electrons in the central valley is very much smaller than of those in the satellite valleys.

It is difficult experimentally to determine the v-E characteristic in the part with negative slope because it is here that the material shows the Gunn instabilities. The v-E characteristic of a Gunn device must therefore be determined under conditions in which no domains can occur. This can be done using microwave fields of very high strength and frequency [8] [9] or using material of very low conductivity [10].

If the frequency of the measuring signal used is sufficiently high (about 35 GHz for material of 1 Ωcm), the field in the material remains in the falling part of the characteristic for such a short time that no domains can form. During the measurement the amplitude of the measuring signal is varied, and what is measured at each amplitude is in fact the conductivity averaged over the range of E values corresponding to that particular amplitude. The v-E characteristic is then derived from the results [8] [9].

As noted in the introduction, no domains can occur in material that has very low conductivity. In material of this type, measurements are made of the transit time of electrons injected in the form of current pulses at the cathode. From these measurements the drift velocity can be directly calculated.

Results relating to the v-E characteristic obtained from microwave measurements [8] and drift-time measurements [10] are presented in *fig. 5*, together with some recent theoretical results [11] [12]. The two experimental methods give results that are reasonably in agreement with the theory.

The v-E characteristic of N-type GaAs is found to be temperature-dependent [10] [13]; this can be seen from *fig. 6*, where measurements carried out at 100 K are compared with others carried out at room temperature. A result calculated by W. Heinle for 110 K is also included. Both theory and experiment indicate that the field at which the characteristic has its maximum decreases with decreasing temperature, and that the ratio

[7] A more detailed treatment is given in textbooks such as: J. S. Blakemore, Solid state physics, Saunders, Philadelphia 1970.

[8] G. A. Acket and J. de Groot, IEEE Trans. **ED-14**, 505, 1967.

[9] N. Braslau and P. S. Hauge, IEEE Trans. **ED-17**, 616, 1970.

[10] J. G. Ruch and G. S. Kino, Phys. Rev. **174**, 921, 1968.

[11] W. Heinle, Physics Letters **27A**, 629, 1968.

[12] A. D. Boardman, W. Fawcett and H. D. Rees, Solid State Comm. **6**, 305, 1968.

[13] G. A. Acket, H. 't Lam and W. Heinle, Physics Letters **29A**, 596, 1969.

of the maximum to the minimum velocity increases. Further, the characteristic for 100 K shows considerable curvature even for fields that are very much smaller than E_{th} so that even at such relatively low fields there is a marked deviation from Ohm's law.

The physical background of this temperature dependence will not be considered here. It should be noted, however, that at temperatures above room temperature the v-E characteristic will in any case be effected by transfers of electrons to the group with the higher minimum energy as a result of thermal excitation, as outlined in the foregoing [14].

In view of recent indications that indium phosphide may also offer prospects for use in microwave oscillators [15], we have investigated the v-E characteristic of N-type InP as well. The results are shown in *fig. 7*. Curve a was obtained from measurements of the wavelength of electric space-charge waves propagating at electron drift velocity [16], curve b was measured with the microwave technique mentioned above [17]. The dashed curve is the result of a calculation [18], included for comparison.

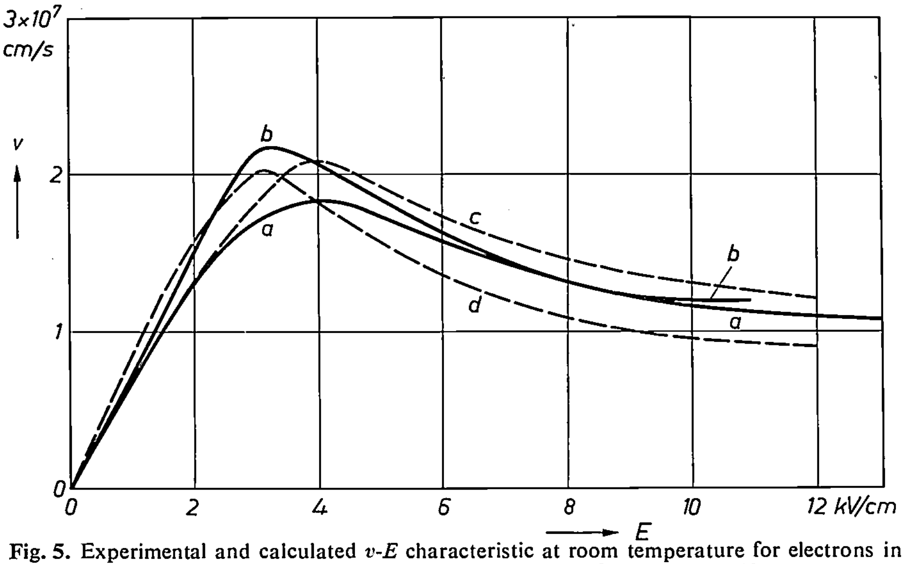


Fig. 5. Experimental and calculated v-E characteristic at room temperature for electrons in gallium arsenide. Curve a relates to microwave measurements [8], curve b to drift-time measurements [10]. Curve c was calculated on the assumption that the velocity distribution of the electrons is a "displaced" Maxwell distribution [11]. Curve d is a Monte-Carlo calculation [12].

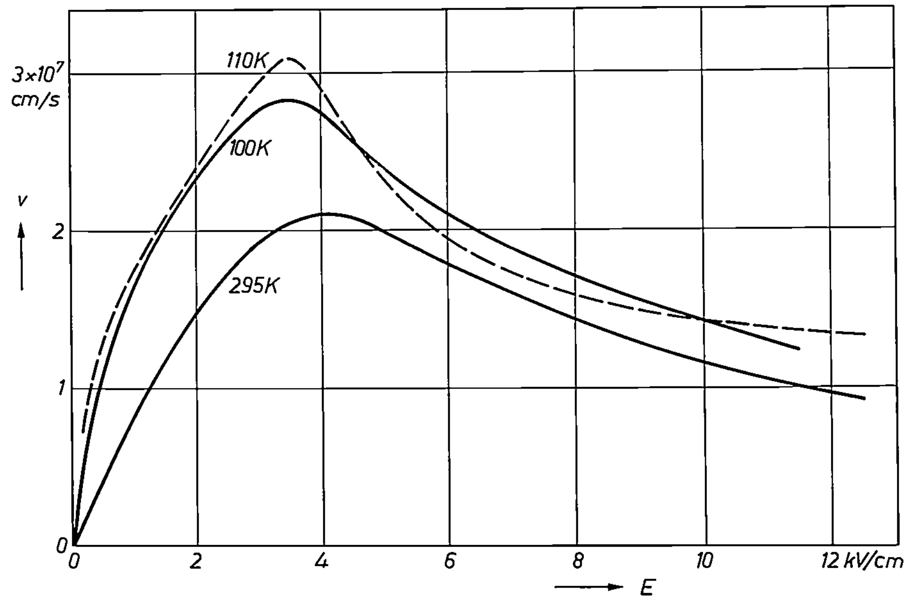


Fig. 6. As fig. 5, but now for different temperatures [13]. The two solid curves give experimental results obtained from microwave measurements at 100 K and 295 K. The dashed curve presents theoretical results obtained by W. Heinle for 110 K.

We have so far tacitly assumed that the average drift velocity of the electrons depends only on the strength of the electric field, implying that even at microwave frequencies the distribution of the electrons among both groups will adjust itself infinitely quickly to

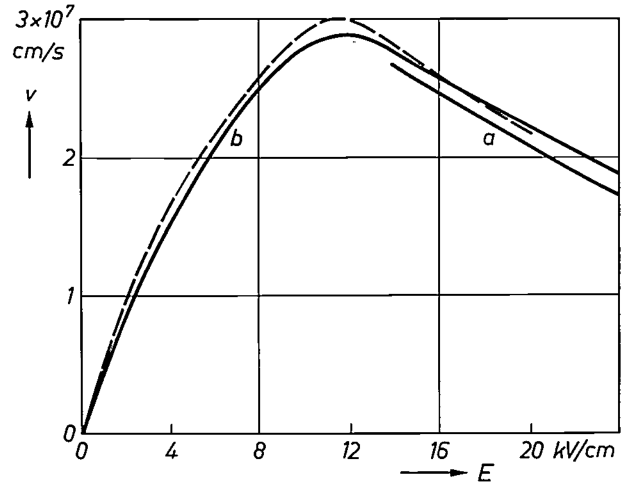


Fig. 7. Experimental (solid) and calculated (dashed) v-E characteristics of N-type indium phosphide at room temperature. Curve a was obtained from wavelength measurements on space-charge waves [16], curve b from microwave measurements [17].

variations of the field. This is only a good approximation to the reality if the time taken to "heat" the electrons in the group with the lower minimum energy and the time they take to transfer to the other group are both short compared with the period of the r.f. field. This is the case at frequencies below about 100 GHz. Calculations by H. D. Rees [19] have shown that the time constant describing the heating of the electrons — called the energy relaxation time — is the longer of the two; from experiments [20] this relaxation time has been found to have values of about 1×10^{-12} s. At frequencies a great deal lower than 100 GHz the oscillations can thus be described with a "static" v-E characteristic. In what follows we shall be concerned with oscillations in a Gunn diode incorporated in a microwave circuit.

Oscillations of a Gunn diode in a microwave circuit

A diode to which a constant voltage is applied, and which gives Gunn oscillations as described above, is not suitable for practical application as an oscillator. The frequency of the oscillations is determined by the transit time and is not easy to change. Moreover, at constant voltage it is impossible to take off microwave energy from the oscillator. These disadvantages can be overcome by incorporating the Gunn diode in a microwave circuit, enabling the voltage across the device to vary periodically with time. In a circuit of this type various modes of oscillation can occur. Before discussing these modes it will be useful to take a closer look at the behaviour of a Gunn diode in which a domain travels from cathode to anode. If we assume that the domain travels through the gallium-arsenide sample unchanged and at constant velocity, we can derive a relation between the maximum field-strength in the domain E_m and the field-strength E_0 outside it. Given the value of E_0 in the current-density field-strength characteristic in *fig. 8*, we obtain E_m by taking the hatched areas as equal. This relation is known as Butcher's rule [21]. If in addition the donor concentration n_d of the gallium arsenide is known and the shape of the domain is assumed to be simple, it is possible to construct a current-voltage characteristic which is applicable during the presence of a domain. This is referred to as the external characteristic (*fig. 9*). A domain can form when the voltage across the diode is above the critical value V_{th}; a domain already formed, however, may continue to exist and travel through the sample at a lower voltage, provided it is higher than the value V_{min}.

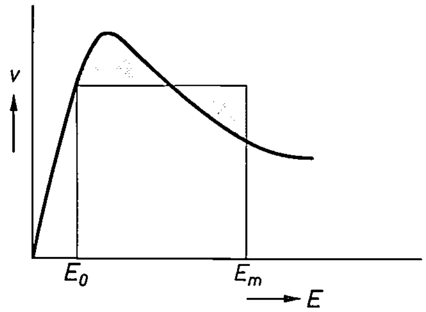


Fig. 8. At a given field-strength E_0 outside a domain the maximum field-strength E_m inside a domain is found by determining the value of E at which the shaded areas in the v-E characteristic are equal (Butcher's rule).

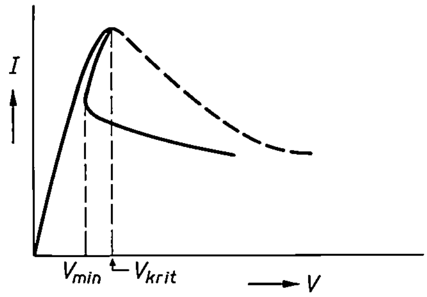


Fig. 9. External characteristic, showing the relation between the current I and the voltage V across a Gunn diode in which a domain is present. A domain can form as soon as V is greater than the value V_{th}. Once formed, a domain can also remain present at a lower voltage, provided it is not lower than V_{min}.

[14] For calculations see J. G. Ruch and W. Fawcett, J. appl. Phys. **41**, 3843, 1970.
[15] C. Hilsum and H. D. Rees, Electronics Letters **6**, 277, 1970.
[16] P. M. Boers, Electronics Letters **7**, 625, 1971.
[17] H. 't Lam and G. A. Acket, Electronics Letters **7**, 722, 1971.
[18] L. W. James, J. P. Van Dyke, F. Herman and D. M. Chang, Phys. Rev. B **1**, 3998, 1970.
[19] H. D. Rees, Solid State Comm. **7**, 267, 1969.
[20] M. T. Vlaardingerbroek, P. M. Boers and G. A. Acket, Philips Res. Repts. **24**, 379, 1969.
[21] P. N. Butcher, Physics Letters **19**, 546, 1965.

Butcher's rule may be derived as follows. At a given total current density J_{tot} the field-strength E and the electron concentration n as a function of the time t and the space coordinate x are given by Poisson's equation and the current equation. Poisson's equation is

$$\frac{\partial E}{\partial x} = \frac{e}{\varepsilon}(n - n_{\mathrm{d}}),$$

where n_{d} is the doping concentration and ε the dielectric constant. The current equation is

$$J_{\mathrm{tot}} = nev(E) - De\,\frac{\partial n}{\partial x} + \varepsilon\,\frac{\partial E}{\partial t},$$

where v is the average velocity of the electrons and D the diffusion constant, assumed to be independent of the field-strength.

A solution of these equations is a domain propagating unchanged and at a velocity v_{d}, when E and n are functions of a coordinate u which is equal to $x - v_{\mathrm{d}}t$. The equations can then be rewritten as:

$$\frac{\mathrm{d}E}{\mathrm{d}u} = \frac{e}{\varepsilon}(n - n_{\mathrm{d}}),$$

$$J_{\mathrm{tot}} = nev(E) - De\,\frac{\mathrm{d}n}{\mathrm{d}u} - \varepsilon v_{\mathrm{d}}\,\frac{\mathrm{d}E}{\mathrm{d}u}\,. \qquad (1)$$

From these equations we eliminate the coordinate u. Putting J_{tot} equal to $n_{\mathrm{d}}ev_{\mathrm{s}}$, where v_{s} is the velocity of the electrons outside the domain, we obtain after some manipulation:

$$\left(1 - \frac{n_{\mathrm{d}}}{n}\right)\mathrm{d}\left(\frac{n}{n_{\mathrm{d}}}\right) = \frac{\varepsilon}{Den_{\mathrm{d}}}\left[v - v_{\mathrm{d}} + \frac{n_{\mathrm{d}}}{n}(v_{\mathrm{d}} - v_{\mathrm{s}})\right]\mathrm{d}E. \qquad (2)$$

Integration of this equation is possible if we take the limits as a point *outside* the domain — where $n = n_{\mathrm{d}}$, $E = E_0$ and $J_{\mathrm{tot}} = n_{\mathrm{d}}ev_{\mathrm{s}}(E_0)$ — and a point (n,E) inside it. We then find:

$$\frac{n}{n_{\mathrm{d}}} - \ln\left(\frac{n}{n_{\mathrm{d}}}\right) - 1 = \frac{\varepsilon}{Den_{\mathrm{d}}}\int_{E_0}^{E}\left\{v - v_{\mathrm{d}} + \frac{n_{\mathrm{d}}}{n}(v_{\mathrm{d}} - v_{\mathrm{s}})\right\}\mathrm{d}E. \qquad (3)$$

We now take as the limit inside the domain the point where the field-strength has the maximum value E_{m}; at that point $\mathrm{d}E/\mathrm{d}u$ is equal to zero, and in accordance with equation (1), n is therefore equal to n_{d}. Equation (3) then becomes:

$$\int_{E_0}^{E_{\mathrm{m}}}\left\{v - v_{\mathrm{d}} + \frac{n_{\mathrm{d}}}{n}(v_{\mathrm{d}} - v_{\mathrm{s}})\right\}\mathrm{d}E = 0.$$

The integration path can still be chosen, since the field-strength at both ends of the domain has the value E_0. For both possibilities the form of $n(E)$ is different, but the integral must nevertheless have the same value. This is only possible if $v_{\mathrm{d}} = v_{\mathrm{s}}$. The relation then becomes:

$$\int_{E_0}^{E_{\mathrm{m}}}(v - v_{\mathrm{s}})\mathrm{d}E = 0,$$

thus proving Butcher's rule, which states that the velocity of the carriers averaged over the field is equal inside the domain to that outside it.

Three modes of oscillation

If in addition to the d.c. voltage we now apply to the Gunn diode an a.c. voltage with a frequency lower than that corresponding to the transit time of a domain, we find the following situation (*fig. 10*). Suppose that at a certain moment the total voltage reaches the value V_{th} and the field-strength reaches the critical value E_{th} so that a domain forms (point A). The current then decreases to a value given by the external characteristic (A'). When the domain arrives at the anode after a time τ, the current again increases (C-C'). The voltage is then too low for a new domain to form, which only happens when the voltage again rises above V_{th}. The current is therefore strongly non-sinusoidal. An important property of the current is that its fundamental component is in the opposite phase to the voltage, which means that when the Gunn diode is incorporated in a tuned circuit (*fig. 11*) the oscillations are self-sustaining.

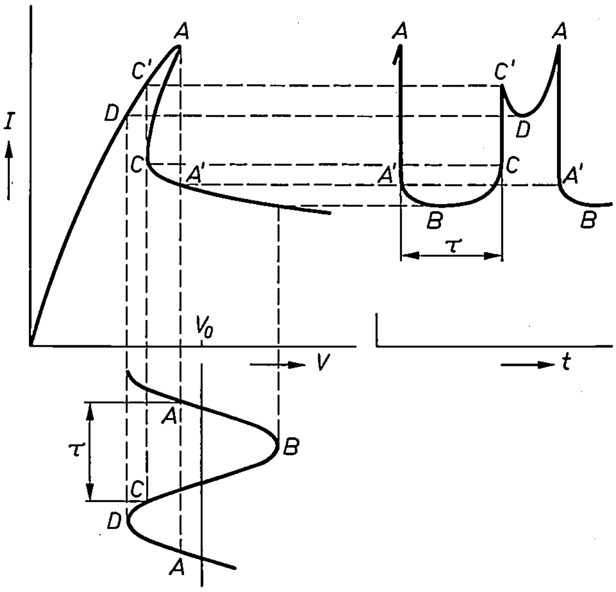


Fig. 10. Variation with time t of the current I in a Gunn diode as a result of an applied d.c. voltage V_0 with an r.f. voltage superimposed on it, whose period is greater than the transit time τ of the charge carriers. The I-t curve (top right) can be derived from the applied voltage (bottom left) by means of the external characteristic (top left).

In another, entirely different type of mode the frequency of the applied r.f. voltage is so high that its cycle is shorter than the transit time. The domains then disappear before they reach the anode, and the frequency of the oscillations can now be adjusted with the microwave circuit. The disadvantage of this mode is that the conversion of d.c. current into microwave power is less effective, since the part of the specimen at the anode end, which is not traversed by the domains, behaves as an additional resistance in the circuit. The adverse effect

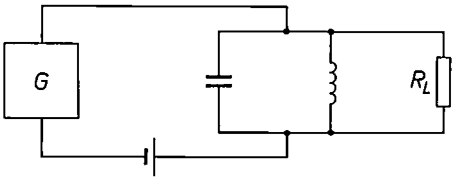


Fig. 11. Gunn diode G in a tuned circuit with LC circuit and load resistance R_{L}.

of this is greater the more the frequency differs from the transit-time frequency. The length of the diode must therefore be chosen in such a way that the transit-time frequency does not differ too much from the desired oscillation frequency. When this requirement is met, the length of the diode for high frequencies is small and the microwave power is low.

This disadvantage, which is inherent in transit-time oscillators, can be overcome by using a third mode of oscillation, known as the LSA mode [22] (LSA standing for Limited Space-charge Accumulation). This mode can occur when the applied d.c. voltage (V_0 in fig. 9) is appreciably higher than V_{th} — for example two or three times higher — and when at the same time the amplitude of the r.f. signal is so high that the instantaneous value of the total voltage nevertheless drops below the value V_{th} for a short time during each cycle. The most fundamental conditions, however, are that the time in which an appreciable space charge (accumulation layer) can form must be longer than the cycle — or at least not much shorter — but that on the other hand the space charge present must disappear rapidly as soon as the instantaneous value of the voltage has dropped below V_{th}. Because of this second requirement, the space charge can never reach a high value, even after a large number of cycles.

It is possible to show that these requirements can only be met if the ratio of the doping concentration to the frequency remains within specific limits, which differ roughly by a factor of 10.

The advantage of the LSA mode is that, because of the relatively slow growth of the space charge, the field has the same strength almost everywhere in the diode and is greater than E_{th}. Furthermore, since the length of the GaAs sample has nothing to do with the frequency and may in principle be freely chosen, a high power output is possible in pulsed operation.

The LSA mode is only a practical proposition in extremely homogeneous material. Inhomogeneities in the donor concentration cause inhomogeneities in the electron concentration, as a result of which the field must remain longer below the critical value for the space charge to disappear than would be necessary if there were no inhomogeneities. This has the result that the effective negative slope of the characteristic decreases so greatly that LSA oscillations are no longer possible at concentration variations of about 10%.

Circuit

If the unit formed by a Gunn diode and a microwave circuit is to be able to oscillate, the two impedances must be equal and opposite. The real part of the impedance of a Gunn diode oscillating in a particular mode is negative, and in many cases the imaginary part is also negative. A complication is that the impedance cannot simply be derived from current and voltage at the oscillation frequency, but that harmonics also have to be taken into account for non-sinusoidal oscillations.

A microwave circuit widely used in our laboratories is the coaxial type as shown in *fig. 12*. The bandwidth of this circuit is very large, which implies that the frequency of the electromagnetic signals capable of propagating in it can be varied in a wide region without changing the configuration of the electromagnetic field perpendicular to the direction of propagation. This makes it possible to experiment with Gunn oscillators for different frequencies in the same circuit.

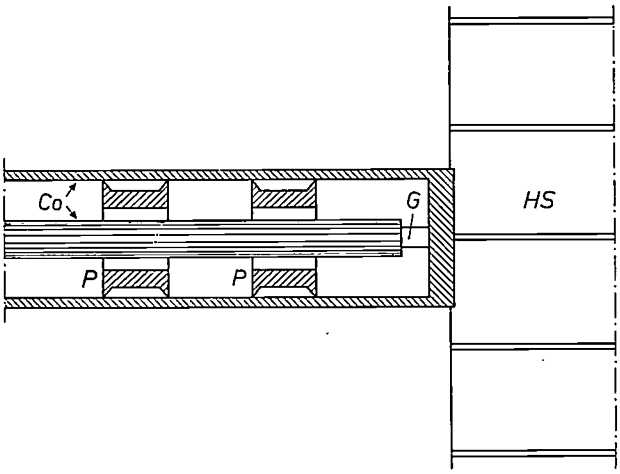


Fig. 12. Microwave circuit in which the Gunn diodes were investigated (schematic). The diode *G*, in its encapsulation, is located at the end of a coaxial line *Co*. The impedance required for resonance is obtained by suitable selection of the dimensions and of the position of the plungers *P*. The required d.c. voltage is applied through the inner and outer conductors of the coaxial line. The r.f. power is taken off at the other end of the circuit (left). On the right there is a copper block with cooling fins.

An impedance equal and opposite to that of the Gunn diode is obtained by means of metal plungers that reflect part of the electromagnetic wave. If the impedance of the Gunn diode is unknown, the correct dimensions and positions for the plungers can be determined experimentally. It is fairly simple to calculate them, however, if the impedance is known from measurements or has been established with the aid of a computer simulation [23].

The fabrication of Gunn diodes of gallium arsenide

The efficiency with which c.w. microwave power can be generated with a Gunn diode of gallium arsenide is somewhere between 3 and 10%. Most of the electrical power supplied to the diode is thus converted into heat. For c.w. operation of the diode this heat must be

[22] J. A. Copeland, Proc. IEEE 54, 1479, 1966.
[23] See the article by De Groot and Mircea [6].

dissipated. This means that the electrical contacts of the Gunn diode must meet two requirements: they must have a high thermal conductivity and they must not produce heat themselves, which means that their electrical resistance must be small.

A junction virtually without resistance between metal and GaAs is only possible if the semiconductor is very strongly doped [24]. The material for Gunn diodes has a doping concentration of about 10^{15} carriers per cm^3, which is not enough for contacts. A more strongly doped junction region must therefore be made to which the metal contact can be applied.

The distance between the contacts on a Gunn diode for generating microwave power at a frequency of 5 GHz is 20 μm. Since it would be extremely difficult to make and process a monocrystalline layer of GaAs as thin as this, an epitaxial layer 20 μm thick is grown on a substrate of strongly doped *N*-type GaAs which is much thicker, e.g. 100 μm [25]. The substrate has a negligible electrical series resistance and is therefore an excellent contact in an electrical respect. Thermally, however, it is not, since gallium arsenide is a poor heat conductor. The second contact must therefore transmit almost all of the heat away, and should therefore be no more than half a micron thick. Epitaxial growth in this case would not be impossible but it would certainly be difficult, since good layers thinner than 2 or 3 microns cannot be made, and it would therefore be necessary to etch away a part of the product to obtain a layer of the required thickness. We have therefore used a different process, which can also be used for making other devices.

It begins, as outlined above, with the epitaxial growth of a monocrystalline layer 10 to 20 microns thick on a strongly doped substrate [26]. A mixture of tin and silver is then deposited at the two ends of the slice by vacuum evaporation. To this metal layer a layer of SiO_2 is applied (0.25 μm) by the pyrolysis of SiH_4 in an oxygen atmosphere at 400 °C. The SiO_2 layer prevents the evaporation of arsenic during the subsequent steps of the process, and it also keeps the tin-silver layer flat during the next step, which is alloying. This is done at 550 °C in a hydrogen atmosphere. For this purpose the gallium-arsenide slice is placed on the heating table with the epitaxial layer turned face downwards. During this step the tin-silver layer melts, and at the highest temperature reached some GaAs dissolves in the molten metal. The further process takes place during cooling, when it is essential to have a temperature gradient in the slice. The side lying face down on the table is hotter than the side exposed to the flow of hydrogen. Consequently the GaAs dissolved in the lowest metal layer settles at the top of the epitaxial GaAs layer, thus forming on this layer a GaAs film which is very strongly doped with metal — mainly tin. On the other side of the slice, the substrate side, the situation is exactly the opposite. This is not a disadvantage, however, since the substrate is in any case heavily doped and therefore requires no extra layer to establish good electrical contact with it.

Once everything has completely cooled down, the SiO_2 is first dissolved, and then the tin-silver mixture is removed by dissolving it in mercury. Next, a metal is deposited by vacuum evaporation on the heavily doped layers, giving contacts with a resistivity of no more than about 10^{-4} Ω/cm^2. On the vacuum-evaporated metal layer — often titanium — a gold layer is then deposited, and finally, by means of photoresist and chemical etching, diodes with a diameter of approximately 200 microns are etched out of the slice.

Each diode is now placed in a gold-plated copper encapsulation, with the epitaxial layer facing the copper. The whole structure is heated to 300 °C and subjected for a few seconds to a pressure of 400 kg/cm^2. This produces a very good connection between the gold of the contact and the gold of the encapsulation. At the same time a gold wire is pressed on to the upper surface, which also adheres firmly to the gold of the surface and thus forms a good electrical connection (*fig. 13*).

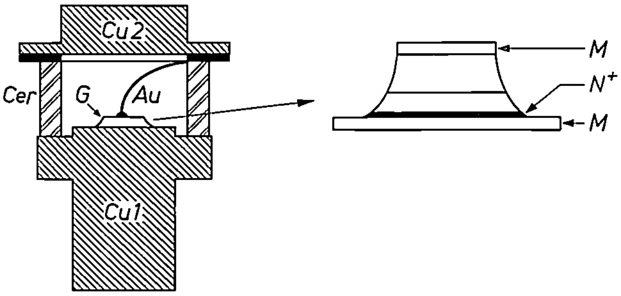


Fig. 13. Section of a Gunn diode in its encapsulation. *G* diode. *Cu1* copper block for heat dissipation. *Cer* ceramic ring providing insulation between the contact electrodes *Cu1* and *Cu2*. *Au* gold wire. The diode (see inset) is mounted on the copper block with the active layer and the thin contact layer N^+ turned face downwards. *M* gold layers providing the contact with *Au* and *Cu1*.

Finally we should note that highly doped layers, as described above, do not necessarily have to be made with tin-silver alloys. Any metal in which gallium arsenide dissolves at not too high a temperature, and to which a doping material can be added, is in principle suitable.

[24] C. A. Mead, Physics of interfaces, in: Symp. Ohmic contacts to semiconductors, Montreal 1968, pp. 3-16.

[25] D. V. Eddolls, J. R. Knight and B. L. H. Wilson, Proc. Int. Symp. on gallium arsenide, Reading 1966, p. 3.

[26] The material used by us was prepared by B. C. Easton and Miss C. Fisher of Mullard Research Laboratories, Redhill, Surrey, England.

Performance

To conclude this article we shall give two representative examples of the results obtained with Gunn diodes of the type described in the microwave circuit mentioned earlier. The oscillation frequencies of the oscillators were 5.0 and 5.94 GHz. *Fig. 14* gives the efficiency and the microwave power of the 5 GHz diode as a function of the applied d.c. voltage V_0, measured with the diode in continuous operation. The measurements on the other diode were carried out during pulsed operation, and are collected in *fig. 15*. The performance in both cases was very satisfactory.

Since more power can be dissipated in the diode during pulsed operation than in c.w. operation, the maximum power in the first case is of course higher. The efficiency is also higher in pulsed operation. This is because the temperature of a diode in c.w. operation is considerably higher than room temperature (about 300 °C), whereas there is scarcely any increase in the temperature of a pulsed diode. We have seen above that the drop in the v-E characteristic is much less marked at a higher operating temperature.

The temperature increase of the c.w. operated diode also partly accounts for the fact that the output power, after having reached a maximum, decreases again with increasing d.c. voltage. Another reason for the falling characteristics at high V_0 is the fact that the resonant frequency and the impedance of the coaxial system are fixed, whereas the oscillation frequency and impedance of the diode vary slightly with V_0. This means that the match between diode and waveguide is not ideal at every voltage.

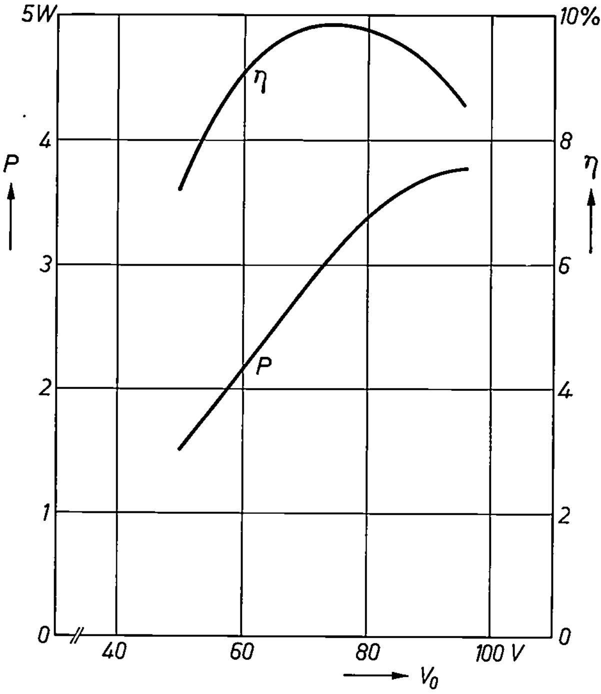


Fig. 15. As fig. 14, but now for pulsed operation at 5.94 GHz. In this case P does not decrease at high V_0, and the efficiency is greater than that of the c.w. diode in fig. 14 because there is now no appreciable increase of temperature.

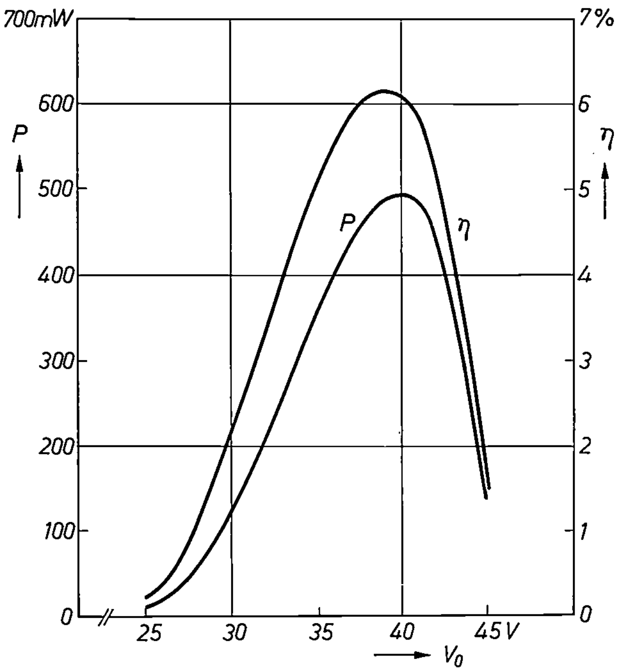


Fig. 14. Output power P and efficiency η of a c.w. Gunn diode at 5 GHz as a function of the applied d.c. voltage V_0. At high V_0 values, P and η decrease with increasing V_0 owing to the excessive heating of the diode and the deterioration of the match between microwave circuit and diode.

Summary. The (Watkins-)Gunn effect is a high-frequency instability occurring in certain semiconductors subjected to a high electric field (e.g. N-type gallium arsenide and indium phosphide) in which the average electron velocity decreases with increasing field in a particular range of field values. The effect is due to the formation of domains of very high field-strength, which travel from cathode to anode in a diode. In general the period of the Gunn oscillation is equal to the transit time of the domains.

A Gunn diode incorporated in a microwave circuit and subjected to an r.f. signal superimposed on the d.c. voltage is able to oscillate in various modes, and the oscillation frequency does not have to be exactly the same as the transit-time frequency.

A technological problem is the fabrication of contacts that have negligible electrical resistance and are able to conduct away the heat generated effectively. Almost all of the heat produced has to be removed via one of the two contacts. This consists of a highly doped layer formed from a layer of molten Sn-Ag in which GaAs is dissolved. A power output of 0.5 W and an efficiency of 6% have been obtained with diodes in c.w. operation, and values of 7.5 W and 10% at 6 GHz with pulsed operation.

Epitaxial growth of gallium arsenide

A. Boucher and B. C. Easton

Gallium arsenide has a special place among the semiconductors used in solid-state microwave electronics. Its electron mobility and band gap are high compared with those of silicon or germanium. By introducing suitable dopant impurities into the material a wide range of electrical conductivity values can be attained. These facts, together with the discovery of the Gunn or transferred-electron effect [1] that occurs in gallium arsenide because of the particular band structure, have stimulated the interest in this material.

Like many semiconductor devices and integrated circuits, gallium-arsenide devices are constructed with their active regions fabricated within a layer grown epitaxially on a monocrystalline substrate of the same material. The electrical properties of the substrate can be selected by suitable doping, so that it can act as either a conducting or an insulating mechanical support. The required properties in the epitaxial layer are also obtained by doping with an appropriate impurity either during growth or by subsequent diffusion. Epitaxial layer thicknesses may range from less than one micron to many tens of microns depending on the particular device.

Gallium-arsenide microwave devices, in which epitaxial material is used, include transferred-electron oscillators, varactor, mixer, tuning and avalanche diodes and field-effect transistors.

The majority of applications require thin N-type layers on highly conducting N^+ gallium-arsenide substrates. Sometimes an N^+ layer is needed on top of such an N layer to provide an ohmic semiconductor contact to the N-type material. Other applications require a number of successive layers with different carrier concentrations, e.g. N^+ substrate, N layer, N^+ layer or a series of alternate N and N^+ layers. This requirement and the need for localized epitaxial deposition for discrete-device or integrated-circuit applications demand a thorough control of all aspects of the epitaxial-growth process. Investigations with this objective have been made at the Laboratoires d'Electronique et de Physique Appliquée (LEP) on some of the fundamental aspects of epitaxial growth (thermodynamics, kinetics and anisotropic effects) [2] [3] and also on new methods which can yield high-performance devices [4]. At Mullard Research Laboratories (MRL) the incorporation and identification of impurities is being studied, and work is also being done on the automation of the epitaxial-growth process. Impurities are identified by electrical characterization and scanning mass-spectrographic techniques [5].

For some applications, e.g. field-effect transistors, semi-insulating gallium arsenide (resistivity higher than 10^7 Ωcm) is needed as an insulating substrate. High-purity gallium arsenide, in which the impurity conduction is due to shallow donor levels, may be given a higher resistivity by creating deep traps by doping the material with chromium.

In recent years the major emphasis in the study of the epitaxial growth of gallium arsenide has been on the preparation of material suitable for Gunn-oscillator devices. Epitaxial layers for this purpose must be of relatively high purity with a free-donor concentration in the region of 10^{15} cm^{-3}, combined with a high electron mobility (about 7500 cm^2/Vs at 293 K).

Two general methods of preparing high-purity epitaxial gallium arsenide are available; the first is by chemical vapour deposition and the second by solution growth from a gallium melt. We have investigated both of these techniques; details will be given and the methods used for characterization of the results by means of electrical and optical techniques will be reviewed.

Epitaxial growth of gallium arsenide

Vapour growth

Most chemical vapour transport reactions for the preparation of epitaxial gallium arsenide involve reactions between gallium, arsenic and hydrogen chloride. These reactants may be derived from the elements or their compounds; the particular combination chosen has strong influence on the purity of the layers produced. D. Effer [6] was the first to use the reactions between hydrogen, arsenic trichloride and gallium to produce layers of a higher purity than any previously prepared by other processes. This has been the method most commonly used for the growth of epitaxial layers of suitable purity for Gunn-device manufacture.

Dr. A. Boucher, formerly with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brevannes, Val-de-Marne, France is now with La Radiotechnique Compelec, S.A., Suresnes, France; B. C. Easton, M.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

Fig. 1 shows a schematic diagram of an epitaxial reactor. High-purity hydrogen from a palladium diffuser unit is passed through an arsenic-trichloride saturator and into a silica reaction tube in a two-zone furnace. Arsenic trichloride is initially reduced in the reaction tube to form arsenic vapour and hydrogen chloride; these reaction products then proceed down the tube to the gallium source (99.9999% purity) contained in a silica boat at 800-850 °C. At this stage arsenic dissolves in the gallium until it is saturated and a solid skin of gallium arsenide is formed on its surface. At the same time the hydrogen chloride reacts with gallium, mainly to form the monochloride. When the source is saturated the gaseous reaction products (principally gallium monochloride, arsenic and hydrogen) pass into the second zone, kept at 750 °C, where deposition occurs on the gallium-arsenide substrate, which has been polished chemically in a bromine-methanol solution. No growth occurs on the surrounding silica ware at this temperature, but free arsenic, gallium arsenide and other by-products of the reaction may be deposited on the wall at the end of the reaction tube. A liner is used there for ease of cleaning.

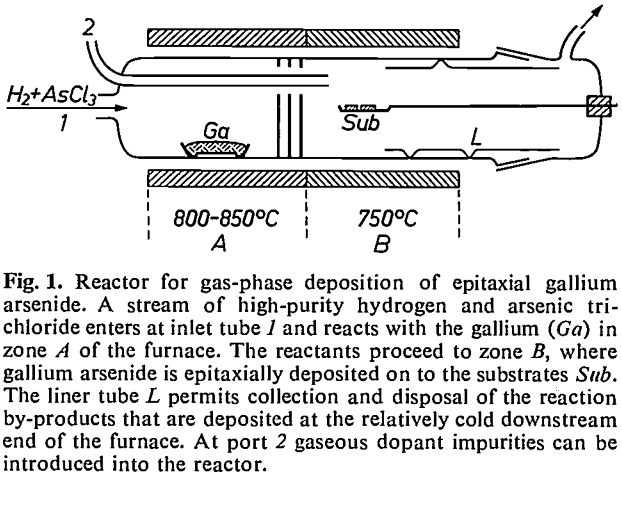


Fig. 1. Reactor for gas-phase deposition of epitaxial gallium arsenide. A stream of high-purity hydrogen and arsenic trichloride enters at inlet tube *1* and reacts with the gallium (*Ga*) in zone *A* of the furnace. The reactants proceed to zone *B*, where gallium arsenide is epitaxially deposited on to the substrates *Sub*. The liner tube *L* permits collection and disposal of the reaction by-products that are deposited at the relatively cold downstream end of the furnace. At port *2* gaseous dopant impurities can be introduced into the reactor.

The principal chemical reactions which occur are summarized as follows:

$$4\,AsCl_3 + 6\,H_2 \rightarrow As_4^{\nearrow} + 12\,HCl$$

$$2\,Ga + 2\,HCl \rightarrow 2\,GaCl^{\nearrow} + H_2^{\nearrow}.$$

The overall equilibrium for the deposition reaction in the temperature range 730-830 °C has been stated as [2]

$$2\,GaCl + \tfrac{1}{2}\,As_4 + H_2 \rightleftarrows 2\,GaAs + 2\,HCl.$$

Layers for Gunn devices are generally grown on substrates cut on a surface 2-3° off a {100} crystal plane. This orientation facilitates dicing and produces lower electrical impurity layers than on {110} or {111} surfaces. Chromium-doped semi-insulating or highly conducting *N*-type substrates, where the dopant may be silicon, tin or tellurium are used. Monitor layers grown on semi-insulating substrates are used for Hall measurements so that resistivity and carrier concentration may be rapidly assessed for each growth experiment. It has been found that these measurements relate closely to the properties of layers grown at the same time on highly doped substrates. Growth rates are in the region of 20-30 microns an hour and layers from 1 to over 100 microns thick may be grown.

The background impurities in an epitaxial system are generally predominantly donors, and free-carrier concentrations down to 10^{14} cm^{-3} may be obtained. The most difficult impurity to eliminate or control during growth is oxygen (usually introduced into the system as water). Its influence on the electrical properties of epitaxial layers is mainly indirect and complex; not only can it affect the reactions controlling deposition but it can react with other impurity atoms in the system. At high levels it has a harmful effect on the crystal quality of the layers.

Most devices require epitaxial material with free-donor concentrations of 10^{15} cm^{-3} or more; the simplest method of doping during growth is to add tin (either as tin-doped gallium arsenide or free metal) to the gallium source, but this approach has the disadvantage that it is inflexible. Doping systems which permit the doping level to be varied, either between growth experiments or during the deposition, rely on the introduction into the gas stream of a gaseous donor impurity such as hydrogen sulphide or selenide between the gallium source and the substrate (see fig. 1). A more recent method developed at LEP, the back-diffusion method, is illustrated in *fig. 2*. It is used to produce tin-doped epitaxial layers, and permits changes in dopant concentration over a wide range during the process.

The main stream of gaseous reaction products is the same as in the process given above. Here however a side stream of hydrogen is introduced into the system, and this hydrogen flows over a boat containing elemental tin. Hydrogen chloride from the main stream diffuses

[1] This effect is treated in detail in the paper by G. A. Acket, R. Tijburg and P. J. de Waard in this issue, page 370.

[2] A. Boucher and L. Hollan, J. Electrochem. Soc. **117**, 932, 1970.

[3] L. Hollan and C. Schiller, J. Crystal Growth **13/14**, 319, 1972.

[4] L. Hollan, J. Hallais and C. Schiller, J. Crystal Growth **9**, 165, 1971, and A. Boucher, J. P. Chané and E. Fabre, Rev. Physique appl. 6, 5, 1971.

[5] J. B. Clegg, E. J. Millett and J. A. Roberts, Anal. Chem. **42**, 713, 1970.

[6] D. Effer, J. Electrochem. Soc. **112**, 1020, 1965.

back into the hydrogen flow and reacts with the tin, which is kept at a temperature of 750 °C. In this way tin is introduced into the system as a chloride. The tin-chloride content is controlled through the flow rate of the hydrogen. Using such techniques layers may be doped in the range 10^{15}-10^{18} free donors cm^{-3}, and epitaxial structures built up of high- and low-doped regions can also be obtained (*fig. 3*).

At MRL an improved epitaxial system has been developed in which a more complex gas-handling system has been constructed, enabling the different growth parameters to be varied and their effect on the layer properties to be studied. It incorporates electrically operated solenoid valves. A programme-timer unit enables the growth process to proceed automatically. The new system provides better control and consistency in the growth process with reduced reliance on operator skill.

Development of the vapour growth has resulted in greater versatility, permitting doping control and automation of the various process operations. It can be argued that the arsenic-trichloride process is inherently unsuitable for precise control since true equilibrium is never attained at the gallium source. To overcome this difficulty alternative procedures for introducing arsenic into the system have been used, e.g. use of elemental arsenic or arsine. In general, however, the alternative processes do not produce layers of sufficiently high purity, largely because sufficiently pure starting materials are not available.

Liquid growth

Two methods have been used for epitaxial growth of GaAs from the liquid phase. Essentially they both depend upon the deposition of GaAs from a saturated solution in gallium. The first, described by H. Nelson [7], is the horizontal method and the second, reported by H. Rupprecht [8], is the vertical method.

E. André and J. M. Le Duc from RTC (La Radiotechnique-Compelec) at Caen have developed a special version of the horizontal method (*fig. 4*). Liquid gallium saturated with gallium arsenide in the range 650-900 °C is brought into contact with the substrate by pulling back a shutter at the bottom of the gallium container. Growth occurs on the substrate as the solution is allowed to cool down. Material has been produced with free-donor concentrations down to 10^{14} cm^{-3} and Hall mobilities in excess of 10^5 cm^2/Vs at 77 K [9].

A schematic diagram of the vertical system used at MRL is shown in *fig. 5*. The saturated gallium is contained in a silica crucible; the substrate is introduced into it from above, suspended from a silica rod, and the growth system is maintained under a pure hydrogen atmosphere. Again, under optimum conditions, layers with free-donor concentrations down to 10^{14} cm^{-3} have been obtained.

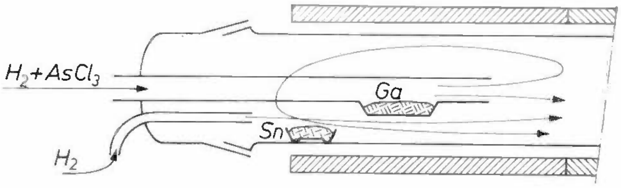


Fig. 2. Back-diffusion reactor for gas-phase deposition of tin-doped epitaxial gallium arsenide. High-purity hydrogen and arsenic trichloride, introduced through inlet *1*, react with the gallium (*Ga*). A proportion of the reactants diffuses upstream in the hydrogen flow, entering at inlet *2*, and reacts with the tin (*Sn*). From the reactant flow containing tin, gallium and arsenic, tin-doped gallium arsenide is deposited on substrates downstream in the furnace. Regulation of the hydrogen flow rate over the tin permits accurate control of the dopant content in the reactant mixture.

In both the horizontal and vertical systems the layer thickness is controlled by the temperature range over which deposition is allowed to occur; in this respect the vertical method permits better control. With cooling rates of 0.1 to 0.2 °C min^{-1} and well defined growth times, thinner layers can be grown by the vertical method than by the horizontal technique.

Tin can be added to the melt to provide additional doping. An advantage of the liquid-growth techniques is that more uniformly doped layers can be grown. Gunn-device assessment of these layers shows that standard *X*-band (about 10 GHz) oscillators prepared from liquid epitaxial material have d.c. to r.f. conversion efficiencies that compare favourably with the best results obtained when vapour-grown material is used.

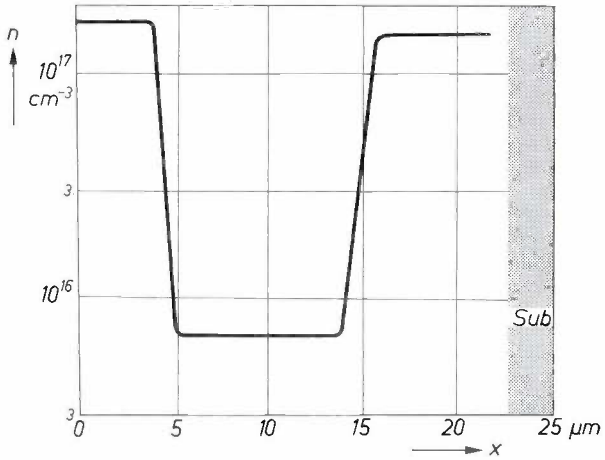


Fig. 3. Typical result of the back-diffusion method. The free-carrier concentration *n* is plotted as a function of the depth *x* below the surface of a multiple-layer structure. *Sub* substrate.

Characterization

Before an epitaxial layer is used in a microwave device it is desirable to determine its suitability by routine measurements. In practice close correlation of the material properties with the device behaviour may be largely masked by subsequent device-fabrication technologies, in particular when making ohmic contacts. The layer-assessment techniques employed at our laboratories are indicated in the following sections.

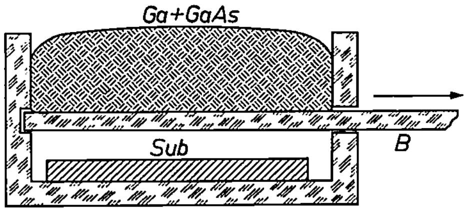


Fig. 4. Apparatus used in a modified version of the horizontal liquid-phase deposition method. The substrate *Sub* is located at the bottom of the sample holder. In the top compartment a charge of gallium is heated under pure hydrogen with an accurately determined quantity of gallium arsenide until the latter is dissolved. The shutter *B* at the bottom of the gallium compartment is then withdrawn and the solution spreads over the substrate. On slowly cooling the solution over a well defined temperature range, an epitaxial layer of accurately controlled thickness is deposited.

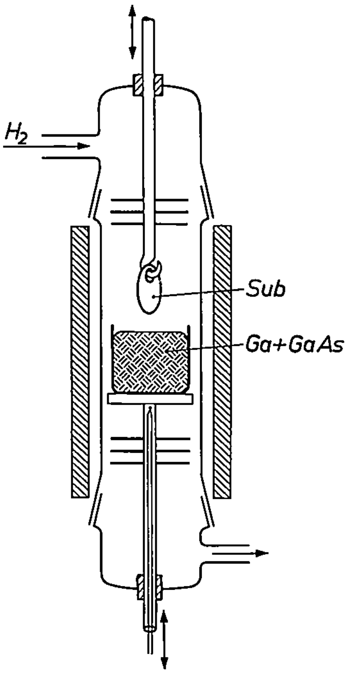


Fig. 5. The deposition of gallium arsenide using the vertical method. A crucible contains gallium arsenide dissolved in gallium. At the temperature at which the solution is saturated, the substrate *Sub* is immersed in the solution. When the solution cools down gallium arsenide is deposited on the substrate. Layer thickness is determined by the temperature range through which the melt is cooled. The system is flushed with pure hydrogen during the process.

Crystallographic and optical methods

Microscopic examination of a surface reveals surface defects, and metallographic staining of a polished or cleaved cross-section reveals the interface between layer and substrate. The perfection and uniformity of the substrate can be studied and the thickness of the layer measured directly. Infra-red interference techniques [10] can be used for measuring non-destructively the thickness of N-type layers on N^+ substrates.

X-ray-reflection topographic examination [11] of layers and substrates is used to show dislocation density, strain and damage arising from cutting or polishing operations; these factors are particularly important for determining the substrate quality required to produce layers with a low density of surface defects. C. Schiller [12] has recently extended the technique to examination of the layer-substrate interface region and has correlated cyrstallographic disturbances at the interface with substrate quality, polishing and pre-treatment before epitaxial growth.

Electrical properties

The Van der Pauw technique [13] for measurement of Hall constant and resistivity is used for layers grown on semi-insulating substrates, and gives values for resistivity, free-carrier concentration and mobility.

Variation of dopant concentration with depth in the epitaxial layer can be found from experiments on a metal-semiconductor diode (Schottky-barrier diode) obtained by evaporating a metal film on to the sample, or more conveniently by applying a mercury contact with a defined area.

The relation between reverse-bias voltage and diode capacity yields the impurity concentration at a depth determined by the bias [14]; the largest depth on which data can be obtained is the depletion-layer thickness at the breakdown voltage of the diode. *Fig. 6* shows results from these measurements. On applying this principle in conjunction with suitable measuring equipment the doping level can be plotted automatically and its uniformity through an epitaxial layer can be investigated. Hall and Schottky-barrier measurements are used as routine assessment procedures. They can be

[7] H. Nelson, RCA Rev. **24**, 603, 1963.

[8] H. Rupprecht, Proc. Int. Symp. on Gallium Arsenide, Reading 1966, page 57.

[9] E. André and J. M. Le Duc, Mat. Res. Bull. **3**, 1, 1968.

[10] P. J. Severin, Appl. Optics **9**, 2381, 1970 and Appl. Optics **11**, 691, 1972 (No. 3).

[11] J. B. Newkirk, J. appl. Phys. **29**, 995, 1958.

[12] C. Schiller, Solid-State Electronics **13**, 1163, 1970.

[13] L. J. van der Pauw, Philips Res. Repts. **13**, 1, 1958; see also Philips tech. Rev. **20**, 220, 1958/59.

[14] General theory has been given in: W. Schottky, Z. Physik **118**, 539, 1941/42. The particular application referred to has been described in: C. O. Thomas, D. Kahng and R. C. Manz, J. Electrochem. Soc. **109**, 1055, 1962.

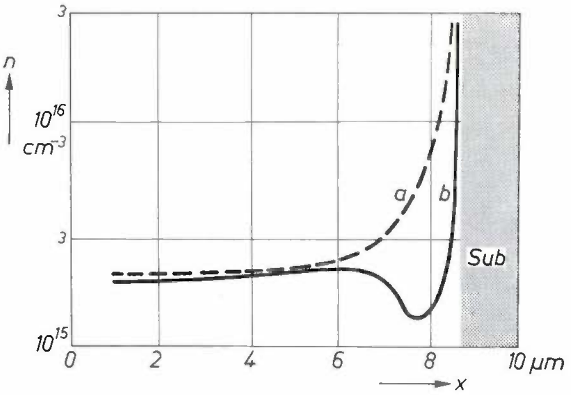


Fig. 6. Concentration *n* of dopant impurities (number of free donors cm^{-3}) as a function of the distance *x* beneath the surface of an epitaxial layer. *Sub* substrate. The curves have been obtained from Schottky-barrier diode measurements. Curve *a* is typical of a layer into which dopant impurity from the substrate has been incorporated during growth. Curve *b* is frequently obtained when silicon- or tin-doped substrates are used and the growth conditions are insufficiently well controlled.

extended by investigating the influence of temperature, and in the case of capacitance measurements, the effect of infra-red radiation. Such studies provide further information [15] about the nature and concentration of the electrically active centres, i.e. donors, acceptors, and traps in the epitaxial layer. Other techniques such as those employing photo- or cathodoluminescence [16] and magnetoresistance are used to complement the Hall and Schottky measurements.

As far as possible it is the aim of the electrical characterization to correlate the results obtained, on one hand with the concentration and nature of impurity atoms, divergence from stoichiometry and crystallographic defects in the epitaxial layer, and on the other with the microwave performance of devices made from particular samples of material.

Finally, although gallium arsenide is the principal compound semiconductor for microwave applications, other III/V compounds with theoretical advantages are being studied at MRL and LEP. Particular examples are indium phosphide and various compositions of the indium-phosphide/gallium-arsenide alloy system. Such materials require techniques for their preparation similar to those established for gallium arsenide, and preliminary studies of the various growth systems have been commenced for these new materials.

Summary. Processes for the epitaxial deposition of gallium-arsenide layers for use in microwave electronics, are under investigation at Mullard Research Laboratories (Redhill), at the Laboratoires d'Electronique et de Physique Appliquée (Limeil-Brévannes) as well as in some other laboratories within the Philips group of companies. Epitaxial layers are deposited both from gas-phase reactions between gallium and arsenic compounds and from a saturated solution of GaAs in liquid gallium. The layers obtained are studied by means of X-ray diffraction, infra-red interference cathodo- or photoluminescence and electrical measurements (conductivity, Hall effect, Schottky-barrier diode measurements and magnetoresistance).

[15] H. I. Ralph and F. D. Hughes, Solid State Comm. **9**, 1477, 1971 (No. 17), and G. A. Acket, Philips Res. Repts. **26**, 261, 1971 (No. 4).

[16] C. M. Wolfe and G. E. Stillman, Proc. 3rd Int. Symp. on Gallium Arsenide and Related Compounds, Aachen 1970, page 3.

Computer calculations of the Gunn effect

J. de Groot and A. Mircea

Since the time of Galileo it has been generally accepted in the physical sciences that the proper understanding of an effect depends on good agreement between experimental results and the results of calculations based on a mathematical model of the situation. If this agreement exists, we may conclude that the model within its limitations is a fair representation of the reality. The model is used not only for making an optimum choice of experimental conditions, but more particularly for obtaining information about quantities that are not directly measurable.

This also applies of course to effects that can be observed in semiconductors. The model may clarify how internal quantities of the material, such as the electric field and space-charge density, vary with space and time, and how they are related to externally measurable quantities such as current and voltage. Model calculations can reveal, for example, the way in which a semiconductor device interacts with the circuit in which it operates and how the circuit design may be optimized.

In this article we shall describe numerical calculations based on a mathematical model of the Gunn effect [1]. First of all we shall consider mathematical models of a semiconductor device, after which we shall deal with the finite-difference method of computation used for the numerical solution of the partial differential equations with their initial and boundary conditions commonly encountered in these models. Finally we give the model of the Gunn effect and some results of numerical calculations that we have carried out on this model.

Mathematical models

The mathematical model of a semiconductor device consists of a set of equations that describe the behaviour of a more or less idealized situation. Like most problems in mathematical physics, these are partial differential equations, i.e. relations between functions and their derivatives with respect to the time and space coordinates. In the simplest case there is only one space coordinate in addition to the time coordinate. This means that the calculations remain limited to those situations in which it may be assumed that a quantity, such as the electric field or the space-charge density, varies in only one direction. A simple model of this type is often sufficient to describe the most important aspects of a physical effect. We assume that this will also be true of the Gunn effect, an assumption that is justified by experimental results.

A set of partial differential equations generally has an infinitely large number of solutions. In order to make the solution of a particular set unique, we must add a number of conditions that relate specifically to the situation to be described, and which usually follow from physical considerations. These are on the one hand the boundary conditions, which are specified requirements for the solution of the equations for given values of the space coordinates, and on the other the initial conditions, which prescribe the form of the solution at the beginning of the effect to be studied.

The differential equations for the Gunn effect are non-linear, i.e. the relations between the functions and their derivatives are not linear. This makes the model rather complicated, because equations of this type cannot usually be solved analytically, and it is necessary to resort to numerical methods. The results are therefore unwieldy, for assuming that a numerical approximation to the solution can be obtained, this approximation is only valid for one set of values of the parameters in the problem. Analytic solutions on the other hand may be so complicated and inconvenient in arrangement that only numerical evaluation gives any insight in the situation.

But the way in which changes in the parameters affect the solution of a problem is the very thing that physicists and engineers often want to know. In an analytical solution it is possible to ascertain the influence of the parameters, but this is certainly not so in the case of a single numerical solution. Then the influence of the parameters can only be found by calculating numerical solutions for various values of the parameters and comparing the results.

The development of digital computers that combine

Ir. J. de Groot is with Philips Research Laboratories, Eindhoven; Dr. A. Mircea is with Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes, Val-de-Marne, France.

[1] This effect, the occurrence under certain conditions of instabilities in the space-charge distribution in some semiconductors, is the subject of an article by G. A. Acket, R. Tijburg and P. J. de Waard in this issue, page 370.

large storage capacity with high computing speed has made it possible, however, to tackle partial differential equations that were previously virtually insoluble. Since the introduction of the first computers an extensive literature has grown up [2], describing the numerical methods required and dealing with their mathematical background. One of the first fundamental publications[3] in this field is much older, however, going back to 1928, when these problems were of purely theoretical interest. In spite of the great interest in the numerical solution of partial differential equations, the complexity of the subject has prevented the theory from being developed to anything like the same extent as that in other fields of numerical analysis.

The finite-difference method

The finite-difference method is one of the most useful and most commonly used methods of numerically solving partial differential equations in a large class of problems which are known beforehand to have a unique solution.

In order to find, using the difference method, a numerical approximation for a function $u(x,t)$ which is the solution of a particular set of partial differential equations in positional coordinate x and time t with the given boundary and initial conditions, we proceed as follows. We select a set of points in the (x,t)-plane such that a lattice is obtained. The lattice points are made equidistant with distances Δx and Δt (*fig. 1*), resulting in a straightforward set of equations. This choice of equidistant points is not necessary, however, and sometimes not even advisable. The points of the regular lattice are written as $(m\Delta x,\ n\Delta t)$, where m and n are integers. At these points we define a discrete function $v_m^{(n)}$, which is an approximation of the values $u(m\Delta x, n\Delta t)$ of the function $u(x,t)$. The function $v_m^{(n)}$ must be chosen such that it tends to the solution $u(x,t)$ of the set of differential equations as Δx and Δt tend to zero.

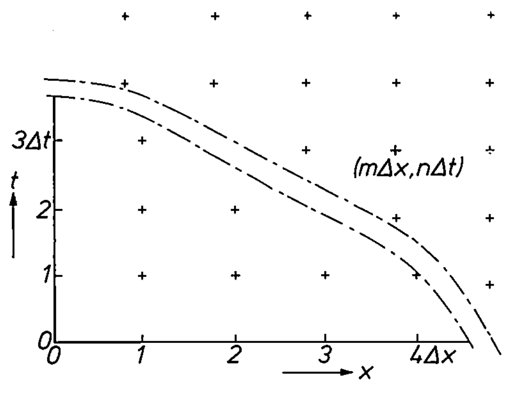


Fig. 1. Lattice of points in the (x,t)-plane, for numerically solving a partial differential equation with two independent variables, x and t.

We can replace the derivatives in the partial differential equations by difference quotients to obtain equations for the function $v_m^{(n)}$. These are referred to as finite difference equations. For example, $\partial u/\partial x$ can be approximated at a lattice point by $(v_{m+1}^{(n)} - v_m^{(n)})/\Delta x$ whereas $\partial u/\partial t$ is approximated by $(v_m^{(n+1)} - v_m^{(n)})/\Delta t$ and $\partial^2 u/\partial x^2$ by $(v_{m-1}^{(n)} - 2\,v_m^{(n)} + v_{m+1}^{(n)})/(\Delta x)^2$. Thus, when Δx and Δt tend to zero, the difference quotients will change to differential quotients and the difference equations will change to the differential equations, and the initial and boundary conditions at the lattice points to the given conditions.

Convergence

The fact that the finite-difference equations tend to the corresponding differential equations as Δx and Δt tend to zero does not necessarily mean that $v_m^{(n)}$, the solution of the finite-difference equations, converges to the solution $u(x,t)$ of the partial differential equations. In particular, if we want to allow all sufficiently smooth functions to be permitted as initial conditions, convergence sometimes will not exist. Problems arise from the fact that the initial conditions may include spatial harmonics of an arbitrarily high frequency. In many such cases there is only convergence if Δx and Δt tend to zero in such a manner that they depend on each other in an exactly defined way.

The approximate function $v_m^{(n)}$ must be the solution of a difference equation that tends to the differential equation as Δx and Δt tend to zero. For a given differential equation a number of different difference equations can be found; some of these have solutions that converge to the exact solution of the differential equation, others under restricting conditions only. We shall illustrate the latter with a heat-conduction problem.

In a homogeneous plate of infinite extent and uniform thickness d we choose the x-coordinate perpendicular to the surface such that $x = 0$ and $x = d$ coincide with the surfaces. The temperature distribution $T(x,t)$ within the plate now satisfies the differential equation:

$$\frac{\partial T}{\partial t} - a\,\frac{\partial^2 T}{\partial x^2} = 0 \quad (0 < x < d,\ t > 0), \qquad (1)$$

assuming that heat transport takes place only in the direction perpendicular to the plate. The thermal diffusivity a is a positive constant. If we now suppose that the two surfaces of the plate remain at the constant temperature 0 and that at the time $t = 0$ the temperature distribution is $T = T_{\max}\,\sin\,(k\pi x/d)$, where k is an arbitrary integer, then the boundary and initial conditions which the solution of equation (1) must satisfy are:

$$\left.\begin{array}{ll} T(0,t) = T(d,t) = 0 & (t \geqslant 0), \\ T(x,0) = T_{\max}\,\sin(k\pi x/d). & (0 \leqslant x \leqslant d). \end{array}\right\} \qquad (2)$$

[2] R. D. Richtmeier and K. W. Morton, Difference methods for initial-value problems, Interscience, New York, 2nd edition, 1967.

[3] R. Courant, K. O. Friedrichs and H. Lewy, Über die partiellen Differenzengleichungen der mathematischen Physik, Math. Ann. 100, 32-74, 1928.

This particular initial-value function allows us to obtain a solution for a very wide range of initial-value functions through Fourier expansion of these functions. The exact solution of equation (1) is uniquely determined by these conditions, and is

$$T(x,t) = T_{\max} \exp\left(-a \frac{k^2 \pi^2}{d^2} t\right) \sin (k\pi x/d),$$

as can be verified by substitution.

In order to solve this same problem numerically we define $m\Delta x = x$; $n\Delta t = t$; $\Delta x = d/M$. Here m and n are integers, and M is the total number of meshes of the lattice which we construct perpendicular to the surfaces of the plate. Instead of the differential equation (1) for the function $T(x,t)$ we now have a finite-difference equation for a function $v_m^{(n)}$, which is defined only at the points of the lattice. This finite-difference equation may be chosen as:

$$\frac{v_m^{(n+1)} - v_m^{(n)}}{\Delta t} - a \frac{v_{m-1}^{(n)} - 2\, v_m^{(n)} + v_{m+1}^{(n)}}{(\Delta x)^2} = 0.$$

With $\lambda = a\Delta t/(\Delta x)^2$ this equation together with the subsidiary conditions mentioned above yields the following three relations for the function $v_m^{(n)}$:

$$\left.\begin{aligned} v_m^{(n+1)} &= \lambda v_{m-1}^{(n)} + (1-2\lambda) v_m^{(n)} + \lambda v_{m+1}^{(n)}, \\ &\qquad\qquad\qquad\qquad (0<m<M), \\ v_0^{(n)} &= v_M^{(n)} = 0, \\ v_m^{(0)} &= T_{\max} \sin (k\pi m\Delta x/d), \quad (0 \leqslant m \leqslant M). \end{aligned}\right\} \qquad (3)$$

If we took a specified value of λ the equations (3) would yield a numerical approximation for $u(x,t)$. We shall try, however, in this fairly simple problem to get an analytical solution to the equations.

On the basis of the exact solution that is available in this case, we can try the expression:

$$v_m^{(n)} = T_{\max} \exp(-\alpha n\Delta t) \sin (k\pi m\Delta x/d), \qquad (4)$$

to see whether it satisfies the equations (3). Here α is a real but unknown number. Substitution of (4) in the first of the difference equations (3) yields an expression for α:

$$\exp(-\alpha\Delta t) = 1 - 2\lambda + 2\lambda \cos(k\pi\Delta x/d) = 1 - 4\lambda \sin^2(k\pi\Delta x/2d).$$

With this expression we can write the solution of the finite difference equation as:

$$v_m^{(n)} = T_{\max} \{1 - 4\lambda \sin^2 (k\pi\Delta x/2d)\}^n \sin (k\pi m\Delta x/d).$$

The question now is whether the solution converges to the exact solution of the differential equation. To verify this, we compare the two solutions at an arbitrary instant of time $t_0 = n_0\Delta t$ and at an arbitrary point $x_0 = m_0\Delta x$ and see what happens when Δx and Δt tend to zero. If x_0 and t_0 are to remain constant, then m_0 and n_0 must tend to infinity. This means that $| 1 - 4\lambda \sin^2(k\pi\Delta x/2d) |^{n_0}$ will also tend to infinity, unless $| 1 - 4\lambda \sin^2(k\pi\Delta x/2d) | \leqslant 1$. If the latter is to be the case for all values (including arbitrarily large values) of k then the condition $\lambda \leqslant \frac{1}{2}$ must hold. This means that in order for the solution of the finite difference equation to converge to the exact solution of the partial differential equation, Δt and Δx must tend to zero in such a way that we always have $\lambda = a\Delta t/(\Delta x)^2 \leqslant \frac{1}{2}$. It turns out that this condition is both sufficient and necessary. If the condition is not satisfied, the solution of the finite difference equation has nothing to do with the solution of the partial differential equation.

For other finite-difference equations that approximate to the same partial differential equations the condition for convergence of the numerical approximation will be different, and it may turn out that it is not necessary to impose any condition on Δx and Δt in order to guarantee convergence of the numerical solution.

Stability

Each step of a calculation carried out in a digital computer consists of a number of elementary arithmetical operations. The result of each operation has to be rounded off, because there are only a limited number of figures available for the storage and read-out of these results. Each operation thus introduces a small error. As the number of steps is usually very large, the final result may therefore contain an unacceptable accumulation of errors. This is referred to as computational instability. To avoid such instability the finite-difference equations must be chosen in such a way as to make them sufficiently insensitive to rounding errors in the individual arithmetical operations. The rounding errors, transferred back to the initial situation, are equivalent in a linear problem to "noise" on the initial-value function. As this noise is random, it may contain fairly high spatial frequencies and consequently a Fourier expansion of this "noisy" function contains spatial harmonics of an arbitrarily high frequency. The requirements for stability of a linear difference equation thus are the same as the conditions for convergence in the case of a sufficiently smooth initial-value function treated above [2].

Computing time

The heat-conduction problem can also provide some idea of the time involved in a computer solution of such a problem.

In our heat-conduction problem Δx was equal to d/M. Suppose that we are interested in the solution at a time $t = \tau$; this requires a number of $\tau/\Delta t$ time steps. To satisfy the condition for convergence, Δt must be proportional to $(\Delta x)^2$. This means that Δt must be proportional to $1/M^2$, yielding a number of lattice points in the (x,t)-plane proportional to M^3. If the behaviour of the solution in the direction of the x-axis is such that M must be large in order to obtain a reasonable accuracy, the number of lattice points soon becomes fairly large. The difference equation has to be solved for each point of the lattice, in each case requiring at least two multiplications and two additions.

Using the finite-difference equation dealt with above, we have performed the calculation for a plate 10 cm thick divided into 100 intervals Δx. The temperature distribution at the beginning of the calculation was given, and we wanted to know the change of this distribution as a function of time. For this purpose we chose $\lambda = 0.45$, thus guaranteeing the convergence and stability of the calculation scheme adopted. If $a = 1$ cm²/s, which is approximately the value for metals like copper and aluminium, the length of a time step is 0.45×10^{-2} s. To follow the variation of the temperature distribution during 500 time steps, i.e. 2.25 s, our difference scheme required 36 s on a medium-size computer. The calculation is thus about 15 times slower than the actual physical effect, as a consequence of the stability and convergence requirements of the computing scheme adopted. If we choose $\lambda = 0.55$ a time step will be 0.55×10^{-2} s, but the computational scheme will then no longer be stable and after only five steps the results will show small errors (*fig. 2*). For such a simple example as we have taken here, however, computational schemes can be found that are less stringent, and indeed may impose no conditions at all for convergence. In this case the magnitude of a time step depends only on the accuracy required.

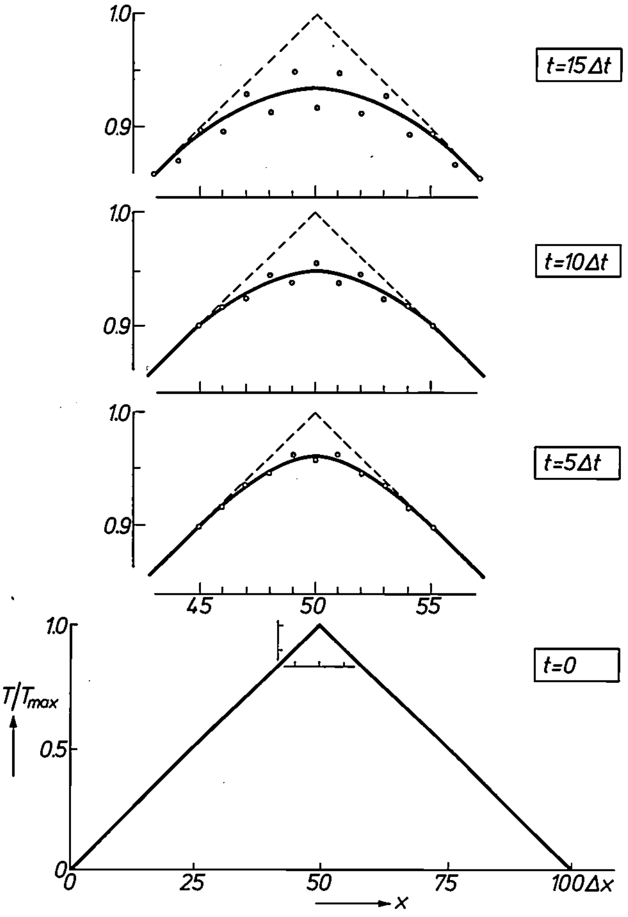


Fig. 2. Temperature distribution in a homogeneous plate of infinite extent and uniform thickness, whose surfaces are kept at a temperature $T = 0$. For a number of values of the time t the temperature is plotted as a fraction of $T_{\max}$, the maximum value of the initial distribution. The thickness of the plate was divided into 100 intervals Δx. The curves give the exact distributions resulting from analytic calculations. The points are calculated numerically by means of the difference scheme given in the text. A time step Δt was chosen to be 0.55×10^{-2} s. In the situation shown this gives $\lambda = 0.55$ and consequently an unstable difference scheme. At $t = 5\,\Delta t$ deviations from the exact values of T can clearly be distinguished. An ever increasing modulation with a period $2\Delta t$ develops on the temperature distribution. At $40\Delta t$ the calculated values of T can even be negative.

A mathematical model of the Gunn effect

We shall now discuss some features of a mathematical model of the Gunn effect as encountered in a Gunn device, consisting of a strip of a suitable semiconducting material provided at both ends with electrical contacts. The simplest model contains, in addition to time, only a single spatial coordinate as independent variable. The model is satisfactory as long as the current density in the device is homogeneous and surface effects on the side face of the device are negligible. An external voltage applied between the contacts produces in the device an electric field in which conduction electrons acquire an average drift velocity v. The relation between v and the electric field E is given by the v-E characteristic in *fig. 3*. This characteristic shows a range of values above a threshold field E_{th} in which the velocity of the electrons decreases with increasing field-strength. As we shall see, the occurrence of this negative slope in the v-E characteristic can give rise to instabilities in the space-charge distribution in the device. The experimental methods of observing the characteristic, and the physical explanation of these observations, are given in another article in this issue [1].

If the shape of the Gunn device is such that the effects may be described as one-dimensional, the relation between the electrical quantities is given by the following three differential equations:

$$\frac{\partial E(x,t)}{\partial x} = \frac{e}{\varepsilon} \left\{ n(x,t) - n_{\text{d}}(x) \right\}, \tag{5}$$

$$J(x,t) = en(x,t)\ v(E(x,t)) - eD\frac{\partial n(x,t)}{\partial x}, \tag{6}$$

$$e\frac{\partial n(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0. \tag{7}$$

Here $E(x,t)$ and $J(x,t)$ are the electric field and convection-current density as functions of the spatial coordinate x, and the time coordinate t. The total current density J_{tot} is found by adding a quantity $\varepsilon\ \partial E/\partial t$, the displacement-current density, to the convection-current density $J(x,t)$. Further, $n(x,t)$ is the free-electron concentration, and $n_{\text{d}}(x)$ is a function representing the doping concentration of the n-type material used. D is the diffusion constant, e is the electronic charge ($e > 0$) and ε is the dielectric constant.

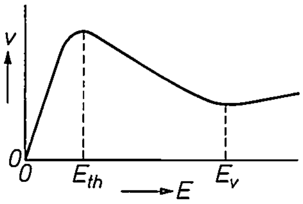


Fig. 3. The average drift velocity v of conduction electrons in gallium arsenide as a function of the electric field E in the material. The range with negative slope between E_{th} and E_{v} can give rise to instability effects in the charge distribution in a Gunn device.

Equation (5) is Poisson's equation, and (7) is the charge-continuity equation; both are linear and are valid as long as the generation and recombination of charge carriers are negligible. Equation (6) states that the convection current consists of a part due to conduction and a part due to diffusion. The conduction current is non-linear due to the shape of the v-E characteristic and to the product $n(x,t)\ v(E(x,t))$ in equation (6). In many cases the contribution of diffusion to the total current can be neglected without seriously affecting the results of the calculations.

The set of coupled differential equations (5)-(7) could give rise to the same problems in a numerical solution as found in the example mentioned above of the heat-conduction equation. There would then have to be a particular ratio between the steps in the coordinates x and t in order to arrive at a convergent and stable numerical solution. However, finite-difference equations can be found, corresponding to the set (5)-(7) that do not impose restrictive conditions on Δx and Δt for convergence.

Instabilities in the space-charge density

As we shall see, a characteristic of the type in fig. 3 can give rise to instabilities. These instabilities have no relation whatever to the instabilities of numerical calculations. We shall now show that equations (5)-(7) admit solutions that have the character of an electrical instability. We assume to begin with that a constant voltage V_0 is present across the device and that the doping concentration n_d is constant throughout the material. The voltage V_0 gives rise to a uniform electric field E_0 in the device, given by $E_0 = V_0/l$, where l is the length of the device. In view of (5) we know that $n(x,t)$ then has a constant value n_0 which must be equal to n_d. From (6) it follows that $J(x,t) = en_0v(E_0) = J_0$ is likewise constant. Is this situation stable? In order to answer this question we substitute in equations (5)-(7) the constant values E_0, n_0 and J_0, augmented by small perturbations, designated by the subscript 1:

$$\begin{aligned} E(x,t) &= E_0 + E_1\,(x,t), \\ n(x,t) &= n_0 + n_1\,(x,t), \\ J(x,t) &= J_0 + J_1\,(x,t). \end{aligned}$$

If we neglect products of perturbations and also assume that the contribution of diffusion to the current is negligible, we can readily derive a differential equation, e.g. for the perturbation n_1 in the space-charge density. This equation has the solution

$$n_1(x,t) = \exp\,(-\sigma_{\rm diff}t/\varepsilon)\,f(x - v(E_0)t), \qquad (8)$$

where $\sigma_{\rm diff}$ is the differential conductivity, defined as $\sigma_{\rm diff} = e\,n_0\,{\rm d}v(E_0)/{\rm d}E_0$, and $f(x - v(E_0)t)$ is an arbitrary but continuously differentiable function of $x - v(E_0)t$. The physical significance of this solution is as follows. If at some instant $t = 0$ there is a small perturbation $f(x)$ of the stable state, this will propagate at the velocity $v(E_0)$ in the positive x-direction and will grow or decay with time depending on whether $\sigma_{\rm diff}$ is negative or positive. In the first case, which is found when $E_{\rm th} < E < E_{\rm v}$ (see fig. 3), the situation is unstable: the perturbation propagates and tends to infinity. In the calculations that led to equation (8), however, it was assumed that the perturbations were small. But in the unstable situation this will very soon no longer be the case, and therefore this approximation will no longer be correct. The non-linearity of the system will then start to play a part and will oppose further growth of the perturbation. The influence of this non-linearity can only be studied by solving the original equations (5)-(7) numerically. This can be done by using the finite-difference method dealt with at the beginning of the article, and we shall presently discuss some results of calculations of this type.

In equation (8), which describes the initial phase of all possible instabilities, $f(x - v(E_0)t)$ may have the form $\sin\,\{2\pi n(x - v(E_0)t)/l\}$. This means that a complete spectrum of oscillations can occur with the fundamental frequency $v(E_0)/l$. These oscillations may be of practical importance if the associated space-charge perturbations are able, in moving through the device, to grow to a value of the order of magnitude of the space-charge density n_0. The time available for such growth is the transit time $l/v(E_0)$. Therefore, keeping to the linear approximation we have already used, we can say that a perturbation is only of practical importance if the argument of the exponential in equation (8) is greater than 1, i.e. if

$$-\frac{\sigma_{\rm diff}}{\varepsilon}\,t = -\frac{e}{\varepsilon}\,n_0\,\frac{{\rm d}v(E_0)}{{\rm d}E_0}\,\frac{l}{v(E_0)} > 1\,.$$

Evidently, this will only apply when ${\rm d}v(E_0)/{\rm d}E_0$ is negative, i.e. only when $E_{\rm th} < E_0 < E_v$.

For a given semiconductor device the maximum value of $|{\rm d}v(E_0)/{\rm d}E_0|$ and the associated values of E_0 and $v(E_0)$ are fixed, and only the values of l and n_0, which is equal to n_d, can be chosen. If a Gunn device is to be used as an oscillator the following requirement [4] must therefore be met:

$$n_{\rm d}l > -\,\varepsilon\,v(E_0) \Bigg/ e\,\frac{{\rm d}v(E_0)}{{\rm d}E_0}\,.$$

The maximum negative value of ${\rm d}v(E_0)/{\rm d}E_0$ of N-type gallium arsenide (GaAs), the usual material for Gunn devices, is approximately 2×10^3 cm^2/Vs, the value of $v(E_0)$ then being about 10^7 cm/s. Since this material has an $\varepsilon_{\rm r}$ of about 10, we must have $n_{\rm d}l > 10^{11}$ cm^{-2} for an oscillator diode.

We have assumed in the foregoing that the contribution of diffusion was negligible. Since diffusion reduces the concentration gradients, it will also change the growth rate, the magnitude and shape of a local perturbation of the space-charge density [5]. This causes the charge accumulation to spread, which is a dissipat-

[4] D. E. McCumber and A. G. Chynoweth, IEEE Trans. **ED-13**, 4, 1966.

[5] B. W. Knight and G. A. Peterson, Phys. Rev. **155**, 393, 1967.

ive process and leads to a higher minimum value of n_dl than the estimated 10^{11} cm^{-2} just given.

We could also have derived a differential equation for the perturbation of the electric field or the current. In that case we would have found perturbations of these quantities corresponding to the space-charge perturbation discussed above.

Accumulation layers and high-field domains

We arrive at a more realistic model if we assume that the doping profile in a device is of the form shown in *fig. 4.* By means of special alloying techniques [6] highly doped, and hence low-ohmic, contact regions have been applied to the ends of a bar of semiconductor material. These regions, called anode or cathode according to their polarity, serve for making non-rectifying contacts with the current leads. In the ideal case the bulk doping concentration is virtually constant, and if the alloying techniques are sufficiently known, the form of the doping-concentration function $n_d(x)$ is known.

We now apply to this device a constant external voltage V. If the resultant electric field somewhere in the device is in the critical region between E_{th} and E_v, space-charge instabilities will build up. This gives rise to a current that reflects the instabilities in the device. Instabilities arise at local space-charge surpluses or deficits in the device, e.g. at the cathode contact in the case of the doping profile given in fig. 4, or they arise from space-charge fluctuations due to noise or to fluctuations in n_d. We have seen that instabilities move from cathode to anode, and therefore no instabilities can arise at the anode contact. An instability consisting of a temporary and local increase of the electron concentration is called an accumulation layer. The space-charge density and electric field in such a layer are shown in *fig. 5.* The maximum in the space-charge density increases with time until non-linearities in the process prevent further growth. The field distribution is found from the charge distribution by means of Poisson's equation, and is such that the integral of E over the length of the device is equal to the applied voltage V.

Fig. 6a shows the formation and propagation of an accumulation layer as obtained by numerical integration of equations (5)-(7) [7]. The accumulation layer forms at the cathode contact. During the movement of the layer through the device the field at the cathode is small, but as soon as the layer disappears in the anode contacı the electric field at the cathode increases again and allows a new accumulation layer to form, which in its turn is amplified and the cycle is repeated. The minimum field-strength at which an accumulation layer can form in N-type gallium arsenide is approximately 3.6 kV/cm. This cyclical process causes a periodic modulation of the current through the device. The current modulation for the propagation of an accumulation layer as in fig. 6*a* has been calculated, and the results are presented in fig. 6*b*.

If the deviation from uniformity in the charge-density distribution is negative, a depletion layer is formed. During the movement of a depletion layer through the device the electron density in the layer decreases. However, since the electron density cannot become negative, this constitutes another limiting mechanism in addition to the non-linearity of the equations. We did not however include this limiting mechanism in our mathematical model.

In practice the doping is never completely uniform; there will always be both positive and negative devia-

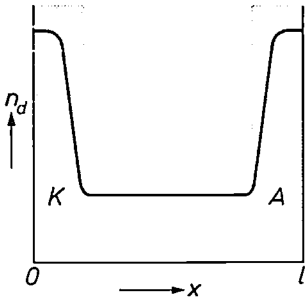


Fig. 4. Variation of the doping concentration n_d as a function of distance x in a Gunn device of length l. K and A regions of relatively low resistivity at the ends of the device, referred to as cathode and anode.

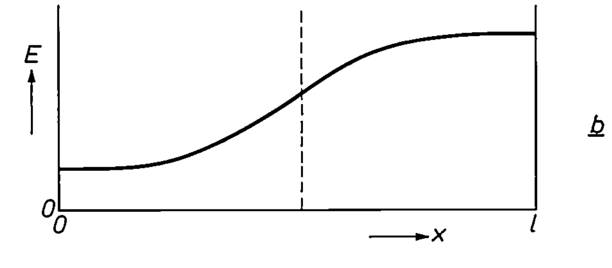


Fig. 5. *a*) Space-charge density n in a Gunn device as a function of the spatial coordinate x, for the case of an accumulation layer, one of the simplest instabilities in the charge distribution. n_d doping concentration in the device. n_{max} maximum charge density in the accumulation layer. l total length of the device. *b*) Corresponding plot of the electric field E in the device.

[6] See the article by Acket, Tijburg and De Waard in this issue, page 370.

[7] H. J. Knoop of Philips Research Laboratories gave valuable assistance in the development of the computer programs for these calculations.

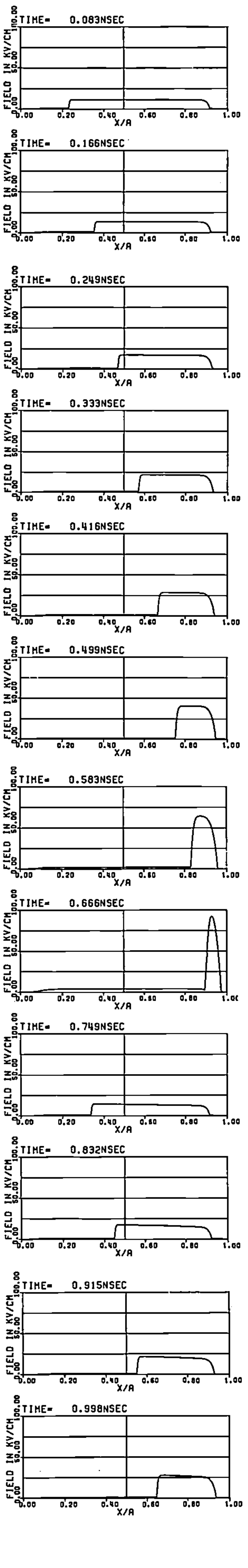
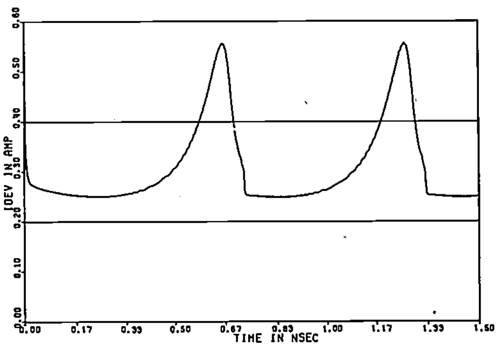


Fig. 6. *a* Electric-field distribution in a Gunn device, numerically calculated by means of a computer, for the case of an accumulation layer moving through the device. The results of the calculation are given after equal time intervals of just over 0.08 ns. X/A normalized distance to the cathode end of the device. *b*) Variation in the current I_{DEV} through the device corresponding to the propagation of an accumulation layer for a constant voltage across the device.

←*a* *b*→



tions from the average concentration. If the local electric field in the device is in the negative dv/dE range, several accumulation and depletion layers can form at the location of these deviations, including the nucleation of an accumulation layer at the cathode contact. An accumulation layer and a depletion layer near to one another will combine to form a moving dipole layer, called a high-field domain. If the device is sufficiently long and the applied voltage sufficiently high, a high-field domain has the chance of growing to its final form on its way towards the anode. *Fig. 7* shows the electron density and the electric field prevailing in a domain at a given moment. When a domain reaches the anode contact and vanishes there, this is manifested in a temporary change of the current $I(t)$ in the external circuit. The external current is not affected, however, by the movement of a completely developed domain through the device. The vanishing of a domain increases the electric field in the device, causing the immediate nucleation of a new domain at local deviations in the doping concentration. Here again, the result is a periodic modulation of the current in the external circuit. *Fig. 8* shows the results of numerical calculations of the nucleation and movement of a high-field domain. This calculation was based on a doping distribution $n_{\mathrm{d}}(x)$, assumed to have a local decrease in n_{d} just behind the cathode contact in order to nucleate the depletion layer associated with the domain.

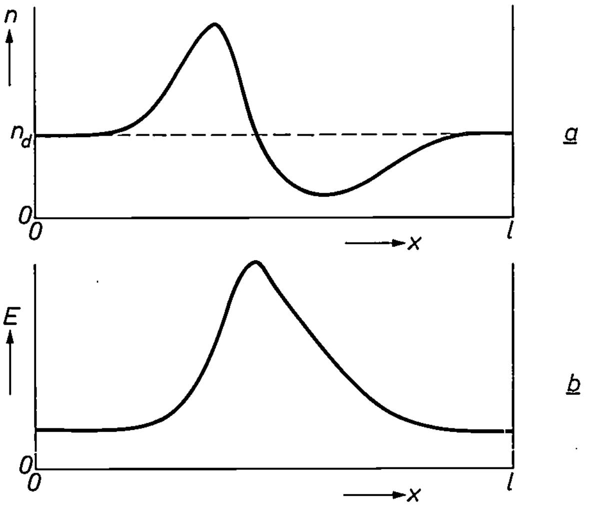
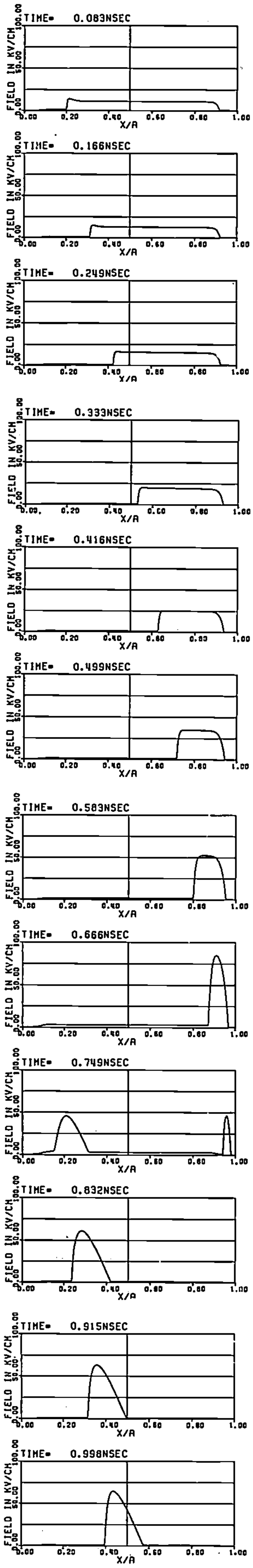


Fig. 7. *a*) Space-charge density n as function of spatial coordinate x in a Gunn device for the propagation of an electric-dipole layer (high-field domain). *b*) The corresponding field distribution in the device.

The mechanism of the periodic formation and disappearance of high-field



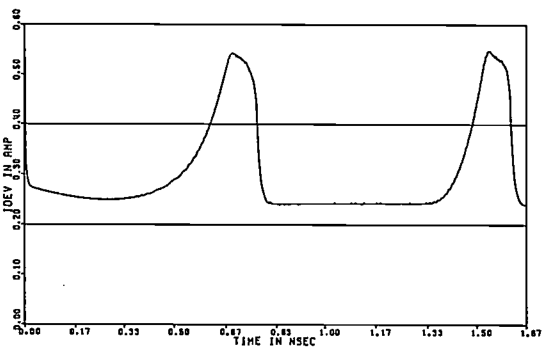


Fig. 8. *Left:* electric-field distribution in a Gunn device, numerically calculated by means of a computer, for the propagation of a high-field domain through the device. The successive plots are given in the same way as those of fig. 6. *Right:* modulation of the current I_{DEV} during the propagation of high-field domains.

domains underlies one of the possible modes of operation of the Gunn oscillator [6]. When the electric field anywhere in the device is in the range between E_{th} and E_{V}, oscillations of this type will only be generated, as we have seen, when the product of n_{d} and l is greater than 10^{11} cm^{-2}. For values of n_{d} in the region of 10^{15} cm^{-3} a device with a length of 10 μm is therefore rather short. If such a device is given a small diffusion constant D, it is nevertheless possible to find periodic solutions of the basic equations (5)-(7). If, on the other hand, the diffusion constant is large, one can only find a solution that tends to a time-independent situation. The example given in *fig. 9* shows what happens in such a case. First an accumulation layer or high-field domain is formed. Once it has arrived at the anode contact, however, diffusion increases and the movement of the instability stops. Behind the space-charge perturbation the electric field is decreased, consequently the situation there becomes stable. Clearly, a time-independent situation of this kind can only exist if the diffusion current can be made so large as to compensate the difference between the external current and the drift current, so that the dielectric displacement current vanishes. This means not only that D/l must be sufficiently large, but also that the effect is partly dependent on the doping concentration [8].

The possible existence of a stable situation, even when the electric field is such that the differential resistance is negative, has so far only been found in computer simulations [9]; it has not yet been demonstrated by experiments. It is thought, however, that stable domains of this kind may provide an explanation for the wide-band amplification observed in short gallium-arsenide devices [8].

All the numerical results described so far were obtained by means of the finite-difference method applied to the basic equations (5)-(7) together with the boundary and initial conditions needed to make the solution unique. To obtain a sufficiently accurate approximation we had to divide the length l of the device into 400 intervals of equal width. In all cases l was 100 μm, except in the case shown in fig. 9, where l was only 12 μm. With a view to the required accuracy it was necessary to make 500 time steps for computing three periods of the oscillations. To give an idea of the computing time needed, each of the results presented in figs. 8 and 9 took about 10 minutes, using a large computer system.

Butcher's equal-areas rule

The basic equations (5)-(7) that give a description of the Gunn effect allow us to derive equations for the onset of instabilities in the space charge in a Gunn diode. Owing to non-linearity of the basic equations, they cannot yield a description of the subsequent growth of the instabilities and one then has to rely on numerical calculations.

The final situation, however, for the fully grown domain is again determined by an analytic expression. This expression, known as Butcher's equal-areas

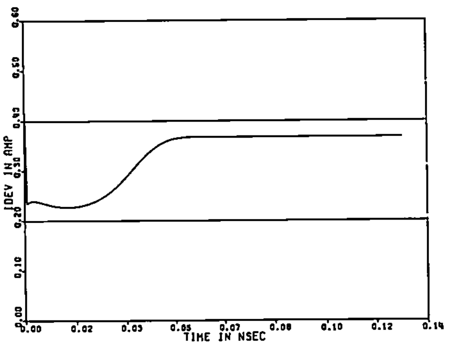
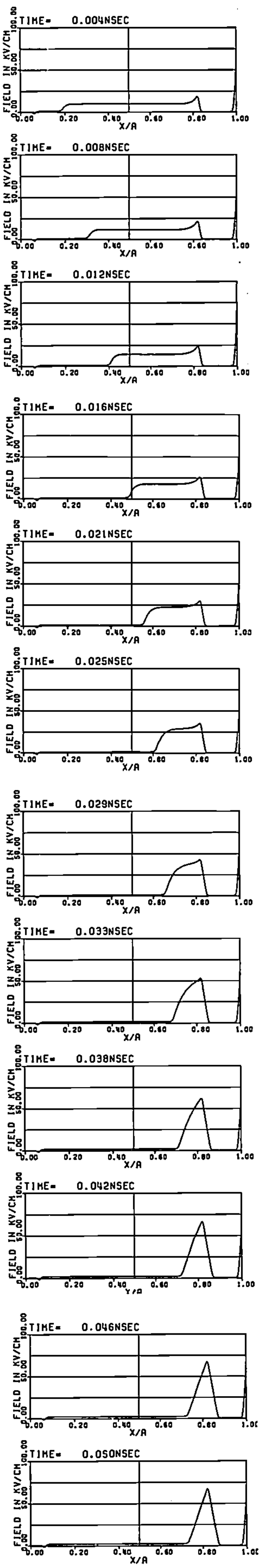


Fig. 9. *Right:* electric field distribution in a Gunn device, numerically calculated by means of a computer, for the case of a high-field domain propagating through the device when the diffusion coefficient is high. The domain, which normally grows during its propagation through the device, now becomes stable because diffusion compensates the growth. The individual plots give the calculated situation after equal time intervals of 0.004 ns. *Left:* plot of current as the domain becomes stable.



rule [10], gives the relation between the maximum field-strength in the domain and the field-strength outside the domain. The importance of the equal-areas rule is in the fact that it allows computation of domain form and domain speed, which are fundamental in a Gunn oscillator [6], without the numerical calculations given above.

Interaction of a Gunn device with a circuit

As we have seen, the propagation of a high-field domain in a Gunn device results in periodic current variations in the external circuit. This means that the device can be used as an oscillator. For power to be extracted from such an oscillator, the device must be incorporated in an electrical circuit. As the device influences the currents and hence the voltages in the circuit, there cannot in practice be a constant voltage across the device. The nucleation and propagation of high-field domains will not therefore take place in the same way as in the foregoing idealized case of a constant voltage V across the device. The practical question is now how to choose the parameters of the system so as to obtain maximum efficiency from the oscillator.

The interaction of the device with the circuit, which is in fact a complicated microwave circuit, can be analysed by means of the simplified circuit given in *fig. 10.* Using the same data as in the example given in fig. 8 we again solve the equations (5)-(7) numerically with boundary conditions determined by the parameters of the circuit. The results are presented in *fig. 11.* Comparison with fig. 8 shows that a domain becomes smaller the nearer it approaches the anode, owing to a decrease in the voltage across the device.

By applying a Fourier analysis the amplitude of the first-harmonic components of the currents and voltages can be found and the efficiency of the device calculated from the result. In investigations of this kind the computer is obviously very useful for studying the effect of a variety of parameters on the efficiency, such as the doping profile and changes of the characteristic with temperature. It is difficult and sometimes impossible to do this by means of experiments.

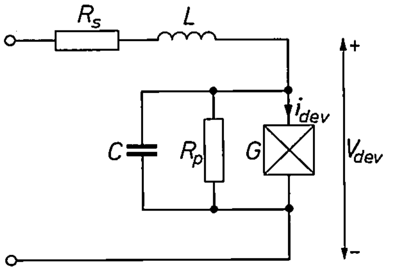
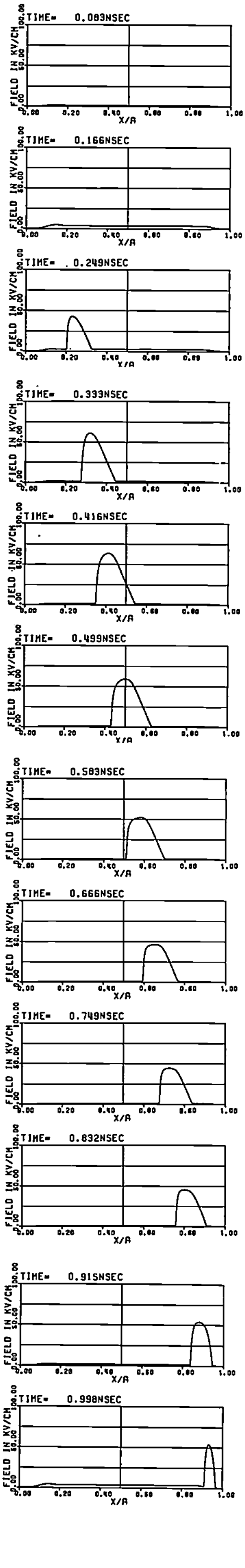


Fig. 10. Circuit in which a Gunn device can generate microwave oscillations. G Gunn device. R_s series resistance. L inductance. R_p parallel resistance. C parallel capacitance. V_{dev} voltage across the Gunn device.

[8] J. Magarshack and A. Mircea, Proc. MOGA Conf., Amsterdam 1970, page 16.19.

[9] A. Mircea, Device Research Conference Rochester, N.Y., U.S.A., 1969.

[10] P. N. Butcher, Physics Letters 19, 546, 1965. This rule is also derived in the article by Acket, Tijburg and De Waard in this issue, page 376.



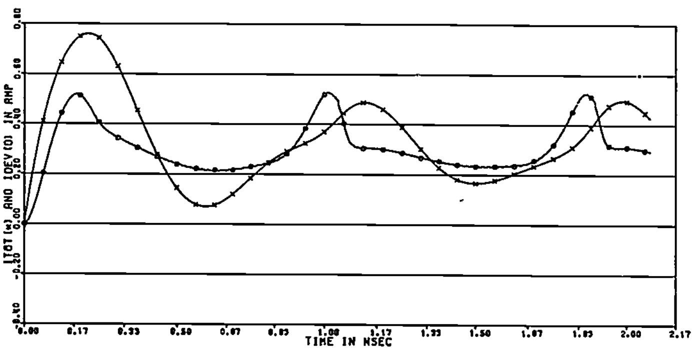


Fig. 11. *Left:* field distribution in a Gunn device, plotted for the circuit in fig. 10, during propagation of high-field domains through the device. The individual plots give the calculated field distribution after equal intervals of about 0.08 ns. *Right:* current I_{TOT} (x) through the circuit and the associated current I_{DEV} (o) through the device.

Limited space-charge accumulation

The modes of oscillation discussed so far are typical transit-time effects, the frequency being almost completely determined by the transit time of a space-charge perturbation in the device. When such a device is coupled to an electrical circuit, the resonant frequency of the circuit must be near the transit-time frequency of the device for optimum operation. To conclude this article we shall discuss a mode of operation where the oscillation frequency is completely determined by the circuit; this is known as the limited space-charge accumulation mode (LSA mode) [6]. This mode was first discovered when performing computer simulations of the Gunn effect [11], but shortly afterwards it was verified experimentally [12]. It was found that the diode material had to be highly homogeneous. Unlike the domain modes discussed in the foregoing, however, the electric field in the device is nearly uniform in this case, even when the field-strength is in the range of negative differential resistance. This means that the negative resistance can now operate over the whole length of the device and not only over the width of a domain. Operation of an oscillator in the LSA mode therefore promises a higher efficiency than in the domain mode.

Stability of the uniform field in the device is assured only if the differential conductivity is positive during part of a period of the oscillation, to prevent the build-up of local charge instabilities. The situation is illustrated in *fig. 12.*

Further study of the LSA mode shows that approximately sinusoidal oscillations can occur and also relaxation oscillations [13]. Experiments have demonstrated [14] that particularly high efficiencies can be achieved with relaxation oscillations. They are obtained with the circuit given in fig. 10, in which, generally speaking, the inductance is high and the capacitance low.

Fig. 13 shows how the behaviour of the circuit depends on one of the param-

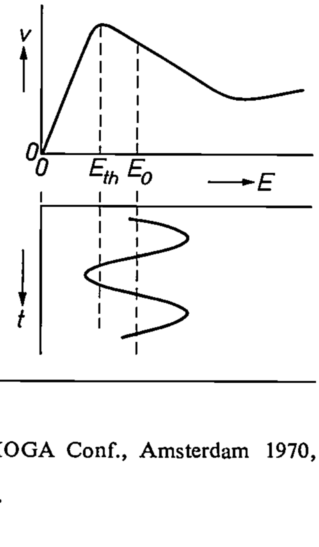


Fig. 12. Way in which the space-charge accumulation in a Gunn device is limited by the field entering the range of stability during part of the oscillations (the field E is then in each period for a short time below threshold value E_{th}).

[11] J. A. Copeland, Proc. IEEE 54, 1479, 1966.
[12] J. A. Copeland, J. appl. Phys. 38, 3096, 1967.
[13] J. de Groot and M. T. Vlaardingerbroek, Proc. MOGA Conf., Amsterdam 1970, page 20.34.
[14] P. Jeppesen and B. Jeppsson, Proc. IEEE 57, 795, 1969.

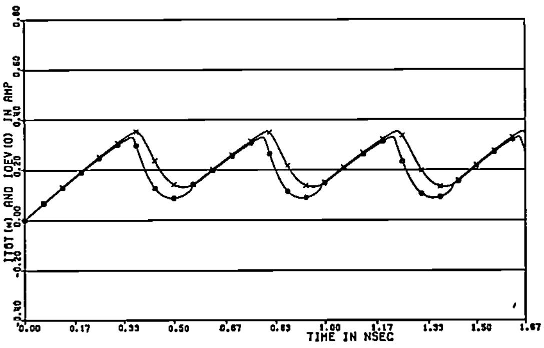
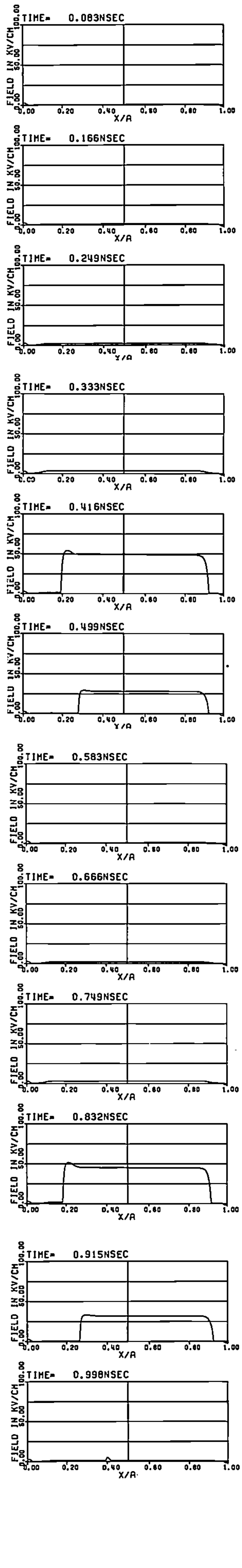


Fig. 14. *Right:* numerically calculated field distribution in a Gunn device for operation in the LSA mode. The individual plots give the calculated field distribution after equal intervals of about 0.08 ns. *Left:* the current $I_{\rm TOT}$ (x) through the circuit and the associated current $I_{\rm DEV}$ (o) through the Gunn device for operation in the LSA mode.



eters in a particular case. The curves were again obtained by numerical integration of the basic equations (5)-(7). For each value of $R_{\rm S}$ values of L and C were chosen that gave the same oscillation frequency in the LSA mode. In the high-field domain mode the frequency is not closely dependent on L and C, since it is almost completely determined by the geometry of the device. It is found that at small values of $R_{\rm S}$ the oscillation is virtually sinusoidal, but as $R_{\rm S}$ increases it changes into a relaxation oscillation. Above a certain series resistance the maximum possible voltage variation across the device is too small to maintain oscillations in the LSA mode, and a sudden transition occurs to the domain mode. If the series resistance increases still further, the voltage across the diode finally becomes too small for the nucleation of domains, and therefore frequency and efficiency drop to zero. *Fig. 14* shows the numerically calculated field distribution in a Gunn device operating in the LSA mode. In these calculations we assumed a local decrease of 2% in the doping distribution in order to see what effect this would have on the limited space-charge accumulation. Although this deviation causes a slight local increase of the electric field, it does not result in the build-up of a high-field domain.

In conclusion we can say that the agreement between the numerical calculations discussed here and the experimental results is on the whole good (*fig. 15*). This means that up to a certain extent the mathematical model used was correct. It is of course obvious that a model can never represent the reality in every detail, since its construction requires certain simplifications of the real situation, and there is always some uncertainty about the values of the various parameters. Microwave circuits, for example, are too complex to be described completely in a model, and a doping profile is never known completely accurately. Because of all this the results of computer calculations as described here can never have more than a qualitative significance, they yield no more than salient features of the effects studied. Nevertheless this information is useful in that it provides

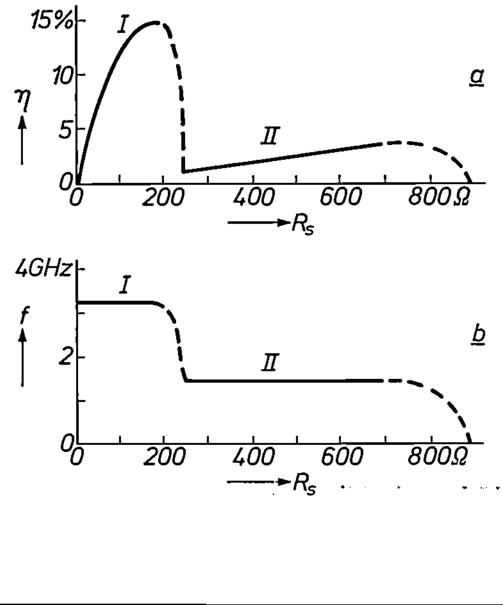


Fig. 13. *a*) Efficiency η of a Gunn oscillator, shown in fig. 11, as a function of the series resistance $R_{\rm S}$. In the region *I* oscillations occur in the LSA mode (limited space-charge accumulation), in the region *II* the oscillations are caused by the propagation of high-field domains. *b*) Oscillation frequency in the case of operation in the LSA mode. In the case of LSA oscillations (*I*) the frequency is kept constant by varying the magnitude of L and C. In the domain mode the frequency is determined by the geometry of the diode and is independent of the circuit parameters.

an understanding of the way in which changes in certain parameters affect the behaviour of the system.

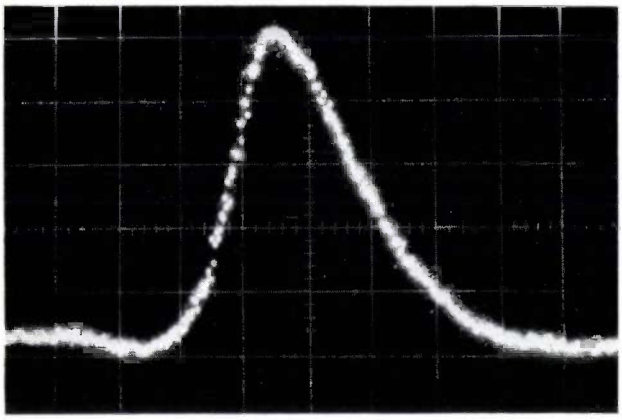


Fig. 15. Field distribution in a high-field domain in a Gunn device, plotted by the method described on page 370 of the article by G. A. Acket, R. Tijburg and P. J. de Waard [6]. The steepness of the leading edge is fundamentally limited by the measuring instruments used, and may therefore in reality have been greater. For the rest the agreement with the calulated distribution given in figs. 8 and 11 is reasonably good.

Summary. In the first part of this article the finite-difference method used for the numerical solution of partial differential equations is discussed. In this method differential quotients are replaced by difference quotients and differential equations by difference equations. When the differences of the independent variables tend to zero a difference equation tends to a differential equation. For the solution of the difference equation to tend to the solution of the differential equation, it is often necessary to impose conditions on the way in which the differences of the independent variables tend to zero. This is illustrated with the example of a simple heat-conduction problem. In the second part the Gunn effect is mathematically described with the aid of three partial differential equations together with boundary conditions. It is shown that there is a range of negative slope in the v-E characteristic of a Gunn device that causes various kinds of instability in the current through the device, such as charge-accumulation layers and high-field domains. These domains can be used for generating microwave oscillations. A time-independent state also exists in which a large part of the device is in fact in an unstable state. The differential equations for all these cases have been solved numerically by computer, and the results are presented and discussed.

Gunn-effect oscillators and amplifiers

J. Magarshack

The discovery and subsequent development of solid-state microwave generators such as Gunn and avalanche diodes has profoundly affected the conception of microwave systems. For the first time sources of microwave energy have become available which are as cheap as a transistor and provide sufficient energy for a small system from a low-voltage d.c. supply. Not only has existing equipment become simpler but a microwave solution is now a feasible alternative in many applications where the cost is important, and where miniature components and simple power requirements are necessary.

At about the same time as these sources of microwave energy became available, people were looking for new solutions at higher frequencies. This is particularly noticeable in telecommunication, where the density of information to be carried by a system is increasing very rapidly with the increase in subscribers to the system and also with the complexity of the signals (video rather than audio signals for instance). This pushes the carrier frequencies well into the microwave region.

In designing a microwave system using Gunn diodes it is necessary to know the characteristics of such diodes, how they perform in a circuit and what circuit will be necessary to satisfy a given specification. We shall therefore consider what sort of power, efficiency, frequency range and noise characteristics we can expect for a Gunn diode in a particular application. These properties, however, depend on the circuit in which the diode is placed, and such a circuit can be best defined if we know the impedance of the diode. We shall therefore begin by describing the impedance of the diode under large-signal conditions and, later on, under small-signal conditions. Large-signal impedances refer to the diode when it is oscillating, in which case the impedance is defined as the ratio of the fundamental frequency component of the voltage to that of the current. This value is a function of the r.f. voltage swing. Small-signal impedance measurements are only possible if the diode is not oscillating and for small enough measuring signals. Under these conditions the diode presents a linear impedance which can be used in a reflection-type amplifier if its resistive component is negative.

After the impedance of the Gunn diode has been determined and its performance predicted, circuits can be designed for particular applications. We shall consider three such circuits, which are often required: an oscillator for frequency-modulated systems, voltage-tuned by means of a varactor, a wide-band tunable oscillator, magnetically tuned by means of a YIG sphere, and a fixed-frequency low-noise stable oscillator.

Finally we shall describe the amplifying possibilities of the transferred-electron device with particular reference to a new mode of amplification that uses diodes that are very similar to those used in oscillators.

Oscillators

Oscillation conditions

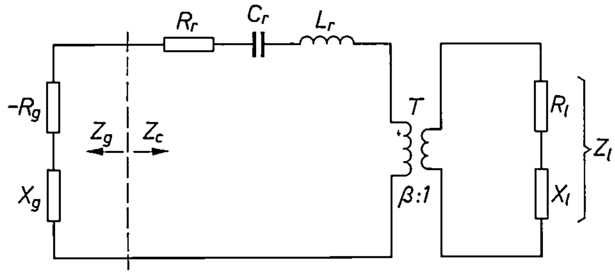


Fig. 1. Gunn oscillator circuit. Z_g Gunn-diode impedance, with $-R_g$ resistive and X_g reactive component. L_r, C_r and R_r represent the resonant cavity. T ideal transformer with impedance transformation ratio β : 1. The load impedance Z_l in the general case consists of a resistive part R_l and a reactive part X_l. The total circuit impedance seen by the Gunn diode is Z_c.

The Gunn diode can be made to oscillate in waveguide, coaxial or microstrip circuits by means of a resonator and a low-pass filter for connection to the d.c. power supply. For a simple analysis of the circuit required we can represent the Gunn diode by large-signal negative resistance $-R_g(V,\omega)$ in series with a reactance $X_g(V,\omega)$. Both resistance and reactance are functions of the voltage swing V and the frequency ω. For convenience we take a series circuit (*fig. 1*). The circuit impedance (Z_c) can be represented by a series-resonant circuit (R_r, C_r, L_r) to which the load ($Z_l = R_l + jX_l$) is coupled by an ideal transformer (impedance transformation ratio β : 1). The condition

J. Magarshack is with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes, Val-de-Marne, France.

for oscillation is that the total impedance in the circuit vanishes, so that

$$Z_g + Z_c = 0, \qquad (1)$$

where the impedances are defined at the oscillation frequency. By equating both the real and the imaginary part to zero we get from (1) two conditions from which the voltage swing V and the oscillating frequency ω can be determined.

We define, in accordance with common practice [1], the unloaded and external Q-factors Q_0 and Q_{ext}:

$$Q_0 = (\omega_0 C_r R_r)^{-1}, \qquad Q_{ext} = (\beta Z_0 \omega_0 C_r)^{-1},$$

where ω_0 is the resonant frequency from $\omega^2{}_0 L_r C_r = 1$ and Z_0 the characteristic impedance of the output line. If we define the normalized quantities r_l and x_l as $r_l = R_l/Z_0$ and $x_l = X_l/Z_0$, the two conditions can be written as:

$$R_g C_r \omega_0 = \frac{1}{Q_0} + \frac{r_l}{Q_{ext}}, \qquad (2)$$

$$X_g C_r \omega_0 + \frac{2\Delta\omega}{\omega_0} + \frac{x_l}{Q_{ext}} = 0\,. \qquad (3)$$

Here it is assumed that the detuning $\Delta\omega = \omega - \omega_0$ is small compared with ω_0, which is always the case in our applications. The equations (2) and (3) can be understood in the following way. As R_g is only a slowly varying function of frequency compared with the other variables, the value that the voltage swing will cause R_g to assume, depending on r_l, can be deduced from (2). The voltage swing together with X_g and x_l will determine the frequency of oscillation through (3).

We shall now consider the stability of the oscillator as regards both amplitude and frequency. We write [2] the fundamental component of the voltage V_g on the Gunn diode as

$$V_g = V \exp \mathrm{j}(\omega + \mathrm{j}\alpha)t.$$

Here α is the growth coefficient of V_g: a positive value of α is seen to represent a decay of V_g, a negative value a growth. For our purpose we assume α/ω to be a small quantity. If we consider that the oscillation condition (1) is momentarily satisfied, we can see whether the oscillation is stable or not by applying small perturbations δV and $\delta\omega$ in amplitude and frequency and calculating the sign of $\alpha/\delta V$. If this quantity is negative the oscillation is not stable since it does not come back to its initial condition. The initial terms of a series expansion of (1) yield:

$$\delta V\left(\frac{\partial Z_g}{\partial V} + \frac{\partial Z_c}{\partial V}\right) + \mathrm{j}(\delta\omega + \mathrm{j}\alpha)\left(\frac{\partial Z_g}{\partial \omega} + \frac{\partial Z_c}{\partial \omega}\right) = 0. \quad (4)$$

Some general implications of this formula, particularly for the study of noise characteristics, have recently been investigated [3] but we shall consider the special case of a linear load, $\partial Z_c/\partial V = 0$ (which is not valid for a varactor however) and where the real parts of both the Gunn impedance and the circuit impedance are independent of the frequency ($\partial R_g/\partial\omega = \partial R_c/\partial\omega = 0$). We then have two equations by equating both the real and the imaginary parts of (4) to zero, from which we obtain:

$$\frac{\alpha}{\delta V} = \frac{\partial X_g}{\partial V}\Big/\left(\frac{\partial X_g}{\partial\omega} + \frac{\partial X_c}{\partial\omega}\right), \qquad (5)$$

$$\frac{\delta\omega}{\delta V} = \frac{\partial R_g}{\partial V}\Big/\left(\frac{\partial X_g}{\partial\omega} + \frac{\partial X_c}{\partial\omega}\right). \qquad (6)$$

We saw from the above that the expression (5) must be positive for amplitude stability; the expression (6) will be used later on to discuss frequency stability. The denominator in these expressions is proportional to the stored energy in the Gunn diode ($\partial X_g/\partial\omega$) and in the circuit ($\partial X_c/\partial\omega$) [4]. The quantity $\partial X_g/\partial V$ is positive since the device reactance is negative (the Gunn diode is capacitive) and the reactance saturates to zero at high levels of voltage swing. (A similar behaviour is found with a klystron [1] or an IMPATT diode [5].) The stable condition for the amplitude is satisfied, therefore, if the sum of the stored energies is positive, which is generally the case unless there are tuned lossy circuits in the load. These may give rise to large negative values of $\partial X_c/\partial\omega$, so that the denominator in (5) and (6) might become negative.

We see then that the diode will oscillate if the circuit which is presented to it satisfies the conditions (2) and (3) and the stability conditions. In order to use these expressions it is necessary to know the functions $R_g(V,\omega)$ and $X_g(V,\omega)$.

Large-signal impedance measurements

The most direct method of obtaining the impedance functions R_g and X_g as a function of frequency and voltage swing is to measure the power output and frequency of the oscillator as a function of the load impedance seen by the diode, and to use the equation (1). This has been done by replacing the diode by an impedance-measuring circuit. The oscillator coaxial circuit is shown in *fig. 2*.

The encapsulated diode is mounted against a short-circuit and the circuit is tuned by two adjustable slugs. The short-circuit and diode are then replaced by a connector and the circuit impedance Z_c that was presented to the diode is measured. These measurements yield the diode impedance including the package impedance, a value that is of direct importance to the user. The separation of package impedance and impedance of the diode proper [6] is of importance to the designer. A simple approximation for a typical package, which remains valid up to 10 GHz and beyond, is that the package is equivalent to a short length of high-impedance line.

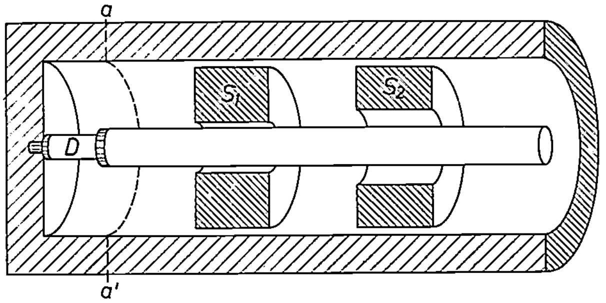


Fig. 2. Coaxial oscillator circuit for large-signal impedance measurement. The Gunn diode *D* is mounted against a short-circuit. The oscillator is tuned by means of the adjustable slugs S_1 (coarse tuning) and S_2 (fine tuning). To the right the circuit is matched to a 50 Ω line terminated by the load. After tuning, the cavity is separated at the plane *aa'* and the left-hand part, including the diode, is replaced by an impedance-measuring circuit.

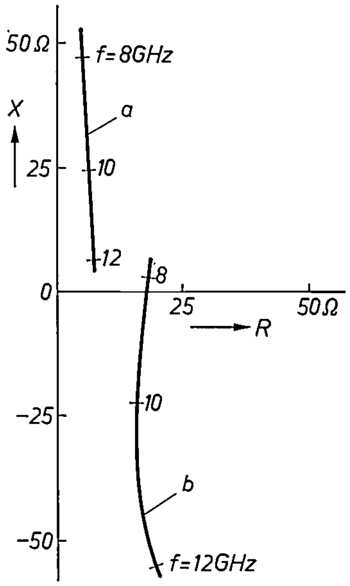


Fig. 3. Optimum loads for two Gunn-diode oscillators. The oscillation frequency *f* is plotted as a parameter along the curves. Curve *a* for a 20 mW r.f. power diode, curve *b* for a 200 mW r.f. power diode. *R* is the resistive and *X* the reactive component of the load.

The coaxial circuit can also be analysed by simple calculations, provided the losses in the cavity can be neglected. This avoids a rather cumbersome measuring technique when routine measurements are involved.

Some typical results of the optimum loads that produce the maximum power for two types of X-band devices are shown in *fig. 3*. The first type is a low-power diode (≈ 20 mW) with 2 W dissipated power and the second is a high-power device (≈ 200 mW) with 7 W dissipated power. From (1) the Gunn-diode impedance is the negative of the load and is therefore well represented by a negative resistance practically independent of the frequency from 8 to 12 GHz, in series with a reactance. The value of the resistance is about −7 Ω for a low-power device and −20 Ω for the high-power diode.

To obtain more detailed information about the variation of the function $Z_g(V,\omega)$ it is necessary to measure not only the load and output power as a function of frequency but to obtain these data for a much wider range of load values.

Results show that a Gunn diode is a well behaved device with values of R_g and X_g which fall to zero as the voltage swing increases. Contours of constant power and constant frequency can be drawn in the impedance plane. The power contours follow the lines of constant resistance, while the frequency contours approximately follow lines of constant reactance. This configuration is known as the Rieke diagram [7].

Some applications of Gunn oscillators

Mechanically tuned oscillators

Microwave oscillators can be tuned by mechanical adjustment of the resonating cavity, as is done in the measurements reported above. The circuit of fig. 2 can be tuned over a standard microwave band by fixing the two slugs together and moving the slug assembly. The power tuning can be achieved without changing the frequency by moving the two slugs slightly in opposite directions. Examples of tuning curves in X, C and Ku bands are shown in *fig. 4* together with the influence of the ambient temperature on the power.

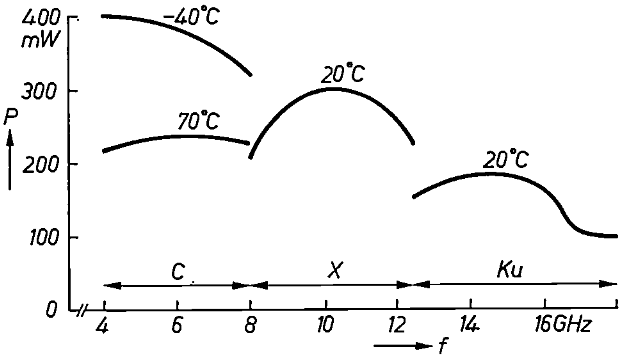


Fig. 4. The power *P* as a function of the frequency *f* in mechanically tuned Gunn-diode oscillators. The frequency range of the standard C, X and Ku microwave bands is indicated. The ambient temperature is indicated for each curve; the two curves for the C-band oscillator illustrate the effect of changes in ambient temperature.

[1] J. C. Slater, Microwave electronics, Van Nostrand, Princeton N.J. 1963, chaps. 3 and 4.

[2] J. Loeb, Ann. Télécomm. 6, 346, 1951.

[3] H.-J. Thaler, G. Ulrich and G. Weidmann, IEEE Trans. MTT-19, 692, 1971 (No. 8).

[4] C. G. Montgomery, R. H. Dicke and E. M. Purcell, Principles of microwave circuits, McGraw-Hill, New York 1948, section 5.23.

[5] D. de Nobel and M. T. Vlaardingerbroek, this issue, page 328.

[6] B. B. van Iperen and H. Tjassens, Philips Res. Repts. 27, 38, 1972 (No. 1).

[7] See for instance Slater [1], section 9.3.

Electronically tuned oscillators

Two methods of electronically tuning Gunn oscillators are currently used, one using a varactor, the other a YIG sphere [8]. Tuning could also be obtained by variation of the bias voltage. This, however, requires a variation of the bias voltage at a comparatively high current level and also gives rise to incompatible requirements for the Gunn diode and the associated microwave circuit. Tuning by means of the bias voltage requires relatively large frequency variations for a given voltage variation (the pushing factor $\partial\omega/\partial V$ must be large), while, on the other hand, the function of the bias voltage is to give stable oscillation and because of this the pushing factor must be as small as possible.

Varactor-tuned sources can be modulated very rapidly but have a limited tuning range and have a further disadvantage that losses and noise due to the varactor increase considerably as the tuning range increases. This can be seen from equation (3). The reactive load variation X_l which is given by the varactor, has a limited range (given by the range $(C_{max}-C_{min})/C_{max}$). For this variation to correspond to a large frequency variation, Q_{ext} must be small and so the losses and the noise will be large. It is possible to increase the tuning range [9] by adding parallel-tuned circuits. The primary application for these circuits is in a frequency-modulated source or frequency stabilization by feedback. In either case the frequency range required is quite small (2%). An example of a varactor-tuned circuit is given in *fig. 5*.

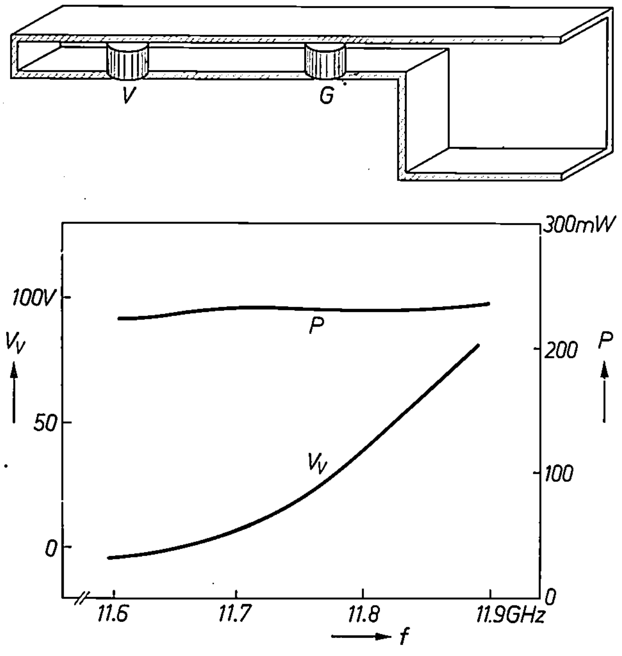


Fig. 5. Varactor-tuned Gunn oscillator. *Above:* the circuit, consisting of a short length of reduced-height waveguide in which the Gunn diode G and the varactor V, both in standard encapsulation, are mounted. *Below:* oscillator power P and tuning voltage V_v on the varactor as a function of the frequency f.

A wide-band tuning range can be obtained by using the magnetic tuning characteristics of YIG spheres. The frequency is linearly tunable by the magnetic field ($f = \gamma H$, where $\gamma = 2.8$ MHz/Oe). Since the Q-factor of a YIG-sphere circuit is fairly high and increases with frequency there is no degradation in noise characteristics as the tuning range is increased. There is the disadvantage that the electromagnet needed to tune the sphere is slow and requires comparatively high powers. The main application of such sources is therefore in frequency-sweep generators. Two methods of YIG tuning have been used. The first (*fig. 6*) uses the YIG sphere as the output coupling transformer to the load. In this case the Gunn circuit and load circuit must be coupled to the sphere in such a manner as to transform the 50 Ω impedance of the output line to 5 to 10 Ω in the Gunn-diode circuit. In the second method (also shown in fig. 6) the functions of tuning and matching the output are separated. The tuning is accomplished by a YIG sphere tightly coupled to the oscillator circuit and so presenting a variable pure reactance to the Gunn diode. The output is coupled through a fixed transformer. The advantage of the first method is that the harmonic content of the signal is filtered by the YIG transformer. R.F. saturation in the YIG sphere occurs at a lower power in the first circuit than in the second, however, and in the first circuit the positioning of the YIG sphere is very critical. An example of the tuning range using the second circuit is shown in *fig. 7*.

Fixed-frequency oscillators; FM noise

Many systems applications, such as local oscillators and Doppler-radar circuits, impose extra conditions on the Q-factor of the source in addition to the power and frequency requirements that have so far been considered. This is the stability of the signal frequency, both long term (drift) and short term (FM noise), which can be conveniently expressed as a function of frequency by a noise spectrum around the carrier. For telecommunication, for instance, the stability requirements in the carrier are imposed by the band separation and give rise to drift specifications of $\Delta f/f \ll 10^{-6}$. In Doppler-radar sources noise is also very important. Here the smallest velocity of a moving object that can be detected depends directly on the bandwidth of the emitted radiation. Hence the noise near the carrier must be very low (≈ -63 dB/Hz at 80 Hz from the carrier).

The FM noise is frequently expressed as a ratio of a mean excursion Δf_{rms} of the frequency to the modulating frequency f_m in a frequency-modulation system, measured in a certain bandwidth.

The noise characteristics of Gunn diodes are generally very unpredictable. The spread in the noise perform-

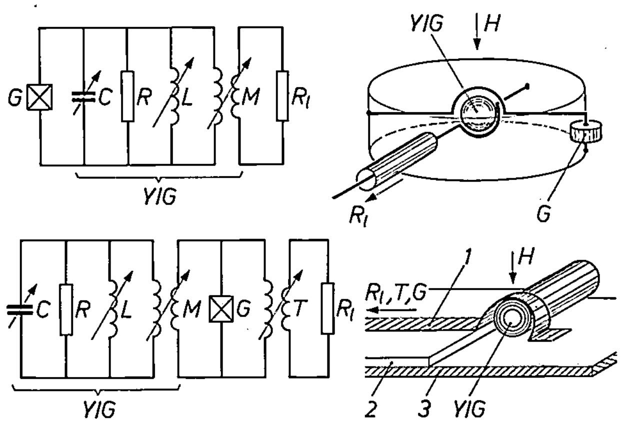


Fig. 6. Two types of YIG-tuned Gunn oscillator. In the equivalent circuits on the left the combination of C, R, L and M represents the YIG sphere. The method of mounting the YIG sphere in the circuit is shown on the right. The upper drawings refer to an oscillator in which the YIG sphere acts not only as the tuning element but also as a tunable band-pass filter coupling the load R_l to the oscillator. The sphere is placed between two loops, one containing the Gunn diode G, the other connected to the load through a coaxial cable. The circuit is enclosed in a metallic cylinder for shielding. The magnetic field H gives the tuning. The lower drawings show an oscillator in which the YIG sphere is used for tuning only, with a separate transformer T matching the load to the oscillator. This circuit is built in microstrip line. *1* microstrip, connected to a gold-band loop surrounding the YIG sphere; the sphere is mounted on a small dielectric rod. *2* substrate. *3* counter electrode. Here again H is the tuning magnetic field.

ance for otherwise similar diodes can be seen in *fig. 8*. These diodes all give about 100 mW at 16 GHz, but the noise varies by a factor of 50. It is therefore necessary for low-noise applications to select diodes, although recent work has shown us that proper attention to the contact technology and the oscillating mode can considerably improve the noise and stability performance [10].

The effect of the circuit on the noise is very important. We can see from equation (6) that an essential requirement for the reduction of frequency fluctuations is to increase $\partial X_g/\partial\omega + \partial X_c/\partial\omega$, an expression which is proportional to the total stored energy. The term $\partial X_g/\partial\omega$, giving the stored energy in the Gunn diode, is fixed, so this only leaves the possibility of increasing $\partial X_c/\partial\omega$. This can be done by either increasing the resonator Q-factor, or increasing the stored energy in the load. Both these methods are currently used to produce low-noise sources.

A useful circuit is obtained when a transmission cavity with a high Q-factor is put in series with the load [11]. This is illustrated in *fig. 9*. The Gunn oscillator is coupled to the load through a half-wavelength line and a high-Q transmission cavity. A stabilizing resistance is put in the line between the oscillator and the cavity. The position of this resistance is chosen such that it damps out all spurious oscillation modes in the

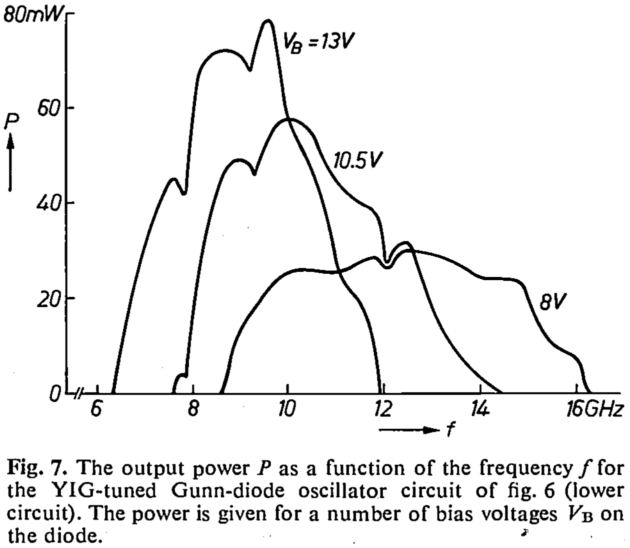


Fig. 7. The output power P as a function of the frequency f for the YIG-tuned Gunn-diode oscillator circuit of fig. 6 (lower circuit). The power is given for a number of bias voltages V_B on the diode.

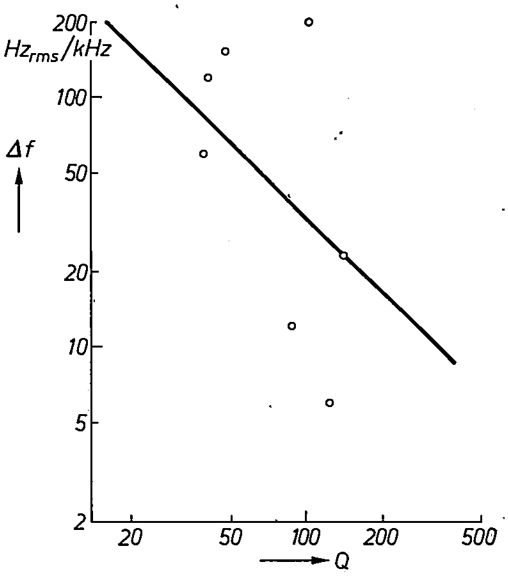


Fig. 8. FM noise Δf, r.m.s. value measured in a frequency band of 1 kHz centred at 100 kHz away from the carrier, as a function of resonator Q-factor for various Gunn-diode oscillators. Dots give measured values, the straight line shows the inverse proportionality between Q and the noise expected theoretically.

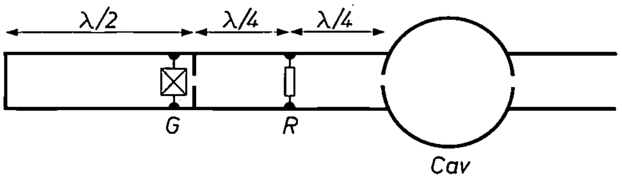


Fig. 9. Diagram of a Gunn-diode oscillator circuit. A stabilizing transmission cavity *Cav* is inserted between the Gunn diode G and the load. R is a stabilizing resistance to damp out unwanted modes in the oscillator.

[8] P. Röschmann, YIG filters; this issue, page 322.

[9] K. Kurokawa, Bell Syst. tech. J. 48, 1937, 1969.

[10] A. Mircea, J. Magarshack and A. Roussel, Device Research Conf., Michigan 1970.
M. Cathelin, J. Magarshack and J. K. Vogel, Proc. MOGA Conf., Amsterdam 1970, page 9.9.

[11] K. Schünemann and B. Schiek, Electronics Letters 7, 618 and 659, 1971 (Nos. 20 and 22), and 8, 52, 1972 (No. 3).

circuit and leaves only the desired mode unaffected. The stability and noise improvement can be expressed as a stabilization factor S, given by the ratio of the frequency fluctuation due to some varying reactances in the load without (index 1) and with (index 2) the stabilization cavity [12]. Thus:

$$S = \Delta f_1/\Delta f_2 = (E_0 + E_1)/E_0,$$

where E_0 represents the stored energy in the oscillator circuit and E_1 that in the stabilizing cavity.

We can rewrite this in terms of $\partial X/\partial\omega$:

$$S = \left(\frac{\partial X_0}{\partial\omega} + \frac{\partial X_1}{\partial\omega}\right) \Bigg/ \frac{\partial X_0}{\partial\omega},$$

which, for the circuit of fig. 9, gives [13]:

$$S = 1 + Q_{1,\mathrm{ext}} \Bigg/ \left\{ Q_{\mathrm{ext}} \left(1 + r_s \frac{Q_{1,\mathrm{ext}}}{Q_{1,0}}\right)\right\}, \qquad (7)$$

where $Q_{1,\mathrm{ext}} = 1/(\beta\ Z_0\ \omega_0\ C_1)$,
$Q_{1,0} = 1/R_1\ \omega_0\ C_1)$,
and r_s = normalized value of the stabilizing resistor.

The relation (7) shows that the stabilizing factor is given approximately by the ratio of the external Q-factor of the oscillating cavity and the stabilizing cavities, specially if the stabilizing Q is much the higher.

A typical result for an oscillator at 16 GHz gives 80 mW of output power, and a drift of 10^{-6}/°C with a gold-plated "Invar" cavity in a high-Q mode as the stabilizing cavity. The temperature range over which the source can operate is —40 °C to +100 °C for a power variation of 2 dB. The noise performance is shown in *fig. 10* with and without stabilization. The stabilization factor is about 100.

Amplification

The condition for linear amplification rather than oscillation in a device with a negative differential bulk mobility depends on the rate of growth for a domain and on the time available (transit time for a perturbation in the space-charge density).

In gallium arsenide this results [14] in the condition on the dopant impurity concentration n_d and the diode length l:

$$n_d l < 10^{12}\ \mathrm{cm}^{-2}. \qquad (8)$$

Above this limit domains can form and consequently oscillations will occur. Amplifying diodes which satisfy the above condition are known as undercritically doped amplifiers [15].

A second type of amplifier uses a planar technology (N-doped on a semi-insulating substrate, where the current is parallel to the interface between two evaporated contacts). It has been shown [16] that domain formation is inhibited if the thickness d of the N layer is such that $n_d d < 1.6 \times 10^{11}$ cm^{-2}. This is because the insulating substrate forms a dielectric short-circuit to the voltage across the active layer. Several useful devices have been constructed on this principle [17], and an interesting variation is reported [18] that gives an extra stabilization by adding a layer with a high dielectric constant on top of the active film.

Some years ago an amplifier based on the following was described in the literature [19]. It has been found that a Gunn diode oscillating at a certain frequency presents a negative resistance at a lower frequency. In the example quoted, there was oscillation at 9.6 GHz and amplification at about 6 GHz. An essential condition in practice is that no power at the oscillating frequency must appear at the output.

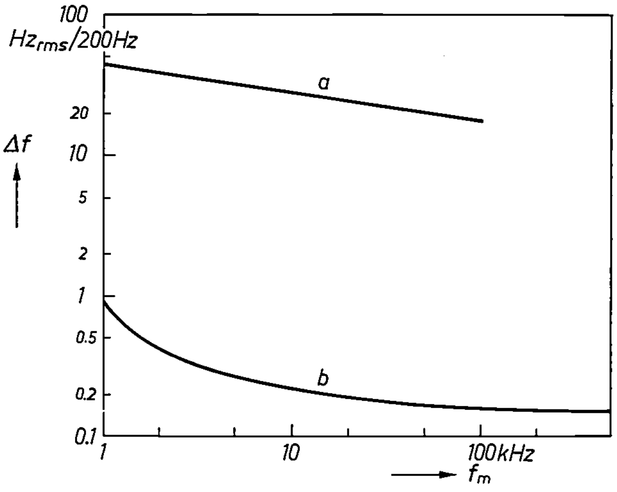


Fig. 10. FM noise Δf, r.m.s. value measured in a 200 Hz bandwidth centred at a frequency f_m away from the oscillating frequency (16.5 GHz) of the Gunn-diode oscillator of fig. 9. Curve *a* without the stabilization cavity, curve *b* with stabilization cavity. A frequency-stabilization factor of about a 100 is obtained at the expense of a reduction in output power by a factor of 3.

We shall deal here with still another type of amplification in overcritically doped diodes, which are very similar to those used for oscillators and where condition (8) is not satisfied. Moving domains do not form, however, and the device can be represented by a stable negative resistance which for sufficiently small signals is independent of the applied signal voltage [20]. The stabilization mechanism here is due to the effect of the contacts, a doping profile or a temperature gradient in the active region. It has been shown [21] that if the contact limits the injection of electrons at the cathode, a high stationary field is produced at the anode, which inhibits moving domains. A. Mircea has shown that even when the contacts are normal, a stable solution to the transport differential equation in which a high field at the anode is also present can be found by computer [22]. This is because in the anode region, where there is a steep doping gradient, the diffusion currents are high and inhibit the formation of a fully grown domain.

Whatever the precise mechanism may be, the devices themselves, when measured in a mount similar to that of fig. 2 but without the matching slugs, show a wide-band negative resistance which is a function of the bias voltage. *Fig. 11* shows the variation of the impedance as a function of bias and frequency. It is seen that the resistance crosses to the negative region as the bias is increased and then passes through a maximum negative resistance before coming back to a positive value. The diode itself remains capacitive and does not pass through a natural resonance (pure resistance) while it is in the negative region; this shows that it is unconditionally stable. The bandwidth of the negative resistance can be seen in *fig. 12*. It can be seen that a single diode can show a negative resistance from 7 to 18 GHz. Such a diode in a simple taper-matched reflection-type amplifier (*fig. 13*) has a constant gain of 10 dB up to input powers of 10 mW. The variation in gain with temperature is shown in *fig. 14*. For different diodes the gain can increase or decrease with temperature. Theor-

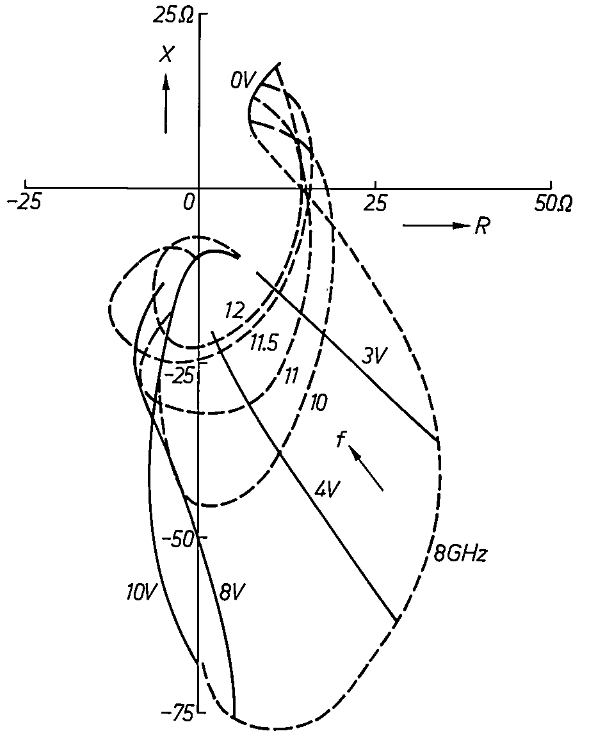


Fig. 11. Small-signal impedance of a stable, amplifying Gunn diode. R resistive component and X reactive component of the diode impedance. The solid lines are for constant bias voltage V_B on the diode, the dotted lines for constant measuring frequency f. On the left of the vertical axis the diode has negative resistance and signal amplification can be obtained.

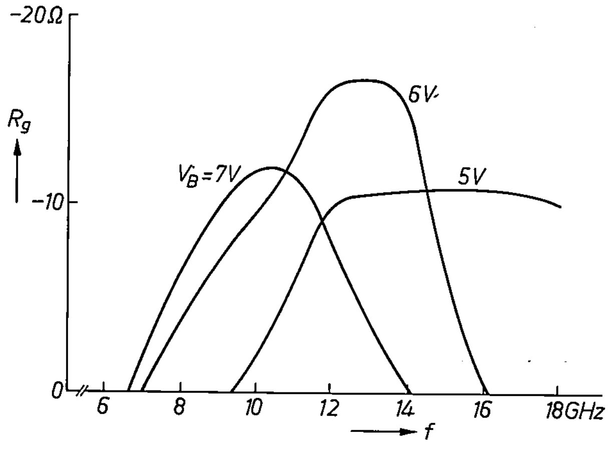


Fig. 12. Negative resistance R_g of a Gunn diode as a function of the frequency f of the measuring signal, for three levels of the bias voltage V_B.

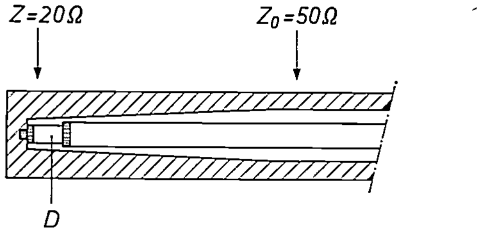


Fig. 13. Reflection-type amplifier. The coaxial outer conductor is tapered, to match the Gunn-diode impedance Z to the transmission-line impedance Z_0. The reflection gain G is given by $G = |Z - Z_0|^2/|Z + Z_0|^2$. For negative Z the value of G is larger than unity and the reflected signal is amplified.

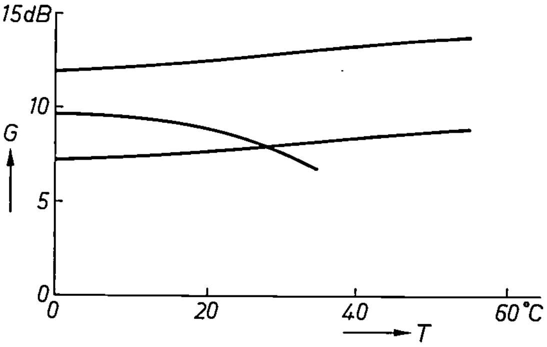


Fig. 14. Variation of gain G as a function of ambient temperature T for three Gunn diodes in a reflection-type amplifier.

[12] I. Goldstein, IRE Trans. MTT-5, 57, 1957.

[13] E. J. Shelton, Jr., Trans. IRE ED-1, No. 4, 30, 1954.

[14] This condition is discussed in detail by G. A. Acket, R. Tijburg and P. J. de Waard, this issue, page 370.

[15] H. W. Thim and M. R. Barber, IEEE Trans. ED-13, 110, 1966.

[16] G. S. Kino and P. N. Robson, Proc. IEEE 56, 2056, 1968.

[17] R. H. Dean, A. B. Dreeben, J. F. Kaminski and A. Triano, Electronics Letters 6, 775, 1970.

[18] K. R. Hofmann, Electronics Letters 5, 289 and 469, 1969.

[19] H. W. Thim, IEEE Trans. ED-14, 517, 1967.

[20] B. S. Perlman, 1970 IEEE Int. Solid-State Circuits Conf., Philadelphia, page 136.
J. Magarshack and A. Mircea, *ibid.*, page 134.

[21] H. Kroemer, IEEE Trans. ED-15, 819, 1968.

[22] See the section on accumulation layers and domains in: J. de Groot and A. Mircea, this issue, page 390.

etically it has been shown [23] that the gain passes through a maximum at a temperature of about 200 °C in the active zone, which would account for a small increase or decrease in gain depending on whether the temperature of the device is greater or less than this value. The measured noise figure is 15 dB.

Although the amplifiers are much less advanced than the Gunn oscillator, some recent wide-band amplifiers announced in the literature [24] show that they are already competitive to low-power travelling-wave tubes.

[23] J. Magarshack and A. Mircea, Proc. MOGA Conf., Amsterdam 1970, page 16.19.

[24] B. S. Perlman, C. L. Upadhyayula and R. E. Marx, IEEE Trans. **MTT-18**, 911, 1970.

Summary. Gunn diodes, for certain values of the bias voltage, present an impedance with a negative resistive component to the microwave circuit in which they are incorporated. This allows these diodes to be used as active elements in oscillator and amplifier circuits. For the design of an oscillator circuit numerical values of the diode impedance are required. These values can be derived from the impedance of a simple oscillator circuit, as is seen from the conditions for oscillation. Further analysis of these conditions yields relations for amplitude and frequency stability of an oscillator. Three types of oscillator circuit are discussed: mechanically tuned, varactor-tuned and YIG-tuned circuits. Finally a fixed-frequency stable microwave source is discussed which incorporates a high-Q cavity for frequency stabilization.

For oscillator applications the level of dopant impurities in the diode must be above a certain minimum value. Diodes with a doping level below this value can be used for amplification (undercritically doped diodes). But if the propagation of high-field domains can be suppressed even overcritically doped diodes can make stable amplifiers. A number of these possibilities are discussed in some detail.

P-I-N switching diodes in phase-shifters for electronically scanned aerial arrays

J. H. C. van Heuven

For some microwave applications the long-established rotating parabolic aerial is gradually being superseded by aerial systems with electronic beam control (phased arrays). These systems consist of a large number of coherently radiating elements, and the beam emitted is produced by interference between the radiation from the individual elements [1]. The phase differences between the radiation of these elements determine the attitude of the beam in relation to the array. Changes in these phase differences thus make it possible to alter the direction of the beam.

The article below is about the variable phase-shifters generally used in phased arrays and the *P-I-N* diodes employed in them as electronic switches. A brief discussion of the details of a phased array is given at the beginning of the article as background material.

Phased arrays

A variable phase-shifter (*fig. 1*) can be incorporated in the feeder of any aerial element to produce the phase differences needed for changing the direction of the beam [2].

The purpose of the phase-shifters is purely and simply to bring about a phase difference between the signals of the successive elements in an aerial array. Phase differences between input and output of the phase-shifter have no effect as long as they are identical for all phase-shifters when in the neutral position. When a particular phase shift is referred to in the following, it is not the absolute shift caused by a circuit that is intended but only the difference in phase shift between different circuits or different states of one circuit.

Since the beam has a finite aperture, a continuous change of direction is not necessary; discontinuous changes equal to the beam aperture are sufficient. Discontinuous phase-shifters can therefore be used, which consist of circuits whose transmission characteristics can be changed in steps by means of switches. The advantage of such stepwise changes is that they are easy to obtain with the required accuracy and reproducibility. Moreover, the switches in such digital circuits can be controlled by a computer. This means that the scanning movement of the beam can be very much faster than would ever be possible with mechanical aerials. It also makes it possible to use the same array simultaneously for different functions. For example it might be used for scanning a large area rapidly but roughly while at the same time carefully following one or more objects in that area.

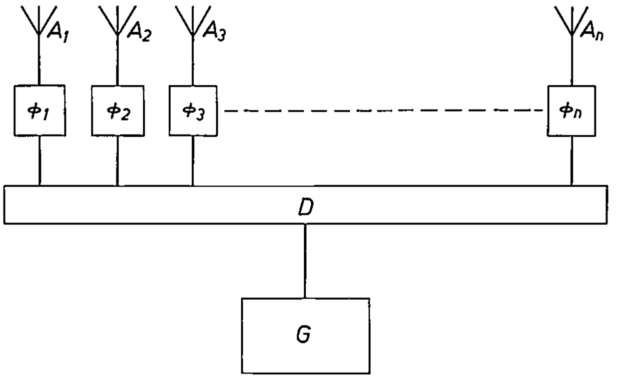


Fig. 1. Block diagram of an electronically scanned aerial array (phased array). G microwave source. D power divider. $\phi_1 \dots \phi_n$ electronically controlled phase-shifters. $A_1 \dots A_n$ aerial elements.

If the phase-shifters are not to be too complicated, the number of phase steps should not be too large and the steps themselves therefore not too small. The smallest phase step in practice is usually 22.5° which, given a distance of half a wavelength between the elements in the array, corresponds to steps of 7° in the beam direction (*fig. 2a*). Smaller steps in beam direction can be obtained without the need for smaller phase steps by causing a group of adjacent elements to radiate in phase and then making the phase difference between successive groups 22.5° (fig. 2*b*).

The elements of an array can be placed in a line; the direction of the beam can then vary in a plane through this line. If the elements are arranged in a two-dimensional matrix, which is usually the case, the beam can be made to cover a large part of the space on one side of the matrix plane.

Ir. J. H. C. van Heuven is with Philips Research Laboratories, Eindhoven.

[1] See A. Meijer, An analogue computer for simulating one-dimensional aerial arrays, Philips tech. Rev. **31**, 2-16, 1970.

[2] A general review of phased arrays is given in: L. J. Hardeman, Microwaves **9**, No. 6, 38, June 1970; see also K. L. Fuller, Philips tech. Rev. 32, 13, 1971 (No. 1).

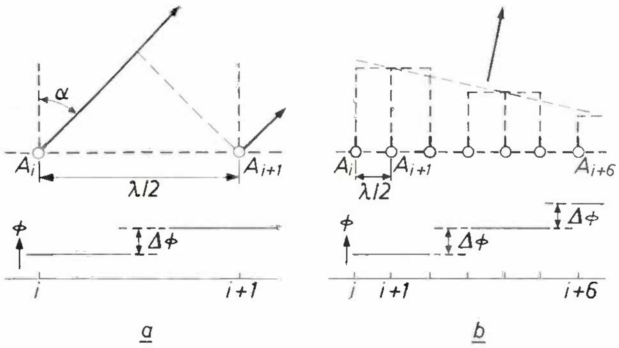


Fig. 2. *a*) The direction of the total beam emitted by aerial elements in a linear array is determined by the phase difference between two consecutive elements. If the spacing between the elements is $\lambda/2$, the emitted beam will be at an angle α to the normal of the system if 180° sin α is equal to the phase difference $\Delta\phi$ between two consecutive elements. *b*) When consecutive groups of elements radiate in phase and the phase difference between consecutive groups is $\Delta\phi$, smaller changes in the direction of the emitted beam can be obtained than when the same phase difference $\Delta\phi$ exists between consecutive elements.

The dimensions of aerial arrays vary considerably, depending on the application and the signal frequency. The smallest arrays that have been produced to date have dimensions of about 50 × 50 cm. Designed specially for aircraft radars, they consist of a few hundred elements, operate at frequencies between 10 and 20 GHz and can deliver a peak power of 10 kW. The largest arrays at present in use measure 30 × 30 m; they contain more than 10 000 elements, operate at frequencies between 300 and 1000 MHz, and can deliver a peak power of more than 1 MW.

With large numbers of elements in an array the phase-shifters have to be inexpensive, since one is usually required for each element. Moreover, their dimensions in at least two directions should not be greater than half the wavelength of the emitted radiation. The reason for this requirement is that the spacing between two aerial elements in an array should be equal to half a wavelength to avoid unwanted side lobes in the radiation pattern of the system. If the phase-shifters are larger than half a wavelength it is difficult to arrange the elements correctly.

The phase-shifters also have to meet certain electrical requirements if they are to be used in aerial arrays. The attenuation and reflection of the signal must be minimal and should as far as possible be the same at all settings of the phase-shifter. In many cases the phase-shifters, and hence the switches used in them, have to be capable of handling a high peak power, and the switches are also required to be fast. As we shall see, the last two requirements are often contradictory.

Finally, for radar applications the elements and all their constituent components are required to be reciprocal, i.e. the characteristics of an element should not depend on the direction of propagation of the signal, since an aerial array must be able to both transmit and receive. If the phase-shifters are not reciprocal, the paths for the transmitted and the received signal must be separated, e.g. by means of circulators.

Before looking at the operation and construction of digital phase-shifters, we shall deal first with the *P-I-N* diode used in them as switches.

The *P-I-N* diode

Like other semiconductor diodes and transistors, this diode owes its name to its construction: a layer of *P*-type and a layer of *N*-type material, and between them a virtually intrinsic region, which is in fact a weakly doped *P*-type or *N*-type region (*fig. 3*).

The d.c. characteristic of a *P-I-N* diode has the same shape as that of a *P-N* diode. Both types of diode also give the same a.c. behaviour at frequencies below 1 kHz; the relation between current and voltage is determined by the d.c. characteristic in both cases, and rectification occurs.

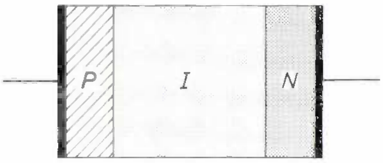


Fig. 3. Schematic cross-section of a *P-I-N* diode. Between two regions of *P*-type and *N*-type conduction there is a layer of intrinsic material *I*. Metal contacts are applied to the *P*-type and *N*-type layers.

If, however, the diode is connected to a d.c. voltage on which is superimposed an a.c. voltage with a frequency above 1 MHz the *P-I-N* diode behaves in an essentially different way (*fig. 4*). There is no longer any rectifying action, and the diode forms a constant impedance to the a.c. voltage, even at high amplitude. The magnitude of this impedance is determined by the d.c. voltage. This is the case even to the extent that an a.c. voltage of large amplitude, superimposed on a low d.c. voltage in the forward direction, gives currents in the reverse direction during the negative periods of the a.c. voltage. Conversely, when the d.c. voltage is in the reverse direction, no currents will flow in the forward direction even though the amplitude of the a.c. voltage would at first sight seem likely to give rise to such currents.

Because the impedance of the *P-I-N* diode differs so considerably from one extreme setting to the other, the device can be used as a switch. It is capable of switching high powers with relatively low losses, and for this reason is particularly suitable for use as a switch in digital phase-shifters.

The high-frequency behaviour of the *P-I-N* diode may be explained as follows. When a d.c. current flows in the forward direction, a reserve of charge in the *I* layer keeps current flowing during the short period when the negative peaks of the a.c. voltage give a voltage in the reverse direction. Conversely, when the d.c. voltage is in the reverse direction, a charge-free zone exists in the *I* layer. During the periods when the peaks of the a.c. voltage give a forward voltage the charge carriers will diffuse over such a short distance in the intrinsic region that no conduction will result. The behaviour of the *I* layer is determined by the thickness of the layer and by the lifetime τ of the charge carriers in it, and therefore the periods referred to will be short when the frequency ω of the a.c. voltage is so high that $\omega\tau$ is much greater than 1. In practice τ is about 1 µs.

The forward d.c. voltage situation can be described mathematically on the basis of the continuity equation for the charge Q in the I layer:

$$\frac{\mathrm{d}Q}{\mathrm{d}T} + \frac{Q}{\tau} = i(t)\,,$$

where $i(t)$ is the current through the layer. If as above $i(t)$ is assumed to be a d.c. current superimposed on an a.c. current: $i(t) = I_0 + I_1 \cos \omega t$, the solution of the equation is:

$$Q = Q_0\,\mathrm{e}^{-t/\tau} + I_0\tau + \frac{I_1\,\tau}{(1+\omega^2\tau^2)^{\frac{1}{2}}} \cos(\omega t + \phi)\,.$$

The first term represents the switching transients, the other two terms give the steady-state solution. Here ϕ is a phase angle, which is not relevant to the present discussion.If the frequency ω is so high or τ so great that $\omega\tau \gg 1$, the solution may be written in the approximate form:

$$Q = I_0\tau + (I_1/\omega) \cos \omega t.$$

If $I_0 \gg I_1\omega$, the first term will completely determine the situation. Since we have assumed that $\omega\tau \gg 1$, this will also be the case when the amplitude I_1 of the a.c. current is several times greater than the d.c. current I_0. The charge density in the I layer is then completely determined by the d.c. current and by the lifetime of the charge carriers and will thus be constant. Hence the impedance of the layer is constant, and the diode behaves as a linear circuit element.

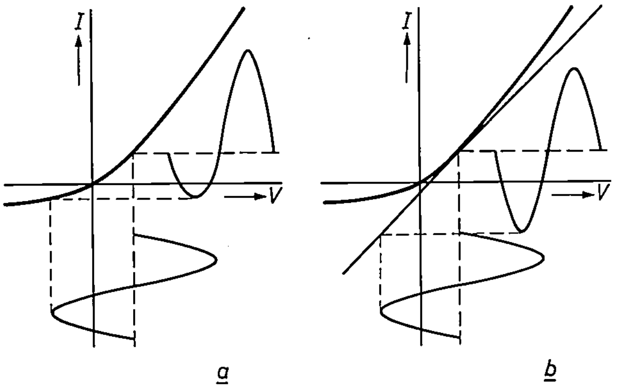


Fig. 4. Relation between current *I* and voltage *V* in a *P-I-N* diode when a d.c. voltage is applied with an a.c. voltage superimposed upon it. *a*) For a.c. voltages at frequencies below 1 kHz the behaviour is determined by the d.c. voltage characteric; rectification occurs. *b*) For a.c. voltages at frequencies above 1 GHz the diode behaves as a linear device whose impedance is determined by the slope of the d.c. voltage characteristic at the operating point.

A reverse d.c. voltage gives rise to a charge-free zone in the I layer, and the device then behaves like a capacitor. Its capacitance will depend on the thickness of the charge-free zone and consequently on the value of the applied d.c. voltage.

This linear behaviour makes the equivalent circuit of the diode at high frequencies very simple, provided the representation is confined to a fixed frequency (*fig. 5*). Apart from the stray inductance L_p and the stray capacitance C_p, both due to the encapsulation of the diode, all we have when the d.c. voltage is in the forward direction is a resistance R_1. When the d.c. voltage is in the reverse direction, we have instead of R_1 a capacitor C_2 in series with a resistance R_2.

To minimize the losses in the diode, and hence in a phase-shifter in which the diode is used as a switch, it is necessary to keep the values of R_1, R_2 and C_2 as small as possible. The quality of a microwave switch can be expressed in a figure of merit q, in which the difference in impedance between the open and closed state is compared with the resistances in the open and closed state. These resistances are a measure of the energy dissipation in the switch. In our case the figure of merit takes the form:

$$q = \left[\frac{(R_1 - R_2)^2 + (1/\omega C_2)^2}{4\,R_1R_2}\right]^{\frac{1}{2}}.$$

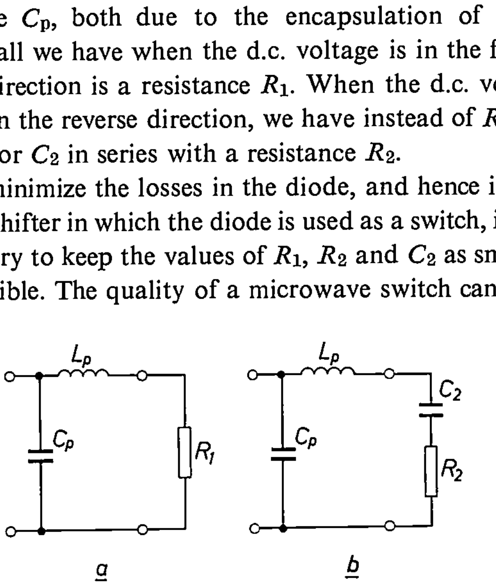


Fig. 5. Equivalent circuit of a *P-I-N* diode for a single fixed high frequency. L_p and C_p stray inductance and capacitance; the values of L_p and C_p depend on the way in which the diode is mounted in the circuit. *a*) When there is a d.c. voltage in the forward direction the r.f. current encounters only a low resistance R_1. *b*) If the d.c. voltage is in the reverse direction, the r.f. current encounters the high resistance R_2 and the capacitance C_2 in series.

In this definition of q the stray elements (in our case L_p and C_p) have no effect on the value of q, provided these elements are loss-free.

The figure of merit used here for a switch, defined above in a special form, can be more generally defined by the expression

$$q^2 = (Z_1 - Z_2)\,(Z_1{}^* - Z_2{}^*)/(Z_1 + Z_1{}^*)\,(Z_2 + Z_2{}^*)\,.$$

Here Z_1 and Z_2 are the complex impedances of the switch in the open and closed states; Z_1^* and Z_2^* are the conjugate complex values of Z_1 and Z_2. A great advantage of the factor q thus defined is that it does not change when a lossless network is connected in cascade (*fig. 6*) [3] [4]. In the situation shown, where the two impedances Z_1 and Z_2 of the switch at one end of the network N are transformed into the impedances Z_1' and Z_2' at the other end of the network, we thus have:

$$q^2 = \frac{(Z_1 - Z_2)(Z_1^* - Z_2^*)}{(Z_1 + Z_1^*)(Z_2 + Z_2^*)} = \frac{(Z_1' - Z_2')(Z_1'^* - Z_2'^*)}{(Z_1' + Z_1'^*)(Z_2' + Z_2'^*)}.$$

The stray elements L_p and C_p in the equivalent diagram of a diode may be regarded as a loss-free network, and therefore it is evident that they do not affect the value of the figure of merit of the switch.

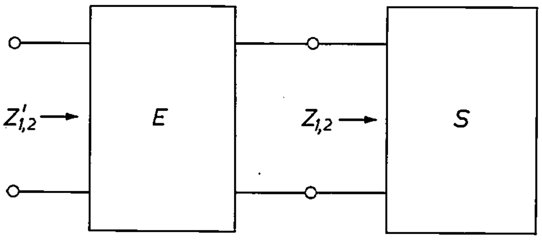


Fig. 6. Microwave switch S with associated network E. The impedance values Z_1 and Z_2 of the microwave switch in the open and closed states are seen at the other end of E as impedances Z_1' and Z_2'. If E is lossless, this network does not affect the value of the figure of merit q of the switch, defined in the text.

Since R_1, R_2 and $1/\omega C$ change in approximately the same way when the area of the *P-I-N* junction is changed, this provides us with a means — subject to certain limitations — of giving these quantities the optimum values for a given circuit.

In addition to the losses in the circuit, which can be expressed in the figure of merit q, an important parameter of the *P-I-N* diode is the switching time, i.e. the time needed to change from the conducting to the non-conducting state. The switching time depends mainly on the thickness of the *I* layer and on the lifetime of the charge carriers in this layer.

This can be understood as follows. If the diode is passing d.c. current in the forward direction (switch closed) charge carriers from the *P* and *N* regions are injected into the *I* region. If the conduction through the diode is to stop once the applied d.c. voltage has been reversed, these charge carriers must first have disappeared from the *I* region. The reverse field attracts charge carriers from the *I* layer. As soon as a charge-free zone has formed in this layer, however, the total voltage will be almost entirely across this zone, and so the rest of the *I* layer will have become virtually field-free. Consequently the field acting on the electrons still present in the *I* layer will be small, so that these electrons can only disappear by recombination or diffusion. The speed of the recombination process is characterized by the lifetime τ of the charge carriers, and the diffusion time is determined by the diffusion path, which is related to the thickness of the *I* layer.

Another consequence of this situation is that diodes capable of switching a high power are usually slower in their response than diodes that switch low power, since high-power switching necessitates the switching of high voltages and currents. To avoid electrical breakdown in an open switching diode, a thick *I* layer is necessary. To keep the losses in the thick *I* layer within reasonable bounds, the charge density in this layer must be sufficiently high, which implies that the lifetime τ will have to be relatively long. A long τ and a thick *I* layer both have the effect of reducing the switching speed.

Again, a diode for switching high currents is usually slower than one that switches low currents. The higher the currents to be handled by a diode, the greater should its cross-section be, and for this reason a high-current diode has a relatively smaller outside surface area. Since the surface recombination of charge carriers is not negligible compared with the bulk recombination, the chance of recombination in low-current diodes is greater than in high-current diodes. This disadvantage can be overcome by doping the material of the *I* layer with selected impurities that act as recombination centres but have no effect on the concentration of the free charge carriers.

A typical switching time for a diode capable of handling a power of 1 kW is about 0.1 microsecond, whereas a diode handling a power of 1 W needs only a few nanoseconds.

Phase-shifters for microwave signals

There are various circuits that can be used as phase-shifters for microwave signals. At Philips Research Laboratories microstrip circuits [5] have been used in which the phase shift is obtained either by altering the reflection coefficient at the termination of a line, or by altering the phase velocity along a line (variable delay line). The use of microstrip meets the requirements of compactness imposed on a phase-shifter for use in the aerial arrays mentioned above. The phase-shifters described here are made on an insulating non-magnetic substrate, and the characteristics of the circuit are varied by means of *P-I-N* diodes [6].

It is found in practice that the variable-reflection type of phase-shifter is better for shifts of 90° and 180°, whereas phase-shifters based on a variable delay line are better for phase shifts of 22.5° and 45°. These two circuits will be discussed first, and next an experimental phase-shifter formed from a combination of these circuits will be described, which can be used to vary the phase of a signal in steps of 22.5° from 0° to 360°.

Phase shift by variable reflection

The reflection of a signal at the termination of a transmission line is determined both in magnitude and in phase by the terminating impedance. Variation of this impedance thus provides a means of making a variable phase-shifter, provided the incident and reflected signals can be separated. The reflection may be characterized by the complex reflection coefficient ϱ, defined as the ratio of the complex values A and B of the incident and reflected signal voltages. It is the usual practice to take the normalized values of A and B, obtained by dividing by the characteristic impedance of the transmission line, Z_0. The reflection coefficient is given by:

$$\varrho = |\varrho|\, e^{j\phi} = \frac{B}{A} = \frac{z-1}{z+1},$$

where ϕ is the phase difference between the incident and the reflected signal, and z the normalized terminating impedance. If we could use an ideal switch as the terminating impedance (*fig. 7a*) we could then change z from $z = 0$ to $z = \infty$, so that the reflection coefficient ϱ would change from -1 to $+1$, and ϕ would change by 180°. For other phase angles it is necessary to use circuits like those illustrated in fig. 7*b*, in which the terminating impedance is changed from a positive to a negative reactance. Since microwave inductors and capacitors are difficult to make in practice, however, a short-circuited transmission line is generally used as the reactance (fig. 7*c*). The length d of this line determines the phase shift that takes place upon the opening or closing of the switch. The length l of the transmission line is irrelevant since we are concerned only with the differential phase shift.

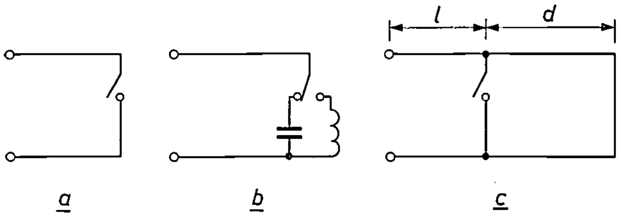


Fig. 7. Reflecting short-circuited transmission line as a phase-shifter. *a*) Transmission line terminated with an ideal switch. *b*) Line terminated with an inductor or a capacitor. *c*) Line with short-circuited termination; the reflcction characteristics are determined by the position of the switch.

Unlike the ideal circuit of fig. 7*a*, which produces a phase shift completely independent of the frequency of the microwave signal, the circuits in 7*b* and *c* give a phase shift that does depend on frequency. The frequency dependence can be minimized, however, by suitable dimensioning of the elements.

In practical circuits of this type the switch is a *P-I-N* diode. As can be seen from the microwave equivalent circuits for this diode (fig. 5), it is certainly not an ideal switch, and for phase shifts of 90° and 180° the circuit in fig. 7*c* has to be used. It is relatively easy to find a relation between the losses occurring in this circuit and the differential phase shift and the figure of merit q of the switch.

This relation can be calculated for a much more general situation [7], but here the calculation will be limited to the case of a variable reflection-type phase-shifter.

If we again normalize the terminating impedance of the line by dividing by Z_0, the characteristic impedance, we arrive at the following expressions for the voltage V across the line and the current I flowing in it:

$$\begin{aligned} V &= A + B = A\,(1+\varrho), \\ I &= A - B = A\,(1-\varrho). \end{aligned} \qquad (1)$$

Here A is the amplitude of the incident signal, B is the amplitude of the reflected signal, and ϱ is the reflection coefficient of the termination of the line. If we let the terminating impedance z assume successively the values z_1 and z_2 and add the subscript 1 or 2 to all associated quantities, then by substituting from (1):

$$\begin{aligned} |V_2I_1 - V_1I_2| &= 2\,A^2\,|\varrho_1 - \varrho_2| = \\ &= 2\,A^2 \left[(|\varrho_1| - |\varrho_2|)^2 + 4\,|\varrho_1||\varrho_2| \sin^2 \frac{(\phi_1 - \phi_2)}{2}\right]^{\frac{1}{2}}. \end{aligned}$$

Assuming now that $|\varrho_1| = |\varrho_2| = |\varrho|$, which implies that the losses in both cases are the same, we can simplify this expression to:

$$|V_2I_1 - V_1I_2| = 8\,P\,|\varrho| \sin\frac{\Delta\phi}{2}, \qquad (2)$$

where $P = \frac{1}{2}A^2$, the power of the incident signal, and $\Delta\phi = \phi_1 - \phi_2$, which is the phase shift when z_1 is switched to z_2. We have thus found an equation that gives a relation between the power P and the associated currents and voltages for a given phase shift $\Delta\phi$.

If the absolute value of the reflection coefficient in both states is identical, the lower limit of the losses involved can be calculated. For this purpose we can rewrite expression (2) as:

$$|I_1I_2| \left|\frac{V_2}{I_2} - \frac{V_1}{I_1}\right| = |I_1I_2|\,|z_2 - z_1| = 4\,A^2\,|\varrho| \sin\frac{\Delta\phi}{2}. \qquad (3)$$

Using the conjugates of voltage, current and the impedances z_1 and z_2 we can write the power dissipated in the impedance z as:

$$\left.\begin{aligned} \tfrac{1}{2}(V_1I_1^* + V_1^*I_1) &= \tfrac{1}{2}\,|I_1|^2\,(z_1 + z_1^*) = A^2(1 - |\varrho|^2), \\ \tfrac{1}{2}(V_2I_2^* + V_2^*I_2) &= \tfrac{1}{2}\,|I_2|^2\,(z_2 + z_2^*) = A_2(1 - |\varrho|^2). \end{aligned}\right\} \qquad (4)$$

The fraction of the incident power that is dissipated on reflection is now given by $1 - |\varrho|^2$. Substituting the equations (4) in (3) and using the figure of merit q defined earlier and introducing the

[3] T. E. Rozzi, J. H. C. van Heuven and A. Meijer, Proc. IEEE **59**, 802, 1971 (No. 5).

[4] K. Kurokawa and W. O. Schlosser, Proc. IEEE **58**, 180, 1970.

[5] J. H. C. van Heuven and A. G. van Nie, Microwave integrated circuits, this issue p. 292.

[6] Phase-shifters in the form of microstrip circuits, with no semiconductor devices, are described by M. Lemke and W. Schilz in: Microwave integrated circuits on a ferrite substrate, this issue p. 315.

[7] See M. E. Hines, Proc. IEEE **55**, 473, 1967, and also the article by Kurokawa and Schlosser quoted in note [4].

notation $\sin^2(\frac{1}{2}\Delta\phi)/q^2 = C$, we find this fraction to be:

$$1 - |\varrho|^2 = 2\,C\,\{(1 + 1/C)^{1/2} - 1\}. \tag{5}$$

This equation thus expresses the losses in the circuit of fig. 7*c* in terms of the figure of merit q of the switch used and the differential phase shift.

If a lossless network is inserted in cascade with the switch there is no change in the expressions in the right-hand sides of equations (2) and (5); this is consistent with the fact that q does not change, as we saw earlier.

If a reflecting line is to be used as a phase-shifter, the incident and the reflected signals must be separated. This can be done by means of a hybrid ring [5], which can readily be made in microstrip. The circuit then has the form illustrated in *fig. 8*. When ports *2* and *4* are terminated by identical impedances there is no coupling between the input port *1* and the output port *3*. If, however, ports *2* and *4* are terminated by unequal impedances, part of the energy entering at *1* will leave the circuit through port *3*. The phase difference between the incident and outgoing signal is determined by the reflection coefficients at *2* and *4*. If ports *2* and *4* are terminated by impedances z and $1/z$, the reflection coefficients at these ports are thus reciprocal and the resultant signals at port *3* will be in phase, so that the output signal there is at a maximum. The impedance $1/z$ terminating port *4* is obtained by transforming an impedance z through a quarter-wave transformer.

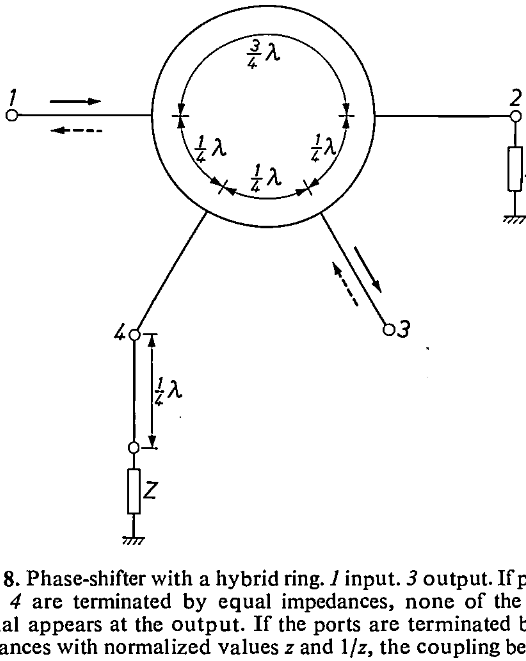


Fig. 8. Phase-shifter with a hybrid ring. *1* input. *3* output. If ports *2* and *4* are terminated by equal impedances, none of the input signal appears at the output. If the ports are terminated by impedances with normalized values z and $1/z$, the coupling between input and output is at a maximum. The impedance $1/z$ at port *4* is obtained by connecting an impedance z through a quarter-wave transformer.

Phase shift with variable delay line

As mentioned above, phase shifts of 22.5° and 45° are more readily obtained with a variable delay line [8] than with reflecting elements. The operation of a variable delay line as a phase-shifter depends on the reduction in the phase velocity of a signal along a transmission line that is obtained when the line is loaded. A signal travelling along a loaded line will consequently lag in phase behind the signal travelling along an unloaded but otherwise identical line.

Let us suppose that a section of transmission line of length l, admittance Y_0 and propagation constant β, is introduced into a transmission line of characteristic admittance Y_1, and that this section is loaded at both ends with an admittance B (*fig. 9a*). Now imagine this section of line to be replaced (fig. 9*b*) by an unloaded section of line of the same length, but with the admittance Y_0' and propagation constant β'. The relation between Y_0 and β and the new parameters Y_0' and β' is given by the equations:

$$Y_0' = Y_0 \left\{1 - (B/Y_0)^2 + 2\,(B/Y_0)\cot\beta l\right\}^{\frac{1}{2}}, \tag{6}$$

$$\cos\beta' l = \cos\beta l - (B/Y_0)\sin\beta l. \tag{7}$$

The circuit that has been introduced will give no reflections provided it satisfies the conditions $Y_0' = Y_1$ and $\beta l = \pi/2$.

If we now substitute an admittance $-B$ for B, the phase shift of the circuit will consequently change by an amount $\Delta\phi$, given by:

$$\Delta\phi = \beta_1' l - \beta_2' l = \pi - 2\arccos(B/Y_0). \tag{8}$$

For a given design value of $\Delta\phi$ for the circuit, equations (6) and (8) yield two equations for Y_0 and B.

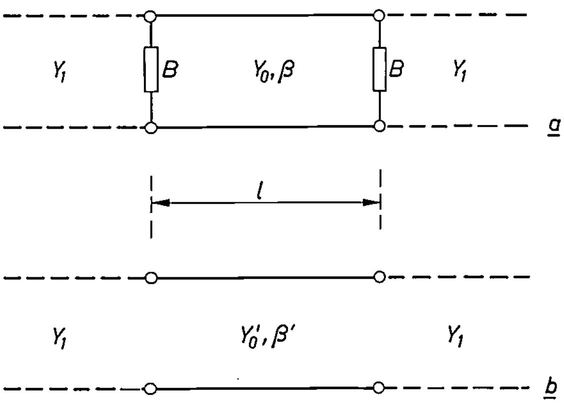


Fig. 9. *a*) Transmission line with characteristic admittance Y_1, in which a section of length l is included with admittance Y_0 and propagation constant β. The ends of this section are loaded with admittances B. *b*) Transmission line incorporating a section equivalent to that in (*a*), but which is not loaded at the ends. The length of both sections of line is the same, but the unloaded section has a characteristic admittance Y_0' and a propagation constant β'.

[8] J. F. White, IEEE Trans. **MTT-13**, 233, 1965, and J. F. White Proc. IEEE **56**, 1924, 1968.

As we have just noted, the requirement $\beta l = \pi/2$ must be met if the circuit is to give no reflections. Because of this the circuit cannot be used for producing large phase shifts. It is therefore used only for the 22.5° and 45° sections in a phase-shifter, and gives better results at these angles than the variable-reflection type of circuit described above.

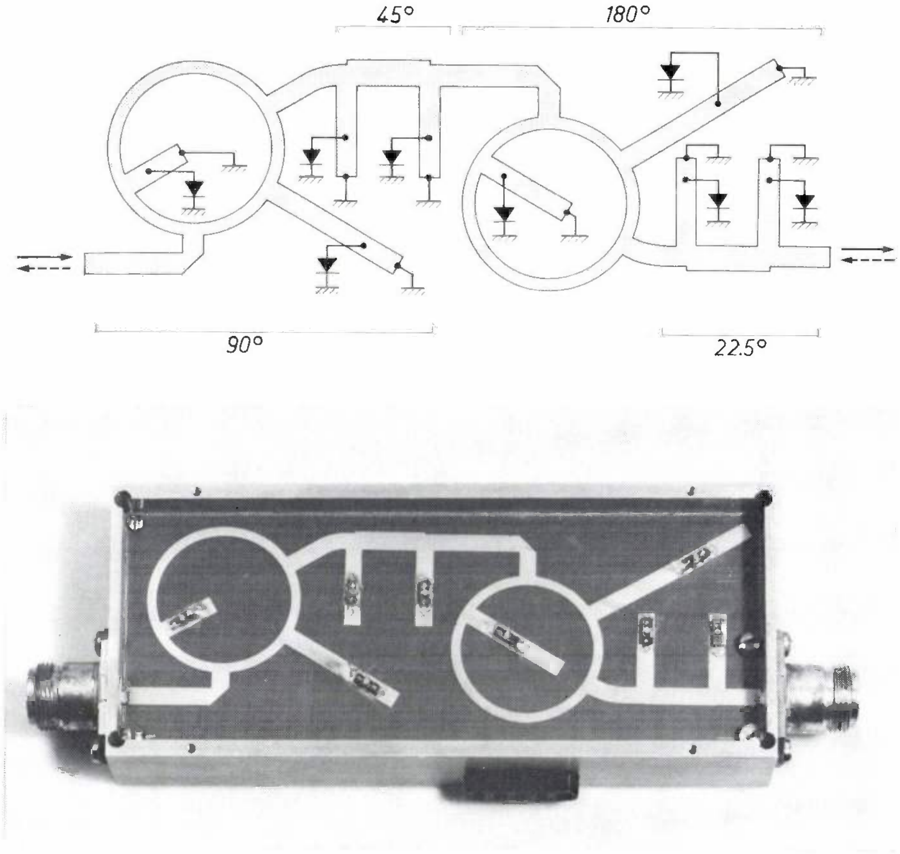


Fig. 10. The experimental phase-shifter for 3 GHz. *Above:* diagram of the phase-shifter. The locations of switching diodes and short-circuits are indicated. The extent to which the phase can be changed in various parts of the circuit is also indicated. *Below:* the phase-shifter. The characteristics of this phase-shifter are summarized in Table I.

An experimental phase-shifter for 3 GHz

The principles described above have been applied in the design of experimental phase-shifters for a frequency of 3 GHz. The entire circuit is in microstrip line (*fig. 10*) and is etched by a photochemical process from a metal layer applied to a PTFE substrate reinforced with glass fibre. The diodes are mounted in an encapsulation (*fig. 11*) whose stray capacitance and inductance are kept as low as possible. To prevent the diode control voltage from being short-circuited by the various short-circuits in the stripline, the diodes are connected to the ground plane of the circuit through capacitors. The complete phase-shifter consists of four sections, each with two diodes. The operation of the sections for 22.5° and 45° phase shift is based on the principle of the variable transmission line. The two other sections are hybrid rings with variable-reflection arms. These give phase shifts of 90° and 180°. In this way all phase shifts between 0° and 360° can be produced in steps of 22.5°. The main features of this digital phase-shifter are summarized in *Table I*.

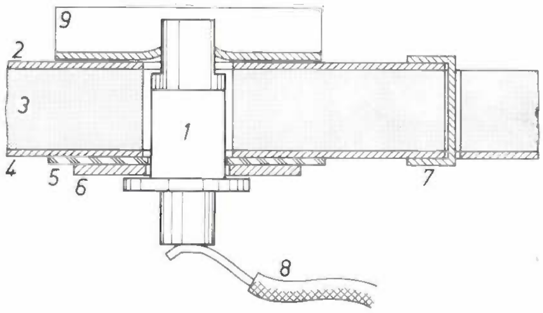


Fig. 11. A switching diode mounted in a microstrip circuit. *1* encapsulated diode. *2* microstrip line. *3* dielectric. *4* ground plane. The insulating sheet *5* and the metal plate *6* form a capacitor, whose function is to ensure that the short-circuited termination *7* of the microstrip does not short-circuit the diode control voltage. *8* supply lead for the control voltage. *9* mounting bracket.

Table I. Some data for the digital phase-shifter for microwave signals, shown in fig. 10.

Frequency	3.0 GHz
Smallest phase step	22.5°
Maximum phase shift	360°
Insertion loss (in all positions)	(1.3 ± 0.1) dB
Reflection (within a 150 MHz band)	< 10%
Transmitted peak power	2 kW
Transmitted average power	20 W
Operating current per diode (in the forward direction; low impedance)	50 mA
Operating voltage (in the reverse direction; high impedance)	12 V
Switching time	0.1 μs
Number of diodes	8
Information content of control signal	4 bits

The total insertion loss of the circuit in fig. 10 is 1.3 dB. The diodes account for 0.4 dB, and the rest is due to the microstrip lines in the circuit. Although the attenuation of the circuit could be reduced still further by using waveguides or other low-loss transmission lines, the device would then lose the advantages that make it so suitable for scanned arrays: it is small, and simple and inexpensive to manufacture.

Summary. A phase-shifter for electronic beam control in a radar system (phased array) should be compact and inexpensive, it should be capable of handling a sufficiently high peak power with low insertion loss and reflection, and should also be reciprocal and fast in operation. These requirements are explained with a short review of the operation of electronically scanned aerial arrays. It is shown that at microwave frequencies a *P-I-N* diode acts as a linear device whose impedance is determined by the d.c. operating point. A *P-I-N* diode can be used as a switch for controlling either the reflection or transmission characteristics of a line; the use of both possible methods for producing a phase-shifter is described. A figure of merit is defined for the switch and used in finding expressions that give the power-handling capabilities of a switch and the losses. Finally a 4-bit phase-shifter for 3 GHz with 8 diodes is described, which gives a maximum phase shift of 360° in steps of 22.5°. The main features of this phase-shifter are summarized in a table.

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 Forschungslaboratorium Aachen GmbH, Weißhausstraße, 51 Aachen, Germany	*A*
Philips Forschungslaboratorium Hamburg GmbH, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany	*H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium.	*B*

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

C. S. Aitchison: Microwave ICs. Looking at recent developments.
Electronics Weekly 26 Jan. 1972, p. 20. *M*

C. S. Aitchison & R. Davies: Gunn-effect power limiter.
IEEE Trans. **MTT-20**, 181-182, 1972 (No. 2). *M*

C. S. Aitchison, R. Davies, I. D. Higgins, S. R. Longley, B. H. Newton, J. F. Wells & J. C. Williams: Lumped-circuit elements at microwave frequencies.
IEEE Trans. **MTT-19**, 928-937, 1971 (No. 12). *M*

C. S. Aitchison, R. Davies, I. D. Higgins, S. R. Longley, B. H. Newton, J. F. Wells & J. C. Williams: Lumped microwave circuits, I. Technology and objectives, II. The five basic circuit elements, III. Filters and tunnel-diode amplifiers, IV. A lumped varactor-tuned Gunn oscillator, V. Degenerate parametric amplifier.
Design Electronics: **8**, No. 11, p. 23 *et seq.*, Sept. 1971; **9**, No. 1, p. 30 *et seq.*, Oct. 1971; **9**, No. 2, p. 42 *et seq.*, Nov. 1971; **9**, No. 3, p. 47 *et seq.*, Dec. 1971; **9**, No. 4, p. 41 *et seq.*, Jan. 1972. *M*

C. Albrecht & D. Dijkstra: Boundary condition for evaluating minority base charge in bipolar transistors.
Electronics Letters **7**, 613-615, 1971 (No. 20). *E*

H. G. Beljers & A. Broese van Groenou: Some effects of Cr^{3+} substitution in Ni and Mg ferrites.
Ferrites, Proc. int. Conf., Kyoto 1970, pp. 536-538; 1971. *E*

E. G. Berns & P. W. van der Zwaan (Philips Lighting Division, Eindhoven): The pyrohydrolytic determination of fluoride.
Anal. chim. Acta **59**, 293-297, 1972 (No. 2).

F. Berz & C. G. Prior: Test of McWhorter's model of low-frequency noise in Si MOSTs.
Microelectronics and Reliability **10**, 429-433, 1971 (No. 6). *M*

G. Bioul & M. Davio: Taylor expansions of Boolean functions and of their derivatives.
Philips Res. Repts. **27**, 1-6, 1972 (No. 1). *B*

J. Bloem (Philips Semiconductor Development Laboratory, Nijmegen): The effect of trace amounts of water vapor on boron doping in epitaxially grown silicon.
J. Electrochem. Soc. **118**, 1837-1841, 1971 (No. 11).

P. M. Boers: Measurements on the velocity/field characteristic of indium phosphide.
Electronics Letters **7**, 625-626, 1971 (No. 20). *E*

B. Bölger: A multiple-beam interferometer with transmission-like fringes in reflection.
Optics Comm. **4**, 313-315, 1971 (No. 4). *E*

P. T. Bolwijn, D. J. Schipper & C. Z. van Doorn: Cathodochromic properties of sodalite.
J. appl. Phys. **43**, 132-137, 1972 (No. 1). *E*

J. van den Boomgaard & F. M. A. Carpay: The eutectoid Co_3Si phase in the Co-Si system.
Acta metall. **20**, 473-476, 1972 (No. 4). *E*

G. A. Bootsma & H. J. Gassen: Room temperature gas adsorption on silicon carbide.
Surface Sci. **29**, 297-299, 1972 (No. 1). *E*

G. A. Bootsma, W. F. Knippenberg & G. Verspui: Growth of SiC whiskers in the system SiO_2-C-H_2 nucleated by iron.
J. Crystal Growth **11**, 297-309, 1971 (No. 3). *E*

H. Bosma: Stochastiek in de elektrotechniek, IX. Het begrip ruismaat en de "temperatuur" van lineaire transmissie-netwerken.
Ingenieur 83, E 23-32, 1971 (No. 9). *E*

H. Bosma: Reflections on non-reciprocal microwave ferrite devices.
Proc. 1971 European Microwave Conf., Stockholm, Vol. 2, pp. B 13/S:1-10. *E*

H. Bouma & L. C. J. Baghuis (Institute for Perception Research, Eindhoven): Hippus of the pupil: periods of slow oscillations of unknown origin.
Vision Res. 11, 1345-1351, 1971 (No. 11).

P. Branquart & J. Lewi: A scheme of storage allocation and garbage collection for ALGOL 68.
ALGOL 68 Implementation, Proc. IFIP Working Conf., Munich 1970, pp. 199-238; 1971. *B*

P. Branquart, J. Lewi & J. P. Cardinael: Analysis of the parenthesis structure of ALGOL 68.
ALGOL 68 Implementation, Proc. IFIP Working Conf., Munich 1970, pp. 37-76; 1971. *B*

J. W. Broer: Abstracts in block diagram form.
IEEE Trans. **EWS-14**, 64-67, 1971 (No. 3). *E*

J. Burmeister: The surface morphology of epitaxial silicon.
J. Crystal Growth **11**, 131-140, 1971 (No. 2). *A*

K. H. J. Buschow: Intermetallic compounds of rare earth elements and Ni, Co, or Fe.
Phys. Stat. sol. (a) **7**, 199-210, 1971 (No. 1). *E*

K. H. J. Buschow: Note on the structure and occurrence of ytterbium transition metal compounds.
J. less-common Met. **26**, 329-333, 1972 (No. 3). *E*

K. H. J. Buschow, A. M. van Diepen & H. W. de Wijn (University of Utrecht): Magnetic properties of $Gd_xTh_{1-x}CuAl$ and nuclear magnetic resonance in GdCuAl.
J. appl. Phys. **42**, 4315-4318, 1971 (No. 11). *E*

K. H. J. Buschow, A. Oppelt (II. Phys. Inst., Technische Hochschule Darmstadt) **& E. Dormann** (II. Phys. Inst., T. H. Darmstadt): Influence of the mean free path of the conduction electrons on the magnetic properties of Gd intermetallics with CsCl structure.
Phys. Stat. sol. (b) **50**, 647-652, 1972 (No. 2). *E*

F. M. A. Carpay & J. van den Boomgaard: The relationship between interlamellar period and growth rate in the unidirectionally decomposed eutectoids Co_3Si and β Ni-In.
Acta metall. **19**, 1279-1286, 1971 (No. 11). *E*

H. B. G. Casimir: Industries and academic freedom.
Research Policy **1**, 3-8, 1971/72 (No. 1). *E*

L. W. Chua: New broadband matched hybrids for microwave integrated circuits.
Proc. 1971 European Microwave Conf., Stockholm, Vol. 2, pp. C 4/5:1-4. *M*

T. D. Clark: Generation and detection experiments using point contact junction arrays.
Physica **55**, 432-438, 1971. *M, E*

P. J. Courtois, F. Heymans & D. L. Parnas: Concurrent control with "readers" and "writers".
Comm. ACM **14**, 667-668, 1971 (No. 10). *B*

D. J. Craik, P. V. Cooper (both with University of Nottingham, U.K.) **& W. F. Druyvesteyn:** Magnetostatic energy coefficients for cylindrical domains.
Physics Letters **34A**, 244, 1971 (No. 4). *E*

H. J. van Daal: The nature of free and bound charge carriers in some transition-metal oxides.
Conduction in low-mobility materials, Proc. 2nd int. Conf., Eilat 1971, pp. 19-30. *E*

H. J. van Daal & K. H. J. Buschow: Kondo effect in intermetallische verbindingen van cerium.
Ned. T. Natuurk. **38**, 69-77, 1972 (No. 5). *E*

M. Davio & J. J. Quisquater: Affine cascades.
Philips Res. Repts. **27**, 109-125, 1972 (No. 2). *B*

J. P. Deschamps: Synchronization of finite automata.
Philips Res. Repts. **27**, 126-139, 1972 (No. 2). *B*

F. Desvignes: Limites fondamentales et critères de qualité des capteurs de rayonnement.
Nouv. Rev. Optique appl. **2**, 121-132, 1971 (No. 2). *L*

G. Dittmer: Electron conduction, electron emission and electroluminescence of MIM sandwich structures with Al_2O_3 insulating layers.
Thin Solid Films **9**, 141-172, 1972 (No. 2). *A*

G. Dittmer: Electrical conduction and electron emission of discontinuous thin films.
Thin Solid Films **9**, 317-328, 1972 (No. 3). *A*

C. Z. van Doorn, D. J. Schipper & P. T. Bolwijn: Optical investigation of cathodochromic sodalite.
J. Electrochem. Soc. **119**, 85-92, 1972 (No. 1). *E*

J. W. F. Dorleijn, W. F. Druyvesteyn, F. A. de Jonge & H. M. W. Booij: Repulsive interactions between magnetic bubbles: consequences for bubble devices.
IEEE Trans. **MAG-7**, 355-358, 1971 (No. 3). *E*

A. Douglas: Examples concerning efficient strategies for Gaussian elimination.
Computing **8**, 382-394, 1971 (No. 3/4). *E*

P. C. Drop & J. Polman: Calculations on the effect of supply frequency on the positive column of a low-pressure Hg-Ar AC discharge.
J. Physics D **5**, 562-568, 1972 (No. 3). *E*

W. F. Druyvesteyn, R. Szymczak (Institute of Physics, Warsaw) **& R. Wadas** (Inst. Phys., Warsaw): Calculation on the behaviour of a cylindrical magnetic domain in a finite plate.
Phys. Stat. sol. (a) **9**, 343-348, 1972 (No. 1). *E*

W. F. Druyvesteyn, D. L. A. Tjaden & J. W. F. Dorleijn: Calculation of the stray field of a magnetic bubble, with application to some bubble problems.
Philips Res. Repts. **27**, 7-27, 1972 (No. 1). *E*

D. den Engelsen: Ellipsometry of anisotropic films.
J. Opt. Soc. Amer. **61**, 1460-1466, 1971 (No. 11). *E*

E. Fabre: Scanning photoluminescence on gallium-arsenide.
Solid State Comm. **9**, 635-638, 1971 (No. 10). *L*

J. P. Fekete: Image recovery microwave mixer.
Proc. 1971 European Microwave Conf., Stockholm, Vol. 1, pp. A 12/1:1-4. *M*

G. B. Gerritsen & G. E. G. Hardeman: An infrared image converter equipped with an array of extrinsic silicon photodetectors.
IEEE Trans. **ED-18**, 1011-1015, 1971 (No. 11). *E*

C. J. Gerritsma, W. H. de Jeu & P. van Zanten: Distortion of a twisted nematic liquid crystal by a magnetic field.
Physics Letters **36A**, 389-390, 1971 (No. 5). *E*

C. J. Gerritsma & P. van Zanten: Periodic perturbations in the cholesteric plane texture.
Physics Letters **37A**, 47-48, 1971 (No. 1). *E*

J. A. Geurst: Continuum theory of type-A smectic liquid crystals.
Physics Letters **37A**, 279-280, 1971 (No. 4). *E*

J.-M. Goethals: On the Golay perfect binary code.
J. combin. Theory A **11**, 178-186, 1971 (No. 2). *B*

R. G. Gossink & J. M. Stevels: Über den Dissoziationszustand von geschmolzenem Natriumdiwolframat ($Na_2W_2O_7$).
Z. anorg. allgem. Chemie **388**, 282-290, 1972 (No. 3).*E*

H. C. de Graaff & J. te Winkel: Relationship between crossmodulation and intermodulation.
Electronics Letters **8**, 33-34, 1972 (No. 2). *E*

C. G. Gray (University of Guelph, Canada) **& H. I. Ralph:** Solution of Boltzmann's equation for semiconductors using a spherical harmonic expansion.
J. Physics C **5**, 55-62, 1972 (No. 1). *M*

G. Groh: Holographic tomography using a circular synthetic aperture.
Appl. Optics **10**, 2549-2550, 1971 (No. 11). *H*

W. S. C. Gurney: Contact effects in Gunn diodes.
Electronics Letters **7**, 711-713, 1971 (No. 24). *M*

G. J. van Gurp: Electromigration in Al films containing Si.
Appl. Phys. Letters **19**, 476-478, 1971 (No. 11). *E*

R. F. Hall, K. E. Johnson & G. T. Sharpless: A tabular data display using a cross-bar addressed glow discharge panel.
IEE Conf. Publn. No. 80: Displays, pp. 91-96, 1971. *M*

P. Hansen: Ferromagnetic resonance in ruthenium-doped yttrium-iron garnet.
Phys. Stat. sol. (b) **47**, 565-572, 1971 (No. 2). *H*

P. Hansen: Spin configurations induced by strong anisotropic ions in yttrium-iron-garnet.
J. appl. Phys. **43**, 650-655, 1972 (No. 2). *H*

P. Hansen & W. Tolksdorf: Spin-wave linewidth of ruthenium and iridium-substituted yttrium-iron-garnet.
Phys. Stat. sol. (a) **6**, K 11-13, 1971 (No. 1). *H*

K. H. Härdtl & R. Wernicke: Lowering the Curie temperature in reduced $BaTiO_3$.
Solid State Comm. **10**, 153-157, 1972 (No. 1). *A*

P. A. H. Hart: Device modelling.
Solid state devices, Proc. ESDERC Conf., Munich 1971, pp. 31-43. *E*

J. Hasker: Astigmatic electron gun for the beam-indexing color television display.
IEEE Trans. **ED-18**, 703-712, 1971 (No. 9). *E*

J. Hasker: Imaging, beam acceptance, and beam-discharge lag in camera tubes.
IEEE Trans. **ED-18**, 1075-1086, 1971 (No. 11). *E*

H. Haug & K. Weiss: Connection between a microscopic and the phenomenological theory of superfluidity.
Lettere Nuovo Cim. (2) **2**, 887-890, 1971 (No. 17). *E*

E. E. Havinga (I, II), **H. Damsma** (I) **& P. Hokkeling** (I): Compounds and pseudo-binary alloys with the $CuAl_2$($C16$)-type structure, I. Preparation and X-ray results, II. Theoretical discussion of crystallographic parameters.
J. less-common Met. **27**, 169-186 & 187-193, 1972 (No. 2). *E*

N. Hazewindus: On the acceptance of a cyclotron axial injection system.
Nucl. Instr. Meth. **96**, 227-233, 1971 (No. 2). *E*

L. Heyne & N. M. Beekmans: Electronic transport in calcia-stabilized zirconia.
Proc. Brit. Ceramic Soc. No. 19, 229-263, 1971. *E*

L. Heyne, N. M. Beekmans & A. de Beer: Ionic conduction and oxygen diffusion in yellow lead oxide.
J. Electrochem. Soc. **119**, 77-84, 1972 (No. 1). *E*

B. Hill: Some aspects of a large capacity holographic memory.
Appl. Optics **11**, 182-191, 1972 (No. 1). *H*

E. E. E. Hoefer: Automatische Auswertung von Röntgenfernsehbildern bei der Prüfung nuklearer Brennelemente.
Materialprüfung **14**, 43-48, 1972 (No. 2). *H*

E. P. Honig & J. H. Th. Hengst: Permeation through $BaSO_4$-precipitate membranes.
Electrochim. Acta **17**, 75-90, 1972 (No. 1). *E*

J. Hornstra & K. H. J. Buschow: The crystal structure of $YbCu_{6.5}$.
J. less-common Met. **27**, 123-127, 1972 (No. 2). *E*

J. Hornstra & E. Keulen (Philips Industrial Equipment Division, Eindhoven): The oxygen parameter of the spinel $ZnGa_2O_4$.
Philips Res. Repts. **27**, 76-81, 1972 (No. 1). *E*

F. D. Hughes, R. F. Headon & M. Wilson: On the measurement of semiconductor carrier concentration profiles.
J. Physics E **5**, 241-242, 1972 (No. 3). *M*

J. P. Hurault: Effet tunnel assisté par des niveaux liés.
J. Physique **32**, 421-426, 1971 (No. 5/6). *L*

B. B. van Iperen & H. Tjassens: Novel and accurate methods for measuring small-signal and large-signal impedances of Impatt diodes.
Philips Res. Repts. **27**, 38-75, 1972 (No. 1). *E*

W. H. de Jeu: Instabilities of nematic liquid crystals in pulsating electric fields.
Physics Letters **37A**, 365-366, 1971 (No. 5). *E*

W. H. de Jeu & J. van der Veen: On the relation between molecular structure and liquid-crystalline behaviour.
Philips Res. Repts. **27**, 172-185, 1972 (No. 2). *E*

F. A. de Jonge, W. F. Druyvesteyn & A. G. H. Verhulst: Multiple hysteresis of a hollow cylindrical domain.
Int. J. Magnetism **2**, No. 2, 7-13, 1972. *E*

H. Jonker, L. K. H. van Beek, H. J. Houtman, F. T. Klostermann & E. J. Spiertz: PD processes for photography at extreme resolution, I. PD-MD based on diazosulphonate, II. PD-D based on diazosulphide.
J. photogr. Sci. **19**, 187-198, 1971 (No. 6), & **20**, 53-63, 1972 (No. 2). *E*

B. A. Joyce & J. H. Neave: An investigation of silicon-oxygen interactions using Auger electron spectroscopy.
Surface Sci. **27**, 499-515, 1971 (No. 3). *M*

Y. Kamp: Cascade synthesis of multivariable impedances of arbitrary degree.
Arch. Elektronik & Übertr.technik (AEÜ) **25**, 362-368, 1971 (No. 8). *B*

A. F. V. van Katwijk (Institute for Perception Research, Eindhoven): Een goed verstaander heeft maar een half woord nodig: de vraag is welke helft?
Langage et l'Homme **1**, 15-19, 1972.

A. W. C. van Kemenade & C. F. Stemfoort: On the formation of β-SiC from pyrolysis of CH_3SiCl_3 in hydrogen.
J. Crystal Growth **12**, 13-16, 1972 (No. 1). *E*

E. T. Keve, K. L. Bye, P. W. Whipps & A. D. Annis: Structural inhibition of ferroelectric switching in triglycine sulphate, I. Additives.
Ferroelectrics **3**, 39-48, 1971 (No. 1). *M*

M. Klinck: Modell für einen fremderregten Gleichstrom-Nebenschlußmotor mit Freilaufdiode.
Elektrotechn. Z. A **93**, 78-81, 1972 (No. 2). *H*

J. T. Klomp: Solid-state bonding of metals to ceramics.
Science of Ceramics **5**, 501-522, 1970. *E*

J. T. Klomp: Recent developments for bonding ceramics to metals.
Powder Metall. Int. **3**, 142-146 & 193-194, 1971 (Nos. 3 & 4). *E*

A. J. R. de Kock: The elimination of vacancy-cluster formation in dislocation-free silicon crystals.
J. Electrochem. Soc. **118**, 1851-1856, 1971 (No. 11). *E*

A. J. R. de Kock, F. M. Beeftink (Philips Electronic Components and Materials Division (Elcoma), Eindhoven) **& K. J. Schell** (Philips Elcoma Division, Eindhoven): Investigation of lithium precipitation in germanium crystals by X-ray transmission topography.
Appl. Phys. Letters **20**, 81-83, 1972 (No. 2). *E*

W. L. Konijnendijk & H. Groenendijk: Nat-chemische bereiding van homogene glasachtige systemen.
Klei en Keramiek **22**, 7-13, 1972 (No. 1). *E*

F. A. Kuijpers: Investigations on the $LaCo_5$-H and $CeCo_5$-H systems.
J. less-common Met. **27**, 27-34, 1972 (No. 1). *E*

F. A. Kuijpers & B. O. Loopstra (R. C. N., Petten, Netherlands): Magnetic structure of $PrCo_5D_4$.
J. Physique **32**, C1/657-658, 1971 (Colloque No. 1, Vol. II). *E*

D. E. Lacklison, H. I. Ralph & G. B. Scott: The Faraday rotation of bismuth calcium vanadium iron garnet.
Solid State Comm. **10**, 269-272, 1972 (No. 3). *M*

H. 't Lam & G. A. Acket: Comparison of the microwave velocity/field characteristics of *n* type InP and *n* type GaAs.
Electronics Letters **7**, 722-723, 1971 (No. 24). *E*

A. J. Lambell: A broadband electronically-scanned linear waveguide array.
Proc. 1971 European Microwave Conf., Stockholm, Vol. 1, pp. **B** 3/6:1-4. *M*

H. de Lang: Coherent diffraction of electrons by standing light waves.
Optics Comm. **4**, 191-194, 1971 (No. 3). *E*

F. K. Lotgering & G. H. A. M. van der Steen: Metal-deficient sulphospinels.
J. Solid State Chem. **3**, 574-581, 1971 (No. 4). *E*

M. H. van Maaren, K. H. J. Buschow & H. J. van Daal: Low-temperature specific heat of $CeAl_3$ and related compounds.
Solid State Comm. **9**, 1981-1984, 1971 (No. 22). *E*

R. Memming & G. Kürsten: Photo- and electrochemical reactions of excited dyes at metal electrodes.
Ber. Bunsen-Ges. phys. Chemie **76**, 4-11, 1972 (No. 1). *H*

R. Metselaar: Radio frequency sputtering of ferrites.
Science of Ceramics **5**, 363-376, 1970. *E*

F. Meyer & J. M. Morabito: Adsorption of organic gases on clean germanium surfaces.
J. phys. Chem. **75**, 2922-2929, 1971 (No. 19). *E*

A. P. J. Michels: The NO-content in the exhaust gases of a Stirling engine.
Proc. 1971 Intersociety Energy Conversion Engng. Conf., Boston, pp. 1010-1023. *E*

I. C. P. Millar, D. L. Lamport & A. W. Woodhead: An experimental X-ray image intensifier incorporating a channel electron multiplier plate.
IEEE Trans. **ED-18**, 1101-1108, 1971 (No. 11). *M*

A. Mircea: A simple criterion for LSA oscillation.
IEEE Trans. **ED-18**, 449, 1971 (No. 7). *L*

A. Mircea, B. Kramer & A. Farrayre: Sources à état solide à grand rendement en hyperfréquence.
Onde électr. **51**, 499-501, 1971 (No. 6). *L*

B. J. Mulder: Recombination diffusion length of minority charge carriers in cuprous sulphide.
Solar Cells, Proc. int. Coll., Toulouse 1970, pp. 131-140; 1971. *E*

J. H. Neave, C. T. Foxon & B. A. Joyce: The dependence of Auger electron yield on primary beam energy at normal and glancing incidence.
Surface Sci. **29**, 411-423, 1972 (No. 2). *M*

J. Neirynck: Maximally flat approximations to the ideal filter as hypergeometric functions.
Electronics Letters **7**, 591-593, 1971 (No. 19). *B*

J. M. van Nieuwland: Extraction of particles from a compact isochronous cyclotron.
Thesis, Eindhoven 1972. *E*

J. M. Noothoven van Goor: Charge carrier densities and mobilities in doped and pure bismuth.
The physics of semimetals and narrow-gap semiconductors, Proc. Conf. Dallas, Texas, 1970, pp. 63-72; 1971. *E*

J. M. Noothoven van Goor: Epilogue to the conference.
The physics of semimetals and narrow-gap semiconductors, Proc. Conf. Dallas, Texas, 1970, p. 563; 1971. *E*

W. J. Oosterkamp, A. P. M. van 't Hof & J. C. Driessen: Laufende Subtraktionen von Bewegungserscheinungen mit Hilfe eines Bildbandspeichers.
Angiographie und ihre Fortschritte, Tagung Baden-Baden 1970, pp. 50-51; 1972. *E*

W. J. Oosterkamp & J. J. F. de Wijk: Exposure meters for protection against X-rays between 10 and 100 kV — method of calibration and characteristics of various types of instruments.
Premier Congrès Européen de l'Ass. Int. de Radioprotection sur la protection contre les rayonnements de faible énergie ou de faible parcours et les effets biologiques de l'irradiation, Menton 1968, pp. 27-30; 1971. *E*

K. J. van Oostrum: Design parameters for non-immersion lenses.
Jernkont. Ann. **155**, 491-492, 1971 (No. 8). *E*

C. van Opdorp: The present state of the theory of photoeffects in heterojunctions.
Proc. Int. Conf. on the physics and chemistry of semiconductor heterojunctions and layer structures, Budapest 1970, Part II, pp. 91-122; 1971. *E*

A. Oppelt, E. Dormann (both with II. Phys. Inst., Technische Hochschule Darmstadt) **& K. H. J. Buschow:** NMR study of ferromagnetic Gd intermetallic compounds with CsCl structure.
Phys. Stat. sol. (b) **51**, 275-282, 1972 (No. 1). *E*

G. den Ouden: Stikstof in lasmetaal.
Rev. Soudure **27**, 85-93, 1971 (No. 2). *E*

R. F. Pearson, A. D. Annis & J. L. Page: Photomagnetic effects in iron garnets.
Ferrites, Proc. int. Conf., Kyoto 1970, pp. 8-13; 1971. *M*

Th. H. Peek: Generalization of Françon's modification of the Savart plate.
Appl. Optics **10**, 2235-2239, 1971 (No. 10). *E*

R. J. van de Plassche: A wide-band operational amplifier with a new output stage and a simple frequency compensation.
IEEE J. **SC-6**, 347-352, 1971 (No. 6). *E*

D. Polder & M. van Hove: Theory of radiative heat transfer between closely spaced bodies.
Phys. Rev. B **4**, 3303-3314, 1971 (No. 10). *E*

J. Polman, J. E. van der Werf & P. C. Drop: Nonlinear effects in the positive column of a strongly modulated mercury - rare gas discharge.
J. Physics D **5**, 266-279, 1972 (No. 2). *E*

H. Rau: Thermodynamics of the reduction of iron oxide powders with hydrogen.
J. chem. Thermodyn. **4**, 57-64, 1972 (No. 1). *A*

J. P. Reinhoudt: On the stability of rotor-and-bearing systems and on the calculation of sliding bearings.
Thesis, Eindhoven 1972. *E*

W. Rey, J. D. Laird ("Dijkzigt" University Hospital, Rotterdam) **& P. G. Hugenholtz** ("Dijkzigt", Rotterdam): *P*-wave detection by digital computer.
Computers & biomed. Res. **4**, 509-522, 1971 (No. 5). *B*

G. Rinzema: De meting van het magneto-optische Faraday- en Kerr-effect.
Polytechn. T. Elektr. **26**, 742-748, 1971 (No. 19). *E*

E. Roeder: Flow behaviour of glass during extrusion.
J. non-cryst. Solids **7**, 203-220, 1972 (No. 2). *A*

J. Roos: Electrostatic sound recording, an experiment.
Proc. 3rd Conf. on Static Electrification, London 1971, pp. 60-67. *E*

J. H. T. van Roosmalen: New possibilities for the design of "Plumbicon" tubes.
IEEE Trans. **ED-18**, 1087-1093, 1971 (No. 11). *E*

J. A. J. Roufs (Institute for Perception Research, Eindhoven): Dynamic properties of vision, I. Experimental relationships between flicker and flash thresholds, II. Theoretical relationships between flicker and flash thresholds.
Vision Res. **12**, 261-278 & 279-292, 1972 (No. 2).

P. Saraga & P. R. Wavish: Edge-coding operators for pattern recognition.
Electronics Letters **7**, 736-738, 1971 (No. 25). *M*

B. Schiek & K. Schünemann (Valvo, Hamburg): Detuning effects and noise in cavity-stabilised Gunn oscillators.
Electronics Letters **8**, 52-53, 1972 (No. 3). *H*

O. Schob (Philips Lighting Division, Eindhoven): Zur Kristallographie der Gleitverformung von Wolfram-Einkristallen im Zugversuch.
Monatsh. Chemie **103**, 255-269, 1972 (No. 1).

K. Schünemann (Valvo, Hamburg) **& B. Schiek:** Comparison between transmission- and reaction-cavity-stabilised oscillators.
Electronics Letters **7**, 618-620, 1971 (No. 20). *H*

K. Schünemann & B. Schiek: Influence of a transmission line on the noise spectra of cavity-stabilised oscillators.
Electronics Letters **7**, 659-661, 1971 (No. 22). *H*

M. F. H. Schuurmans, W. van Haeringen & H.-G. Junginger: On the fitting of electronic energy bands to symmetrized plane waves.
Solid State Comm. **10**, 549-552, 1972 (No. 6). *E, A*

M. H. Seavey: Acoustic resonance in the easy-plane weak ferromagnets α-Fe$_2$O$_3$ and FeBO$_3$.
Solid State Comm. **10**, 219-223, 1972 (No. 2). *E*

R. J. Sluyter & P. J. van Gerwen: Dual single-sideband modulation for wideband data transmission.
Electronics Letters **7**, 640-642, 1971 (No. 21). *E*

L. A. Æ. Sluyterman & M. J. M. de Graaf: The effect of salts upon the pH dependence of the activity of papain and succinyl-papain.
Biochim. biophys. Acta **258**, 554-561, 1972 (No. 2). *E*

L. A. Æ. Sluyterman & J. Wijdenes: Cyanuration of papain. Activity and fluorescence of the products.
Biochim. biophys. Acta **263**, 329-338, 1972 (No. 2). *E*

E. C. Snelling: Ferrites for linear applications, I. Properties, II. Performance requirements.
IEEE Spectrum **9**, No. 1, 42-51, Jan. 1972, & **9**, No. 2, 26-32, Feb. 1972. *M*

J. L. Sommerdijk: On the excitation mechanisms of the infrared-excited visible luminescence in Yb^{3+},Er^{3+}-doped fluorides.
J. Luminescence **4**, 441-449, 1971 (No. 4). *E*

J. L. Sommerdijk, W. L. Wanmaker & J. G. Verriet: Infrared-excited visible luminescence in oxidic lattices doped with Yb^{3+} and Er^{3+}.
J. Luminescence **4**, 404-416, 1971 (No. 4). *E*

W. T. Stacy & C. J. M. Rooymans: A crystal field mechanism for the noncubic magnetic anisotropy in iron garnet: oxygen vacancy ordering.
Solid State Comm. **9**, 2005-2008, 1971 (No. 23). *E*

J. M. Stevels: Repeatability number, Deborah number and critical cooling rates as characteristic parameters of the vitreous state.
J. non-cryst. Solids **6**, 307-321, 1971 (No. 4). *E*

T. G. W. Stijntjes (Philips Electronic Components and Materials Division (Elcoma), Eindhoven), **A. Broese van Groenou, R. F. Pearson, J. E. Knowles & P. Rankin:** Effects of various substitutions in Mn-Zn-Fe ferrites.
Ferrites, Proc. int. Conf., Kyoto 1970, pp. 194-198; 1971. *E, M*

T. G. W. Stijntjes, J. Klerk (both with Philips Electronic Components and Materials Division (Elcoma), Eindhoven), **C. J. M. Rooymans, A. Broese van Groenou, R. F. Pearson, J. E. Knowles & P. Rankin:** Magnetic properties and conductivity of Ti-substituted Mn-Zn ferrites.
Ferrites, Proc. int. Conf., Kyoto 1970, pp. 191-193; 1971. *E, M*

A. J. A. van Stratum & P. N. Kuin: Tracer study on the decrease of emission density of osmium-coated impregnated cathodes.
J. appl. Phys. **42**, 4436-4437, 1971 (No. 11). *E*

S. Strijbos: The manufacture of carbon-film resistors employing fluidized-bed technology.
Philips Res. Repts. **27**, 186-195, 1972 (No. 2). *E*

A. L. Stuijts: New fabrication methods for advanced electronic materials.
Science of Ceramics **5**, 335-362, 1970. *E*

A. L. Stuijts: Control of microstructures in ferrites.
Ferrites, Proc. int. Conf., Kyoto 1970, pp. 108-113; 1971. *E*

A. L. Stuijts, D. Veeneman & A. Broese van Groenou: Preparation of ferrous-zinc ferrites with high saturation magnetization.
Ferrites, Proc. int. Conf., Kyoto 1970, pp. 236-238; 1971. *E*

J. van Suchtelen: Product properties: a new application of composite materials.
Philips Res. Repts. **27**, 28-37, 1972 (No. 1). *E*

T. J. B. Swanenburg: Observation of negative conductance in an interdigital electrode structure on an oxidized Si-surface.
Physics Letters **38A**, 311-312, 1972 (No. 5). *E*

T. L. Tansley: Forward bias conduction of anisotype heterojunctions.
Proc. Int. Conf. on the physics and chemistry of semiconductor heterojunctions and layer structures, Budapest 1970, Part II, pp. 123-150; 1971. *M*

T. L. Tansley: Heterojunction properties.
Semiconductors and Semimetals, editors R. K. Willardson & A. C. Beer, publ. Academic Press, New York, Vol. 7A, pp. 293-368, 1971. *M*

A. Thayse: A variational diagnosis method for stuck-faults in combinatorial networks.
Philips Res. Repts. **27**, 82-98, 1972 (No. 1). *B*

A. Thayse: Testing of asynchronous sequential switching circuits.
Philips Res. Repts. **27**, 99-106, 1972 (No. 1). *B*

A. Thayse: A fast algorithm for the proper decomposition of Boolean functions.
Philips Res. Repts. **27**, 140-150, 1972 (No. 2). *B*

J. B. Theeten, J. Bonnerot (La Radiotechnique, Suresnes, France), **J. L. Domange** (Ecole Nationale Supérieure de Chimie de Paris) **& J. P. Hurault:** Very low temperature LEED from Si(III).
Solid State Comm. **9**, 1121-1123, 1971 (No. 13). *L*

D. L. A. Tjaden: Dependence of digital recording properties upon physical parameters for metallic recording surfaces.
IEEE Trans. **MAG-7**, 544-545, 1971 (No. 3). *E*

D. L. A. Tjaden & A. M. A. Rijckaert: Theory of anhysteretic contact duplication.
IEEE Trans. **MAG-7**, 532-536, 1971 (No. 3). *E*

H. van Tongeren: Electron temperature and radial density distribution of Cs ground-state atoms in the positive column of a Cs-Ar d.c. low pressure discharge.
Physics Letters **37A**, 317-318, 1971 (No. 4). *E*

J. D. B. Veldkamp: On the tensile strength of ribbon-like SiC whiskers.
J. Mat. Sci. **6**, 1486-1492, 1971 (No. 12). *E*

J. D. B. Veldkamp & W. F. Knippenberg: Nieuwe gewapende materialen.
Klei en Keramiek **22**, 62-79, 1972 (No. 3). *E*

C. H. F. Velzel: Inverse Fourier spectroscopy.
Optics Comm. **4**, 121-124, 1971 (No. 2). *E*

J. G. Verhagen, A. Liefkens & G. W. Tichelaar: Gas shielding for CO_2 welding.
Metal Constr. Brit. Welding J. **4**, 47-50, 1972 (No. 2). *E*

A. G. H. Verhulst, T. Holtwijk, W. Lems & U. Enz: Pulse induced suppression of flux switching in photomagnetic garnets.
IEEE Trans. **MAG-7**, 729-732, 1971 (No. 3). *E*

J. C. Verplanke: Full-energy peak efficiency calibration of a well-type Ge(Li) gamma-ray detector.
Nucl. Instr. Meth. **96**, 557-560, 1971 (No. 4). *E*

G. Verspui, W. F. Knippenberg & G. A. Bootsma: Lanthanum-stimulated high-temperature whisker growth of α-SiC.
J. Crystal Growth **12**, 97-105, 1972 (No. 2). *E*

J. F. Verwey & B. J. de Maagt: Emitter avalanche currents in gated transistors.
IEEE Trans. **ED-19**, 245-250, 1972 (No. 2). *E*

A. G. van Vijfeijken: Technological forecasting: techniques.
Long range forecasting, some aspects and tools, editor F. C. Romeijn, pp. 26-45, 1971. *E*

A. T. Vink, A. J. Bosman, J. A. W. van der Does de Bye & R. C. Peters: Optical properties of excitons bound to neutral Si_{Ga}-donors in GaP and the degeneracy of the Si_{Ga}-donor ground state.
J. Luminescence **5**, 57-68, 1972 (No. 1). *E*

M. T. Vlaardingerbroek: Theory of oscillator noise.
Electronics Letters **7**, 648-650, 1971 (No. 21). *E*

J. Vlietstra: Computers en ontwikkelingslanden, I, II, III.
Informatie **14**, 177-181, 303-308 & 368-374, 1972 (Nos. 4, 6 & 7/8). *E*

H. J. W. M. Volman: Berekening en constructie van het radiaal hydrostatisch staplager.
Polytechn. T. Werktuigbouw **27**, 46-52, 1972 (No. 2). *E*

L. Vriens: Scattering of polarized electrons from magnetic materials.
Phys. Rev. B **4**, 3088-3093, 1971 (No. 9). *E*

J. W. ter Vrugt, W. L. Wanmaker & J. G. Verriet (Philips Lighting Division, Eindhoven): $Ba_3Y_2MoO_9$, a new molybdate with perovskite structure.
J. inorg. nucl. Chem. **34**, 762-763, 1972 (No. 2).

K. Walther: Ultrasonic attenuation and Mn^{55} nuclear acoustic resonance in MnTe.
Phys. Rev. B **4**, 3873-3885, 1971 (No. 11). *H*

K. Walther: Magnetic-field dependence of the ^{55}Mn-nuclear acoustic resonance in MnTe.
Physics Letters **38A**, 149-150, 1972 (No. 3). *H*

J. H. Waszink & M. Kauffmann: Experimental investigation of the positive column of a GaI_3-Ar discharge.
J. appl. Phys. **42**, 4848-4854, 1971 (No. 12). *E*

J. H. Waszink & J. A. J. M. van Vliet: Measurements of the gas temperature in CO_2-N_2-He and CO_2-N_2-H_2O-He discharges.
J. appl. Phys. **42**, 3374-3379, 1971 (No. 9). *E*

H. Weiss: Autocorrelation of the wave front in the focus of a homogeneous sphere.
Optik **34**, 475-481, 1972 (No. 4). *H*

K. Weiss: Zur diffusionsbedingten Sinterung eines binären Oxids.
Ber. Dtsch. Keram. Ges. **48**, 489-493, 1971 (No. 11). *E*

F. F. Westendorp: Domain-wall energy and coercive force of cobalt rare-earth permanent magnet materials.
J. appl. Phys. **42**, 5727-5731, 1971 (No. 13). *E*

M. V. Whelan: Recombination-generation currents at a thermally oxidized silicon surface.
Proc. Int. Conf. on the physics and chemistry of semiconductor heterojunctions and layer structures, Budapest 1970, Part V, pp. 221-229; 1971. *E*

P. W. Whipps: Growth of high-quality crystals of KTN.
J. Crystal Growth **12**, 120-124, 1972 (No. 2). *M*

P. W. Whipps: Stability regions for the growth of barium-strontium niobate crystals.
J. Solid State Chem. **4**, 281-285, 1972 (No. 2). *M*

G. Winkler, P. Hansen & P. Holst: Variation of the magnetic material parameters and lattice constants of polycrystalline yttrium-iron garnet by incorporation of nonmagnetic ions.
Philips Res. Repts. **27**, 151-171, 1972 (No. 2). *H*

P. L. Wodon: Methods of garbage collection for ALGOL 68.
ALGOL 68 Implementation, Proc. IFIP Working Conf., Munich 1970, pp. 245-262; 1971. *B*

L. E. Zegers: Common-bandwidth transmission of data signals and wide-band pseudonoise synchronization waveforms.
Thesis, Twente 1972. *E*

Contents of Philips Telecommunication Review **30**, No. 2, 1972:

F. L. van den Berg: The Lincompex radio-telephone terminal equipment of the RY 741 series (pp. 45-58).
H. van Kampen: The DS 714 system for telex (pp. 59-64).
A. J. M. Dingjan: The 8TR 602 pulse code modulation system (pp. 65-76).
S. de Vleminck & Y. Lebon: Digital telegraph and data transmission system type 3TR 1500 (pp. 77-86).
M. Etienne & J. P. Defeuilley: UHF shipborne transceiver ERM 7000 (pp. 87-90).

Contents of Philips Telecommunication Review **30**, No. 3, 1972:

P. Bikker: Frequency synthesizer RY 746 for HF receivers and transmitters (pp. 93-102).
H. L. Bakker: The 60 MHz coaxial transmission system 8TR 341 (pp. 103-112).
J. W. Scholten & H. Rebel: 60 MHz translating equipment 8TR 340 (pp. 113-126).
W. Fabich: Line conditioning for data transmission (pp. 127-134).

Contents of Electronic Applications Bulletin **31**, No. 1, 1972:

H. E. van Brück: Organization of ferrite core memories (pp. 2-27).
Intruder detector using a Gunn effect oscillator (pp. 28-36).
P. A. Neeteson: Integrating digital voltmeter: operating principles and accuracy (pp. 37-58).
J. M. Lavallee & J. Merrett: Reversible d.c. motor drive with regenerative braking (pp. 59-70).

Contents of Electronic Applications Bulletin **31**, No. 2, 1972:

J. E. Marquerinck: Optimizing the signal-to-noise ratio in a television camera pre-amplifier. An exercise in computer-aided design (pp. 74-132).

Contents of Electronic Applications Bulletin **31**, No. 3, 1972:

A. H. Hilbers: Large-signal behaviour of r.f. power transistors, Part 1. Analysis of the equivalent circuit (pp. 135-150).
E. Xanthoulis: Synthesis of active low-pass filters (pp. 151-186).
B. J. M. Overgoor: Error sources in analog multipliers (pp. 187-204).

Contents of Mullard Technical Communications **12**, No. 114, 1972:

T. G. Giles & J. E. Saw: Simple doppler radar using the CL8630 Gunn effect oscillator for the observation of small rotating objects (pp. 114-119).
I. Thomas: Design of a microwave oven cavity using two magnetrons, type YJ1371 (pp. 120-123).
J. A. Tijou: Audio amplifiers giving up to 3 W for use with TBA750 (pp. 124-128).
R. E. F. Bugg: Thyristor power supplies for television receivers: design considerations (pp. 129-144).

Contents of Mullard Technical Communications **12**, No. 115, 1972:

L. Hampson: Application of inverter-grade thyristors in a single-phase forced-commutation inverter (pp. 146-158).
Telecommunications Group: Transistorised 25 W v.h.f. a.m. transmitter for frequencies between 144 and 174 MHz (pp. 159-168).

Volume 32, 1971, No. 9/10/11/12 pages 281-420 Published 22th January 1973