An analogue computer for simulating one-dimensional aerial arrays

A. Meijer

In various fields of transmitting and receiving technology such as radar, radio astronomy, telemetry, and space flight, aerial arrays are used rather than single aerials. The mathematical treatment of such systems is rather complex, even when the treatment is confined to linear systems and the interaction between the elements is neglected. The article presents two methods by which the array factors of symmetrical, unequally spaced linear aerial arrays can be simulated and displayed electronically, thus providing a rapid understanding of the radiation behaviour of the system.

Introduction

Interest in aerials has increased considerably in the last ten or twenty years. The aerial is always a vital part of the system: a radar is made or marred by the quality of its aerials, without good aerials space flight would not have reached its present state of development, and radio astronomy would have made little progress without the aerial systems specially developed for it.

For many applications an aerial is desired that transmits in just one direction or receives from just one direction — an ideal that can be achieved only imperfectly. It has been found that the application of the well known *interference principle* familiar from optics can give a solution with some attractive aspects, particularly for radar. For example, the direction of radiation of the aerial can be varied electronically — and hence rapidly — without the inertia of mechanically rotating aerials. Another possibility is rapid adaptation of the radiation characteristics to the often variable environment in which the radar targets are situated. By using the interference principle more versatile radar aerials can be designed.

Since an aerial whose operation is based on the interference principle consists of more than one element, we shall henceforth speak of an *aerial array*.

Exact calculation of the radiation characteristics of an aerial array is extremely difficult, because the radiation characteristics of a single aerial in free space is different from that of the same aerial when it is part of an array. Here the various elements affect one another's performance in a very complicated way, depending on the type of aerial and on the geometry of the system. The problem is made very much simpler if this interaction is neglected. The errors introduced by neglecting the interaction become far less significant the farther apart the elements are located. Although there is usually a difference between the calculated radiation characteristics and the measured ones — even with widely spaced elements — the information obtained can be of value in designing an aerial array.

When the interaction between the elements is neglected, we may consider the aerial array as a group of elements each of which has a completely defined radiation pattern. This radiation pattern, which is never omnidirectional, is known as the *element factor*. If all the elements have the same element factor, we can find the radiation pattern of an aerial array by starting from a hypothetical array consisting of *omnidirectional* elements. This will be explained later. The radiation pattern of such a system is called the *array factor*. This array factor depends only on the feed to the individual elements and on the geometry of the array. The radiation pattern of the aerial array is then given by the product of the array factor and the element factor.

In this article we shall only deal with arrays whose elements are located on a straight line. They do not however all have to be at the same spacing. Unequally spaced linear aerial arrays of this type have become increasingly important in recent years. They were first used in radio astronomy, to obtain the same radiation pattern from a smaller number of aerials than are needed when the spacing between the elements is equal [1]. Unequally spaced aerial arrays also have features of interest for radar systems. Some practical linear arrays can be seen in *figs. 1a* and *1b*.

Even with the limitation to linear arrays and with the interaction between the elements neglected, the theor-

Ir. A. Meijer is with Philips Research Laboratories, Eindhoven.

[1] G. W. Swenson Jr. and Y. T. Lo, IRE Trans. AP-9, 9, 1961.

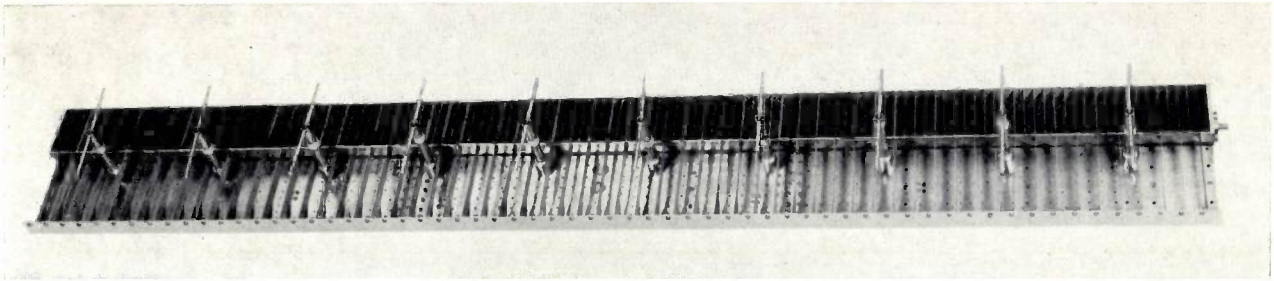


Photo: N.V. Hollandse Signaalapparaten, Hengelo (O), The Netherlands

Fig. 1*a*. A linear equally spaced aerial array for radar.

etical treatment of unequally spaced arrays is very difficult and as yet no satisfactory solution has been found. Any other method that can provide information about the relationship between the parameters describing such an array and its radiation pattern is therefore very welcome.

In the following it will be shown that the array factor of an unequally spaced linear array can be simulated electronically in two ways, making use of different analogies: the *space-time analogy*, which displays the array factor as a time phenomenon, and the *space-frequency analogy*, which displays the array factor as a frequency spectrum.

The space-time analogy has been found particularly useful for studying aerial systems whose direction of radiation can be varied electronically. This is not very easy with the space-frequency analogy. This analogy does however have the advantage that certain element factors can also be simulated.

Of the two simulation methods the one based on the space-frequency analogy can more easily be made to work with existing equipment. The space-time analogy, however, requires an analogue computer specially designed for this purpose. Since there is considerable interest in electronically steerable aerial systems, we have built such a computer.

In both methods the effect on the array factor of a change in the feed or in the arrangement of the aerial elements can be directly displayed on the screen of a cathode-ray tube.

In the following we shall deal first with the principles of simulation. A description of the electronic circuits of the analogue computer for the space-time simulation is then given. Finally, to illustrate the principle, some array factors obtained with the aid of the computer are presented. To complete the picture, a few results obtained by space-frequency simulation are included for comparison.

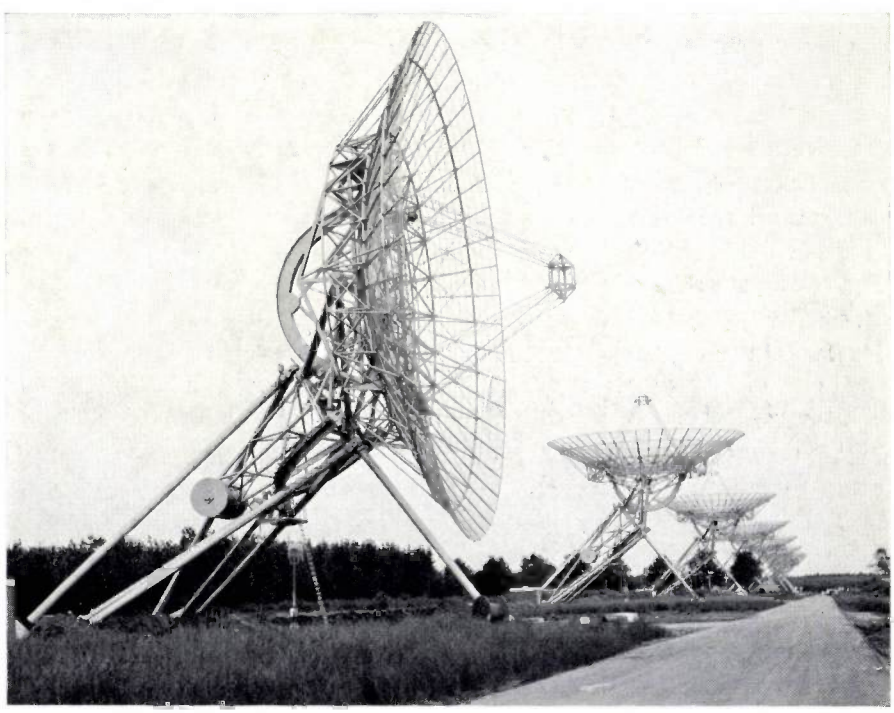


Fig. 1*b*. The radio telescope at Westerbork (The Netherlands), which was brought into use in 1969. To give a better view of the construction, the first element of the aerial has been turned out of position.

Theoretical principles of simulation

The array factor

Let us consider a linear array whose aerials are all fed from a single source by means of power dividers. Each aerial feed line is assumed to include an attenuator and a phase shifter, so that the feed to each aerial can easily be controlled in both amplitude and phase (*fig. 2*).

We shall now consider the *far field* of such a system, that is to say we consider a point far enough away from the aerial array that the array can be treated as a point source. This implies that the relative differences in path length from our point to the aerial elements no longer have any effect on the relative differences in the field strength of the individual waves reaching the point. The differences in path length do however affect the phase differences between the individual waves. The improvement in radiation behaviour that can be obtained with respect to that of a single-element aerial is due to these path-related phase differences.

It follows from what we have said that the far field of an aerial element may be regarded as a scalar quantity and that the total far field is found from a summation of the fields of the individual aerial elements, taking account of both amplitude and phase. The total far field of a system consisting of N aerial elements is thus given by the expression:

$$F_{\mathrm{tot}}(\Theta,\varphi) = \sum_{i=1}^{N} F_i(\Theta,\varphi), \quad \ldots \quad (1)$$

where $F_i(\Theta,\varphi)$ is the far field of the i-th element; the angular coordinates Θ and φ determine the direction of view (see *fig. 3*). The field $F_i(\Theta,\varphi)$ is determined by the strength a_i and the phase ψ_i of the signal supplied to the i-th element, by the element factor $e_i(\Theta,\varphi)$ of the i-th element and by the path-related phase difference with respect to a reference source. Assuming that the reference source is located at the point $s = 0$, as in fig. 3, and the i-th element at the point $s = s_i$, then this path-related phase difference is equal to $ks_i \sin\Theta$ rad, where k is the wave number ($k = 2\pi/\lambda$). The far field of the i-th element is thus given by:

$$F_i(\Theta,\varphi) = a_i \exp(\mathrm{j}\psi_i)\, e_i(\Theta,\varphi) \exp(\mathrm{j}ks_i \sin\Theta), \quad (2)$$

so that the total far field is:

$$F_{\mathrm{tot}}(\Theta,\varphi) = \sum_{i=1}^{N} a_i \exp(\mathrm{j}\psi_i)\, e_i(\Theta,\varphi) \exp(\mathrm{j}ks_i \sin\Theta). \quad \ldots \quad (3)$$

If all the elements have the same element factor, we can put this in front of the summation sign and write equation (3) as follows:

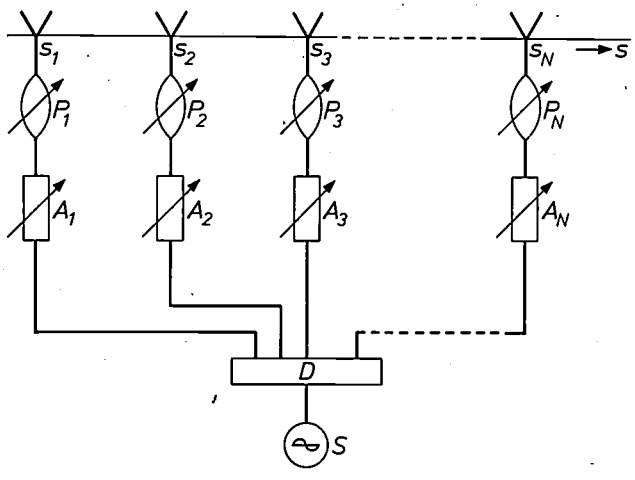


Fig. 2. Diagram of a linear aerial system. The elements are located on a straight line (coordinate of position s) and are fed via a power divider D from a common source S. Incorporated in each feed line is a phase shifter P_i and an attenuator A_i.

$$F_{\mathrm{tot}}(\Theta,\varphi) = e(\Theta,\varphi) \,.\, E(\sin\Theta), \quad \ldots \quad (4)$$

where

$$E(\sin\Theta) = \sum_{i=1}^{N} a_i \exp(\mathrm{j}(ks_i \sin\Theta + \psi_i)). \quad (5)$$

This function is the array factor.

It can be seen that the array factor may be regarded as the far field of a row of omnidirectional aerials; consequently it does not depend on the nature of the elements.

The aerial designer now has to find a set of amplitudes a_i, phases ψ_i and locations s_i that will give an "acceptable" array factor. Owing to the great number of variables it is by no means certain whether an array factor arrived at more or less fortuitously that has "good" characteristics will also be the best one. What we mean by "acceptable" or "good" will be seen later from the discussion of the results obtained.

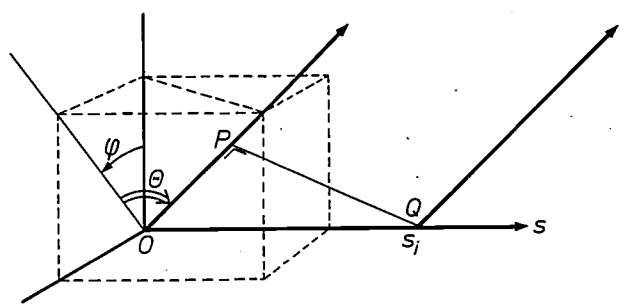


Fig. 3. The path-related phase difference, expressed in radians, of the signal of an element at Q at the position $s = s_i$ and a reference source at the origin, is 2π times the number of wavelengths contained in the projection OP of the distance OQ in the viewing direction (angular coordinates Θ and φ).

The space-time analogy

Fig. 4 illustrates schematically the method of simulation based on the space-time analogy. The generators G_i ($i = 1, 2, \ldots, N$) give voltages that vary with time as a cosine function and have the frequencies f_i. The initial phases ψ_i of these signals are set by means of phase-shifting networks. The amplitudes a_i of the

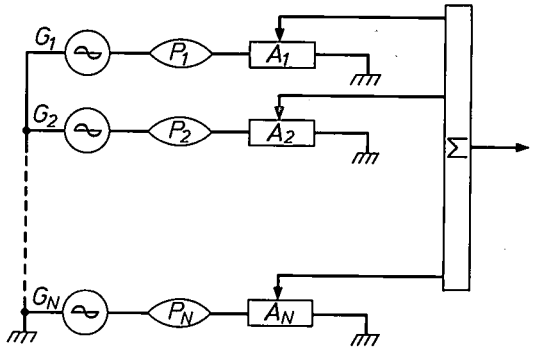


Fig. 4. Illustrating the simulation of array factors in the space-time analogy. G_i cosine voltage generators. P_i phase-shifting networks. A_i attenuators. Σ summation network.

signals can also be freely chosen. The i-th signal thus has the form:

$$V_i(t) = a_i \cos(2\pi f_i t + \psi_i). \quad \ldots \quad (6)$$

All the N signals of this nature are fed to a summation network Σ. The output signal $V_R(t)$ of this network is thus given by:

$$V_R(t) = \sum_{i=1}^{N} V_i(t) = \sum_{i=1}^{N} a_i \cos(2\pi f_i t + \psi_i). \quad \ldots \quad (7)$$

Substituting

$$t = T \sin \Theta \quad \ldots \quad (8a)$$

and

$$f_i = s_i/\lambda T, \quad \ldots \quad (8b)$$

we obtain for (7) the expression:

$$V_R(T \sin \Theta) = \sum_{i=1}^{N} a_i \cos(k s_i \sin \Theta + \psi_i), \quad (9a)$$

which is identical with the real part of (5). If *sinusoidal* voltages are generated by the sources G_i, we then obtain the *imaginary* part of equation (5):

$$V_I(T \sin \Theta) = \sum_{i=1}^{N} a_i \sin(k s_i \sin \Theta + \psi_i). \quad (9b)$$

By adding the squares of V_R and V_I we obtain the square of the array factor, represented as a function of time. It appears from the substitutions (eq. 8) that the time has been made equivalent to the space variable $\sin \Theta$. This is why we refer to this principle as the space-time analogy.

The constant T, which may at first sight seem to have been introduced simply to preserve the correct dimensions, also has a physical significance: the functions $V_R(t)$ and $V_I(t)$ only have physical significance in the t-interval $[-T, T]$, since a t-value outside this interval symbolizes a value of $\sin \Theta$ greater than 1, i.e. an imaginary angle. For this reason the interval $[-T, T]$ is also referred to as the "visible" region of t.

In the manner just described any array factor can be simulated with the aid of two sets of function generators. In nearly all aerial systems, however, the elements are arranged symmetrically about the centre of the system, and it is therefore sufficient to use only one set of generators. Mathematically this symmetry may be described by the condition:

$$s_i = -s_{N-i+1}.$$

This also establishes that the origin $s = 0$ of the s-coordinate coincides with the centre of the aerial system. The system then consists of M pairs of elements symmetrically arranged about the centre, possibly with an extra element at the centre (this will be the case if N is odd, that is to say if $N = 2M + 1$). If also each pair of elements is supplied with a signal of the same amplitude ($a_i = a_{N-i+1}$) but in opposite phase ($\psi_i = -\psi_{N-i+1}$), then equation (5) shows that each such element pair contributes to the function $E(\sin \Theta)$ a real amount $2a_i \cos(k s_i \sin \Theta + \psi_i)$, where $i = 1, 2, \ldots, M$. The array factor then becomes:

$$E(\sin \Theta) = a_0 + 2 \sum_{i=1}^{M} a_i \cos(k s_i \sin \Theta + \psi_i). \quad (10)$$

Here a_0 is the amplitude of the signal fed to the central element.

If we compare (10) with (9a) we see that the circuit shown in fig. 4 is sufficient for generating a function of time that is an adequate simulation of the array factor of an aerial system that is unequally spaced but symmetrically arranged about the centre, with equal amplitudes and opposite phases for each pair of elements. The term a_0 in equation (10) can easily be supplied by adding a d.c. source of the appropriate voltage.

A diagram explaining the simulation principle is shown in *fig. 5* for the case where $\psi_i = 0$. In fig. 5*a* the amplitudes and positions of the elements are presented schematically. The array factor as a function of $\sin \Theta$ then has a form like that shown in fig. 5*b*.

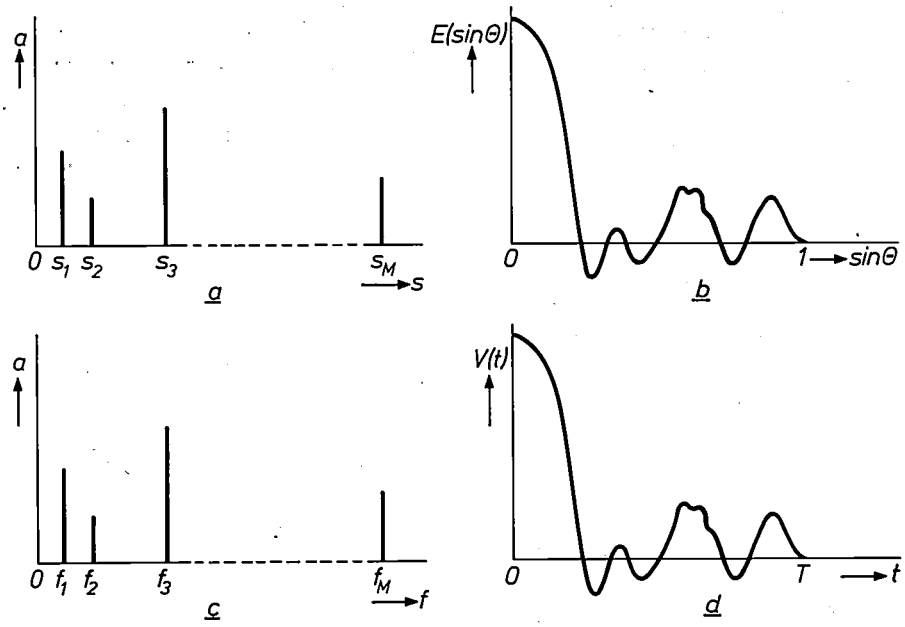


Fig. 5. The space-time analogy. *a*) The right-hand part of a symmetrical, unequally spaced linear aerial array represented schematically. The signal amplitude a_i is plotted vertically, the distance s_i from the aerial elements to the centre is plotted horizontally. *b*) The array factor $E(\sin\Theta)$ for the aerial array in (*a*). The function is symmetrical with respect to $\sin\Theta = 0$. *c*) Amplitudes a_i and frequencies f_i of a set of cosine generators with which the aerial array is simulated. *d*) The time function $V(t)$ of which (*c*) is the frequency spectrum. The generated function only has a physical significance (Θ real) for $t < T$.

The space-time analogy consists in the correspondence of (*b*) and (*d*) to each other when (*a*) and (*c*) correspond.

Since the system is symmetrical, this far field will also be symmetrical with respect to $\sin\Theta = 0$; it is therefore sufficient to draw this field for values of $\sin\Theta$ between 0 and 1.

Fig. 5*c* gives the amplitude and the frequency of each cosine generator. The time function is shown in fig. 5*d*. The space-time analogy is found in the fact that figs. 5*a* and *b* are equivalent to figs. 5*c* and *d*. If we change the frequency of the second cosine generator, for example, then the effect of this on the time function in fig. 5*d* is the effect which a corresponding displacement of the second pair of elements would have on the far field in fig. 5*b*.

The space-frequency analogy

The conclusion to be drawn from fig. 5 is that the configuration of the aerial array (fig. 5*a*) can be regarded as the spectrum of the far field (fig. 5*b*), since fig. 5*c* is of course the spectrum of the time function of fig. 5*d* (see also equation 7). Now a time function and its spectrum are Fourier transforms of one another. This means that the far field of an aerial array can be described as the Fourier transform of the configuration of the array. This conclusion is the basis for the space-frequency analogy, whose principle is illustrated in *fig. 6*.

In fig. 6*c* a time function is shown that consists of

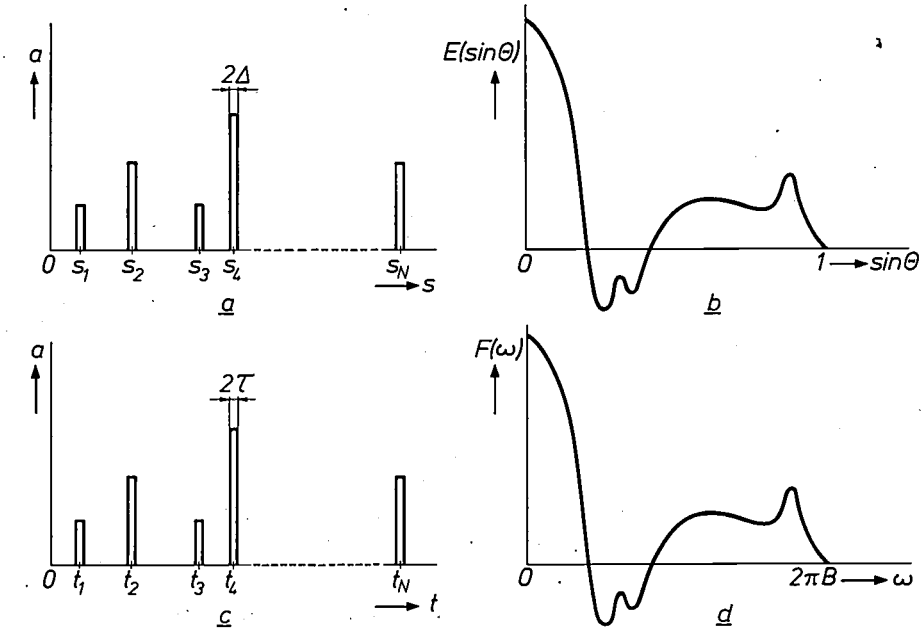


Fig. 6. The space-frequency analogy. *a*) An unequally spaced linear aerial array, represented schematically. The elements are line sources of length 2Δ. The signal amplitude a_i is plotted vertically, the distance s_i of the element from the origin, which is taken to be at the lefthand end of the array is plotted horizontally. *b*) The array factor of the aerial system as a function of $\sin\Theta$. This diagram is the Fourier transform of the function shown in (*a*). *c*) A series of pulses of width 2τ, with which the aerial array is simulated. The pulse amplitudes are plotted vertically, the times at which the pulses occur are plotted horizontally. *d*) Frequency spectrum of the time function of (*c*). The spectrum has physical significance (Θ real) only when $f < B$. If (*a*) and (*c*) correspond, so do (*b*) and (*d*).

a series of pulses of peak amplitudes a_i, which occur at the times t_i. The width 2τ of the pulses is small compared with the distance between them. The frequency spectrum $F(\omega)$ (fig. 6*d*) is the Fourier transform of this time function. Fig. 6*a* is an accurate reproduction of fig. 6*c* and represents our aerial system, in which signals of amplitude a_i are fed to elements of finite length 2Δ located at the points s_i; in other words the elements must now be treated as *line sources* instead of omnidirectional point sources. Since the far field of this system as a function of sin Θ (fig. 6*b*) is the Fourier transform of this configuration, this far field must have exactly the same curve as the frequency spectrum in fig. 6*d*.

The space-frequency analogy thus amounts to the generation of a signal whose variation with time is a faithful representation of the configuration of the array to be studied. In this case the array need not be symmetrical. The frequency spectrum of this signal is then the representation of the far field of the aerial array.

consisting of a series of pulses with the form of a half-cosine function (*fig. 7*).

Fig. 8 shows a diagram illustrating the simulation method just described. Each time the pulse generator *PG* delivers a pulse, the pulse shapers PS_i generate a pulse of a particular shape. This pulse is delayed t_i

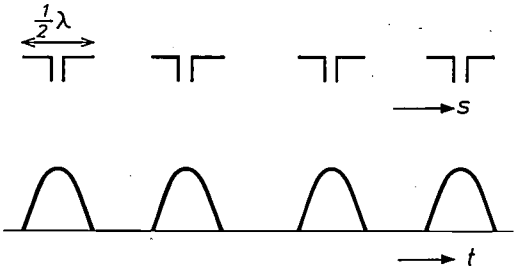


Fig. 7. Example of the way in which the element factor can be taken into account in the far field by giving the pulse signals a particular shape. *Above:* Aerial system consisting of four ½λ dipoles. *Below:* Series of pulses representing the current distribution in the elements and simulating this aerial array via the space-frequency analogy.

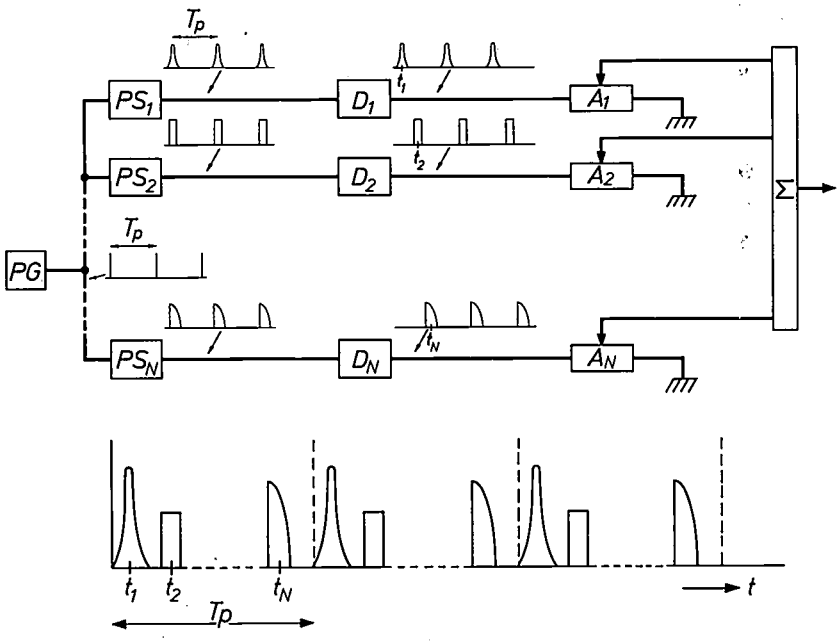


Fig. 8. Diagram of arrangement for simulating an array factor on the space-frequency analogy principle. Every T_p seconds a pulse generator *PG* delivers a clock pulse that causes the pulse shapers PS_i to deliver a pulse of a particular shape. This pulse is delayed by the circuit D_i (t_i seconds) and if necessary attenuated to the amplitude a_i by A_i. The signals are then added in a summation network Σ and the sum signal goes to a spectrum analyser (not shown). An example of a sum signal is shown under the diagram.

In certain cases the space-frequency analogy allows the element factor to be taken directly into account by giving the pulses a suitable shape. For example, the far field of a group of half-wavelength dipoles can be represented by the frequency spectrum of a pulse train

seconds in a delay network D_i and is then attenuated to the required amount by an attenuator A_i. All the pulses are then fed to a summation network Σ, and the resulting signal goes to a spectrum analyser.

Since the spectrum analyser must be supplied with

a periodic signal, the pulse train representing the aerial system must be periodically repeated. The repetition frequency is $2\pi/T_p$, where T_p is the period of the pulse train. This means that the screen of the spectrum analyser shows a line spectrum with lines at the frequencies that are the harmonics of this repetition frequency. The envelope of this line spectrum then represents the array factor.

From the mathematical treatment of this simulation principle (see the small print below) it appears that the space-frequency analogy is only suitable for simulating the far field of an aerial system all of whose elements are fed in phase. However, the equipment can be modified in such a way that the simulation method can also be applied for an array whose elements are fed in different phases, but we shall not go any further into this here.

The calculation also shows that it is not necessary to display the whole frequency spectrum. Only frequencies between particular values $-B$ and $+B$ (see fig. 6) represent real Θ-values. The frequency interval $[-B, B]$ is thus the visible region of the frequency spectrum, just as in the space-time analogy a time interval $[-T, T]$ represented the visible region.

The space-frequency analogy can be verified mathematically as follows. Since the signal is periodic with a period T_p (see fig. 8), it can be expressed in a Fourier series:

$$f(t) = \sum_{-\infty}^{+\infty} F_n \exp\left(-j\frac{2\pi nt}{T_p}\right). \quad \ldots\ldots (11)$$

The Fourier coefficients F_n are found by using the expression:

$$F_n = \frac{1}{T_p}\int_0^{T_p} f(t) \exp\left(j\frac{2\pi nt}{T_p}\right) dt. \quad \ldots (12)$$

In view of the shape of the signal, this integral can be written as the sum of N integrals:

$$F_n = \sum_{i=1}^{N} \frac{1}{T_p}\int_{t_i-\tau}^{t_i+\tau} f_i(t) \exp\left(j\frac{2\pi nt}{T_p}\right) dt =$$

$$= \sum_{i=1}^{N} \frac{1}{T_p} a_i\, e_i\left(\frac{2\pi n}{T_p}, \tau\right) \exp\left(j\frac{2\pi nt_i}{T_p}\right), \quad \ldots (13)$$

where

$$e_i\left(\frac{2\pi n}{T_p}, \tau\right) = \int_{-\tau}^{+\tau} f_i(t) \exp\left(j\frac{2\pi nt}{T_p}\right) dt$$

represents the element factor of the i-th element.

Now the frequency spectrum of the aperiodic signal is proportional to

$$F(\omega) = \sum_{i=1}^{N} a_i\, e_i(\omega,\tau_i) \exp(j\omega t_i), \quad \ldots\ldots (14)$$

so that the Fourier coefficient F_n has been shown to be no more than the sampling of the frequency spectrum at the frequency $2\pi n/T_p$.

To establish the analogy of equation (14) with the expression for the array factor (equation 5) we assume that all the pulses are equal. In this case $e_i(\omega,\tau_i)$ is identical for each pulse, and this factor can be put in front of the summation sign. We then obtain:

$$F(\omega) = e(\omega,\tau) \,.\, E(\omega),$$

where

$$E(\omega) = \sum_{i=1}^{N} a_i \exp(j\omega t_i). \quad \ldots\ldots (15)$$

Substituting

$$\omega = 2\pi B \sin\Theta \quad \ldots\ldots\ldots (16a)$$

and

$$t_i = s_i/\lambda B \quad \ldots\ldots\ldots (16b)$$

in this expression, then for the case $\psi_i = 0$ — i.e. where the elements are fed with signals of the same phase — equation (5) is identical with equation (14).

The analogue computer based on the space-time analogy

We shall now look a little more closely at the electronic circuits of the analogue computer for the space-time analogy [2]. The basic circuit diagram of this equipment has already been shown in fig. 4; a photograph of the equipment can be seen in *fig. 9*.

We saw earlier that the function generators G_i in fig. 4 have to be *cosine* generators of variable frequency. If we want to simulate unequally spaced aerial arrays, the frequencies of these generators must not be harmonics of one another, which means that the generated time function is not periodic. In order nevertheless to obtain a stationary picture on an oscilloscope, the part of the function in which we are interested is periodically repeated. This is done by making all the generators run for T seconds only, starting them again after $T + \Delta T$, stopping them again after $2T + \Delta T$, and so on. The time T is chosen such that only the right-hand half of the visible region is described (between $\sin\Theta = 0$ and $\sin\Theta = 1$, see fig. 5). The start and stop signals are derived from the sawtooth generator of the oscilloscope, as shown in *fig. 10*. The time ΔT is made equal to the flyback time of the oscilloscope trace.

At the moment the start pulse appears, all the generators must immediately start generating a cosine function of the right phase, frequency and amplitude. There must be no switching transients; the generated function must be undistorted during the period T, and it must also be accurately reproducible during a very large number of sawtooth periods. In addition it must be possible to modify the function in a fast and reliable way.

This is done by deriving all the frequencies of the cosine functions from the same train of high-frequency pulses, which come from a central pulse generator, called the master clock. New series of pulses of lower frequencies are derived from this pulse train by means of frequency dividers. These pulse trains, whose frequencies correspond to the positions of the elements

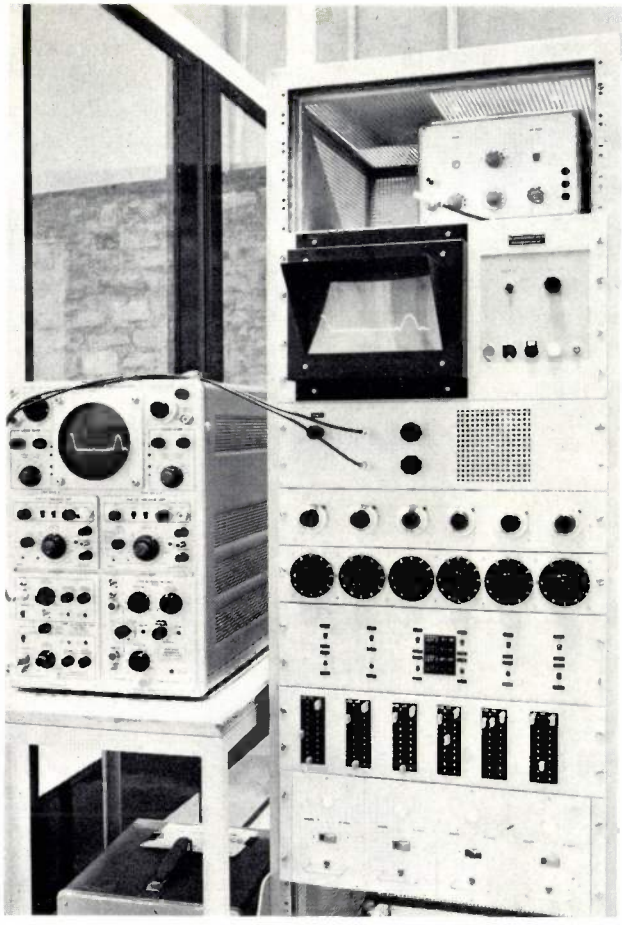


Fig. 9. The analogue computer for the simulation of array factors by the space-time analogy (right). The system is capable of simulating aerial arrays with up to 13 elements. The array factor can be displayed on the oscillosope screen (left) and also on the upper panel of the computer. The master clock is a separate unit at the top. Under the display panel can be seen, from top to bottom, the modulus shaper, the amplitude controllers, two panels with phase-selector switches and, on the second panel up, six frequency-selector switches.

of the aerial array being studied, are fed to function generators. As we shall see later, each pulse that arrives at the input of a function generator initiates a step in the output voltage, the successive steps being of the correct heights for the waveform to approximate to a cosine function. It is clear that the frequency of the generated function will be lower than that of the pulse train at the input, and as many times lower as the number of steps in which the function is approximated. We have taken this number of steps as 24, which gives a favourable compromise between the time needed for generating a complete cosine function and the accuracy of the approximation. The function obtained in this way can be started and stopped at any time without giving rise to switching transients, and since all the generators are driven by a common pulse generator there is no frequency drift between the signals.

The initial phases of the cosine functions generated are adjusted in the following way. We arrange that the first m pulses go straight to the cosine generator, to be followed by the pulses from the frequency divider of frequency f_i'. The cosine generator then starts with the value $\cos 2\pi m/p$ (p being the number of steps in which the cosine function is approximated), so that a cosine

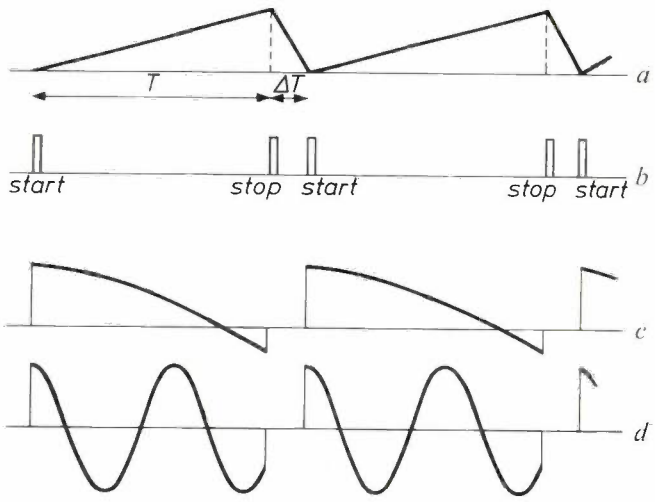


Fig. 10. The signals from the cosine generators are made periodic by applying start and stop signals derived from the sawtooth voltage of the oscilloscope. *a*) Sawtooth voltage. T is the sweep time of the oscilloscope, ΔT the flyback time. *b*) Start and stop pulses derived from the sawtooth voltage. *c*) and *d*) Examples of the cosine functions of different frequencies generated in the period 0-T.

[2] An important contribution to the construction of the analogue computer was made by J. Hopstaken, formerly with this laboratory.

signal is generated of frequency f_i'/p and initial phase $2\pi m/p$ rad. There are other well known analogue-computer techniques for generating cosine functions, but for various reasons they were not so suitable for our purpose.

A more detailed circuit diagram of the analogue computer is given in *fig. 11*. The unit *M* (on the left) is the master clock, whose output signal (frequency about 100 kHz) goes to the control units CD_i. These control units supply the cosine generators CG_i at the appropriate instants with the required pulse trains, as described above. First they supply a pulse train for setting the initial phases of the cosine generators, and then pulse trains of lower frequencies corresponding to the positions of the symmetrically arranged pairs of elements. The cosine generators CG_i transform the pulse trains from the frequency dividers into step-approximated cosine functions. After any attenuation that may be necessary, the signals are summed in the summation network Σ and supplied to the oscilloscope *O*. Since often only the moduli of the array factor are required, the signals are passed through a modulus shaper *MS*. This is a rectifying circuit which reverses the polarity of all voltages below a certain level (i.e. reverses them with respect to that level), so that the output signal is the modulus of the sum of the input signals (see equation 5), plus a constant amount which represents the signal amplitude of the central element of the aerial system. The unit *St* is a circuit that derives the start and stop pulses for the control circuits from the sawtooth voltage of the oscilloscope.

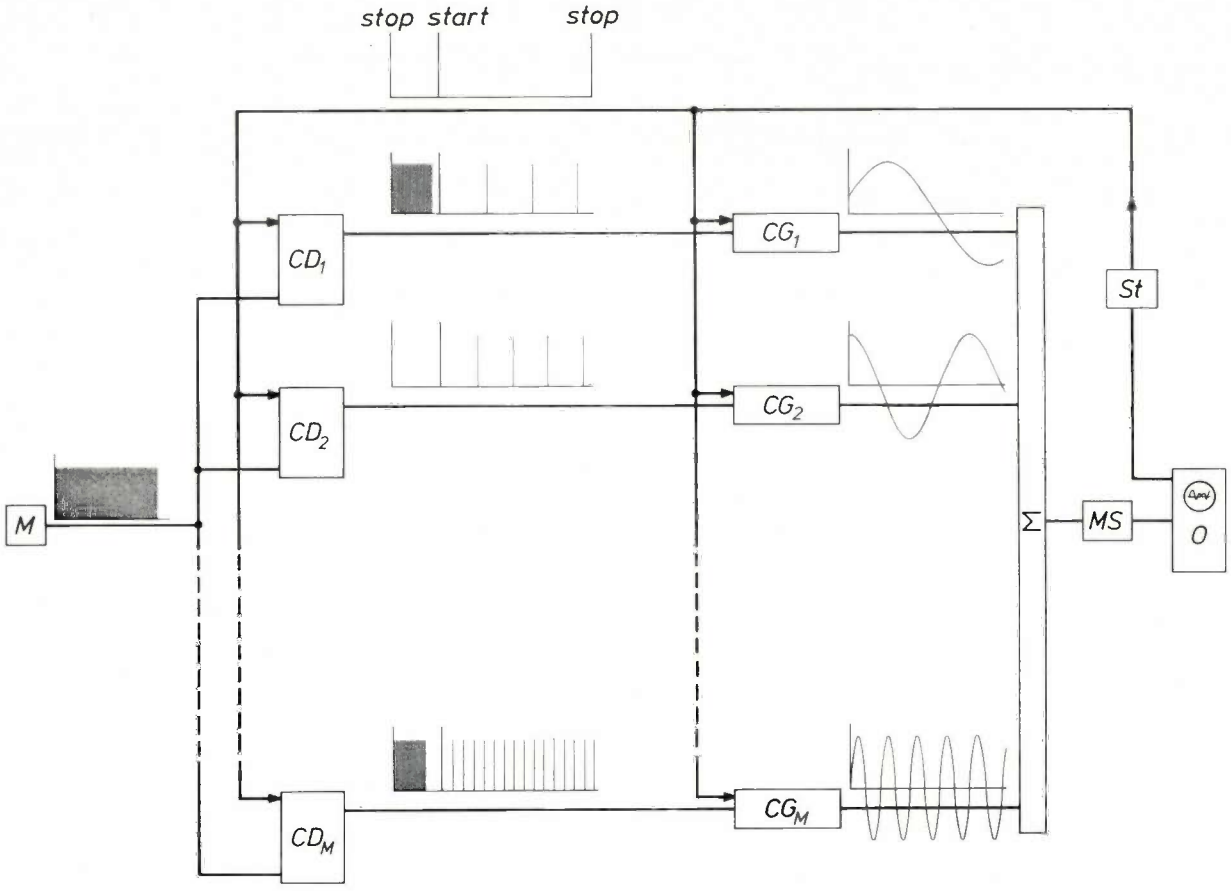


Fig. 11. Basic diagram of the analogue computer for simulating array factors by the space-time analogy, with schematic indication of the various signals. *M* master clock. CD_i control units. CG_i cosine generators. *St* circuit supplying the start and stop signals for the frequency dividers and the cosine generators. Σ summation network, in which the incoming signals are attenuated if necessary. *MS* modulus shaper. *O* oscilloscope.

The circuit arrangement of the control units is shown in *fig. 12*. The vital unit is the decimal counter *C*, which consists of a number of decade counters of the ring type [3]. The counter *C* has two functions in the control unit; it acts as a frequency divider and it counts the

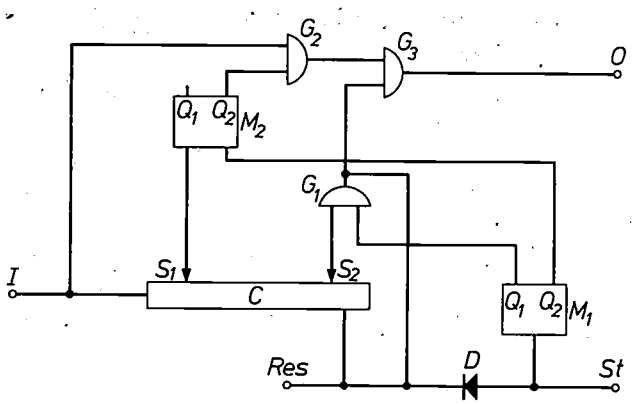


Fig. 12. Control unit for the cosine generator. The situation before the appearance of a stop pulse is as follows. The Q_2 terminal [4] of the bistable circuit M_2 is low, so that the AND gate G_2 is closed. The Q_1 terminal of the bistable circuit M_1 is high, and therefore the AND gate G_1 is open. The stop pulse now appears at *St*. This causes M_1 to change its state: Q_1 becomes low and G_1 closes. At the same time Q_2 of M_2 goes high, so that G_2 opens. The result is that the master-clock pulses entering at I, which are counted by the digital counter C, are now transferred straight to the output terminal O by means of the OR gate G_3. After the selector switch S_1 has counted m clock pulses, M_2 changes state, causing G_2 to close and thus preventing the clock pulses from reaching O. This state continues until the start pulse arrives at *St*. This causes M_1 to change state and G_1 then opens. (G_2 remains closed because M_2 does not respond to a trailing edge.) At the same time the counter C is reset to zero, as the starting pulse also appears at the reset terminal *Res* of the counter via the diode D. After the transfer of each predetermined number of pulses, counted off by the selector switch S_2, the digital counter is reset. (The diode D prevents this reset pulse from making M_1 change state.) In this way a series of pulses of the required frequency is obtained at the output terminal O. After the next stop pulse has appeared at *St*, the cycle starts again.

transferred from the one bistable circuit to the next one on the arrival of a trigger pulse at the trigger gate T. The output voltage from each bistable circuit is fed through a diode to the resistance network. The voltage V_0 appearing at the generator output in any situation is equal to the fraction of this voltage that is determined by R_i and R_0. The values of the resistors R_i are chosen to make the output voltage give a stepped approximation of a cosine function.

The actual circuit is somewhat more complicated than the one shown in fig. 13. It contains only seven bistable circuits but is nevertheless able to approximate the cosine function in 24 steps. This is done by means of a two-way shift register and by using the Q_2 terminals of the bistable circuits. We shall not go into the details here.

Some radiation patterns obtained with the two methods of simulation

We shall now give as an example some radiation patterns obtained with the analogue computer described above, and in comparison some patterns obtained with an arrangement based on the space-frequency analogy. These were not simulations of the radiation pattern of aerial arrays designed for any particular applications, but simulations of arrays that were chosen because they give a good illustration of the method.

Fig. 14 shows the computer-simulated radiation pattern of three aerial arrays, of two, four and six elements, and fed with signals of equal amplitude and phase. All the elements were spaced at 2λ.

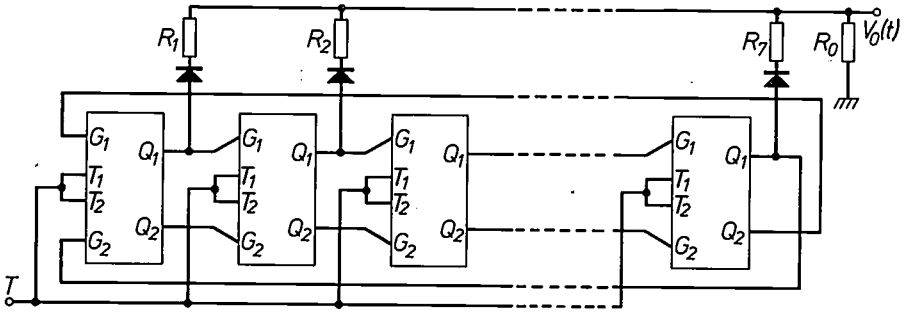


Fig. 13. Simplified circuit diagram of the cosine generator. All bistable circuits except one, e.g. the i-th, are at a low level (Q_1 terminal at 0 volts). The Q_1 terminal of the i-th bistable circuit is at a high (negative) level (−6 volts), so that only the voltage divider consisting of R_i and R_0 is operative. $V_0(t)$ is thus equal at this moment to $-6R_0/(R_0 + R_i)$ volts.

number of clock pulses m corresponding to the required initial phase of the cosine function.

The cosine generator is a combination of a shift register [3] and a resistance network. The principle is illustrated in *fig. 13*. The shift register is composed of bistable circuits (flip-flops), arranged in such a way that only the Q_1 terminal of one bistable circuit is at the high voltage at any given moment. This state is

Fig. 14*a* gives the output signal of the summation

[3] C. Slofstra, The use of digital circuit blocks in industrial equipment, Philips tech. Rev. **29**, 19-33, 1968.

[4] We use the standardized letter code for the various terminals; see for example the article by Slofstra [3] and our fig. 13.

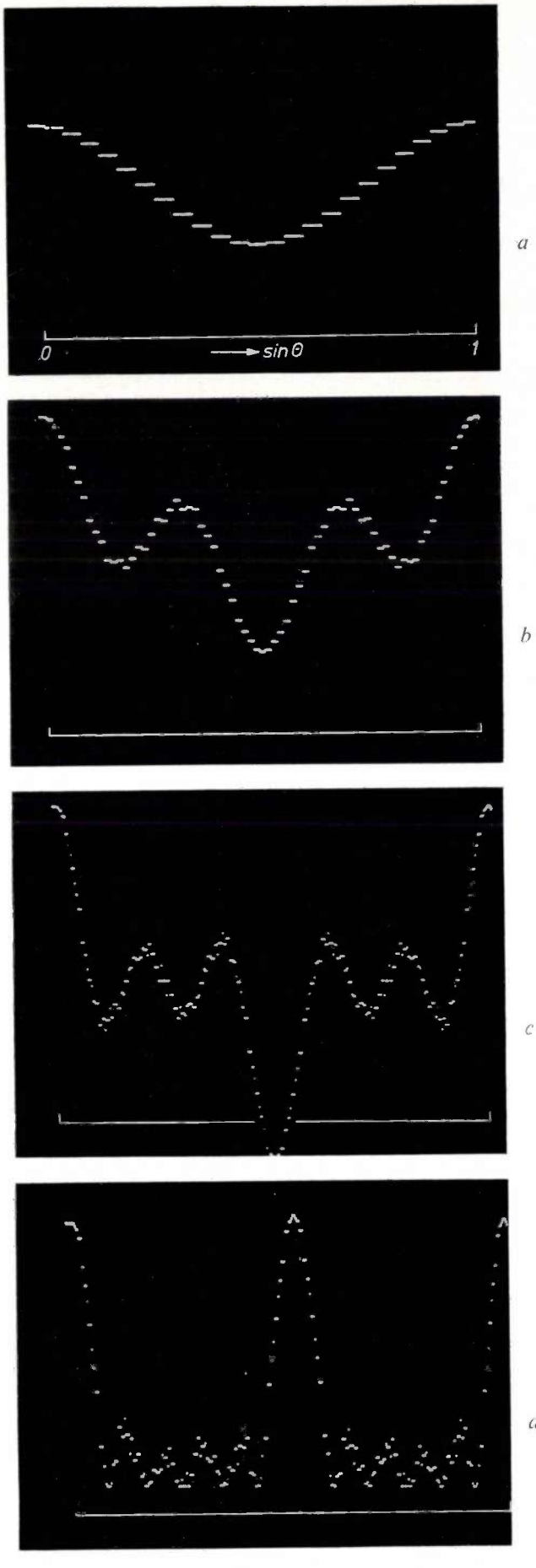


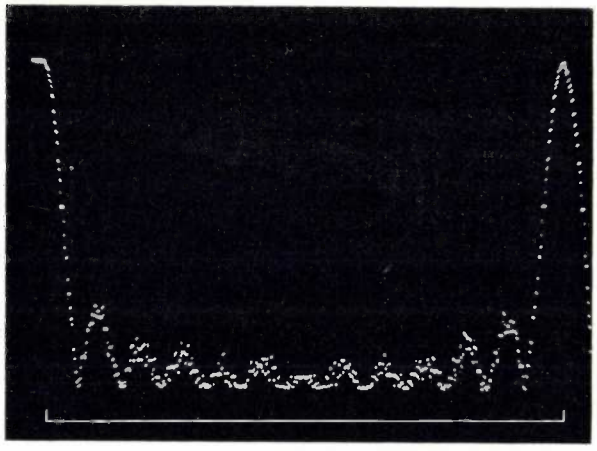
Fig. 14. Radiation patterns (space-time analogy) of three aerial arrays whose elements are identical and have a spacing of 2λ. *a*) Diagram for a two-element system. *b*) Diagram for a four-element system. *c*) Diagram for a six-element system. *d*) The same system as in (*c*) but now showing the modulus of the signal.

circuit as a function of sin Θ for an array of two elements. This signal is produced by a cosine generator of frequency f_0. As the spacing between the elements is 2λ, a whole period of the cosine function is generated (see eq. 8b). It can clearly be seen that the cosine function is approximated in 24 steps per period. Fig. 14*b* shows how the radiation pattern changes when we go from two to four elements. This change makes it necessary to include in the computer a second cosine generator, which supplies a signal of frequency $3f_0$ to the summation network. Fig. 14*c* gives the corresponding diagram of the aerial array with six elements. In fig. 14*d* the more usual presentation for the aerial array of fig. 14*c* can be seen: the output signal of the modulus shaper plotted as a function of sin Θ.

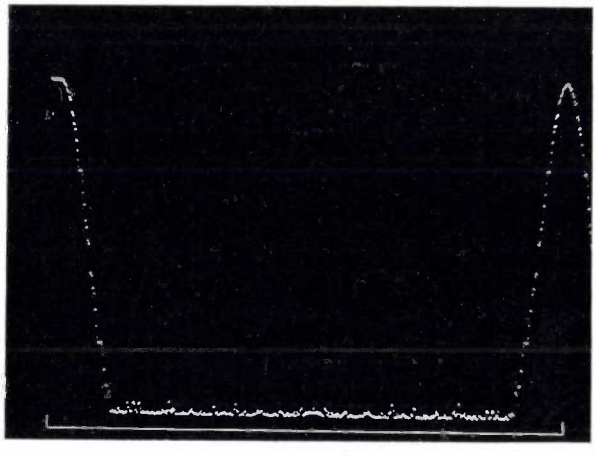
A similar radiation pattern is presented in *fig. 15a*, but here for a system with 13 symmetrically arranged elements at a spacing of one wavelength. Fig. 15*b* shows that it is possible to make the level of the side lobes very low for this kind of system. This diagram was obtained by feeding the elements with signals of different amplitude.

The signal amplitudes a_i for the elements were chosen on the basis of a solution given by T. T. Taylor [5] for a line source.

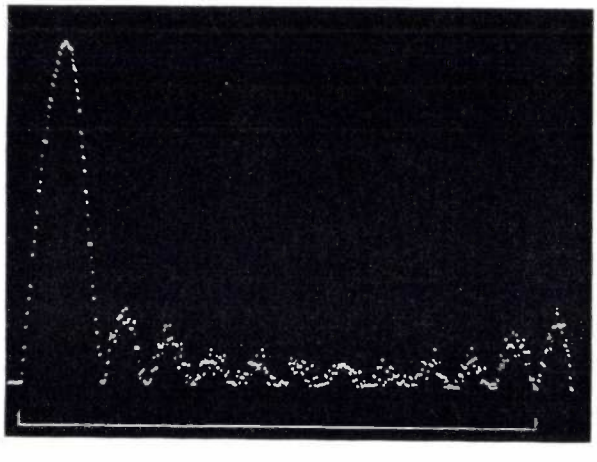
[5] T. T. Taylor, IRE Trans. **AP-3**, 16, 1955.



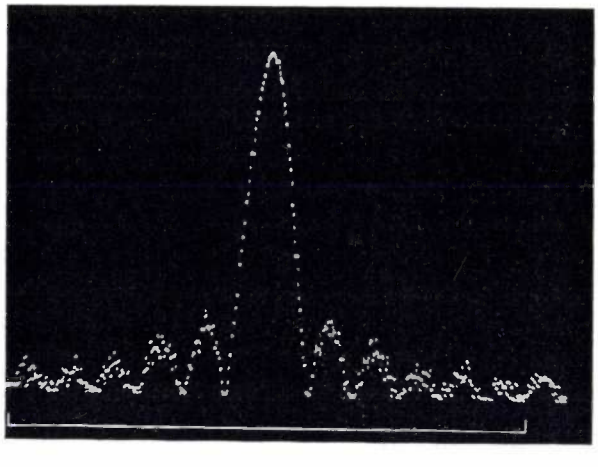
a



b



c



d

Fig. 15. Radiation pattern (space-time analogy) of an equally spaced linear aerial array with 13 elements at a spacing of one wavelength. *a*) All elements are fed in phase with signals of the same amplitude. *b*) The elements are fed with signals of different amplitude; the side lobes are now lower. *c*) The signals fed to the elements are of the same strength but the phase for each element is displaced by 27° with respect to that of the previous one. *d*) The phase of each element is displaced by 162° with respect to the previous one.

For a side-lobe level of —35 dB he gives the curve shown in *fig. 16*, representing the relative signal amplitude $f(s)$ as a function of position. The values a_i that we have chosen are proportional to the ordinate values of $f(s)$ at the locations $s = n\lambda$ (n = 0-6), for the right-hand aerial elements.

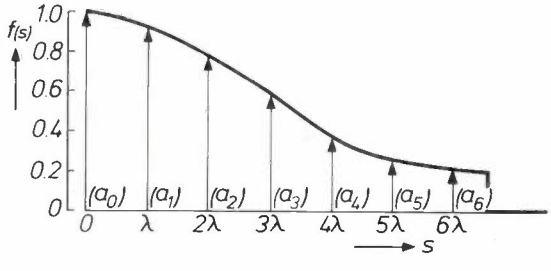


Fig. 16. Curve showing how the signal amplitude $f(s)$ of a line source should vary with position to give a side-lobe level of —35 dB.

The patterns shown in figs. 15*c* and 15*d* show how the direction of radiation of an aerial array can be varied by means of the phases of the signals fed to the elements. Fig. 15*c* relates to a 13-element aerial system for which the feed for each element was 27° different in phase from the preceding one over the whole length. In fig. 15*d* this relative phase difference is 162°. We see that the whole pattern remains intact, only the main beam being displaced. This means in fact that the whole

aerial pattern has been rotated with respect to the aerial array. As we noted in the introduction, this can be a very attractive feature for radar systems.

Figs. 17 and *18* demonstrate what can be achieved by varying the positions of aerial elements radiating with equal amplitude and phase. If the elements are spaced at $\frac{1}{2}\lambda$, which is often the case in equally spaced aerial arrays, then the higher-order maxima fall outside the visible region, and the side-lobe level decreases with increasing Θ. Fig. 17*a* shows a radiation pattern for such an aerial array with 13 elements. By spacing the elements unequally the side-lobe level can be reduced from —13 dB to —20 dB (see fig. 17*b*).

Fig. 18 illustrates how maxima of higher order can be suppressed by suitably arranging the elements of an aerial array; this is important when one very narrow main beam is required without the side-lobe level having to be very low.

Finally, we shall present some patterns obtained by means of the method of simulation based on the space-frequency analogy.

Fig. 19a shows the radiation pattern of an equally spaced aerial array consisting of seven line-source elements 0.1λ long and at a spacing of 2λ. They are fed in equal amplitude and in phase. With the bandwidth of the spectrum analyser set to 100 kHz the system is represented — as eq. 16b will show — by a series of seven pulses of width 1 μs and 20 μs apart. The oscillogram of fig. 19*a* was obtained by repeating these pulses at a repetition frequency of 3.5 ms, so that the picture consists of 350 lines.

The elements in the array of fig. 19 are so short that the element factor has no effect. If they are made 10 times longer — corresponding to a pulse width of 10 μs — then the element factor does have an effect. It can be calculated that the first zero of the element

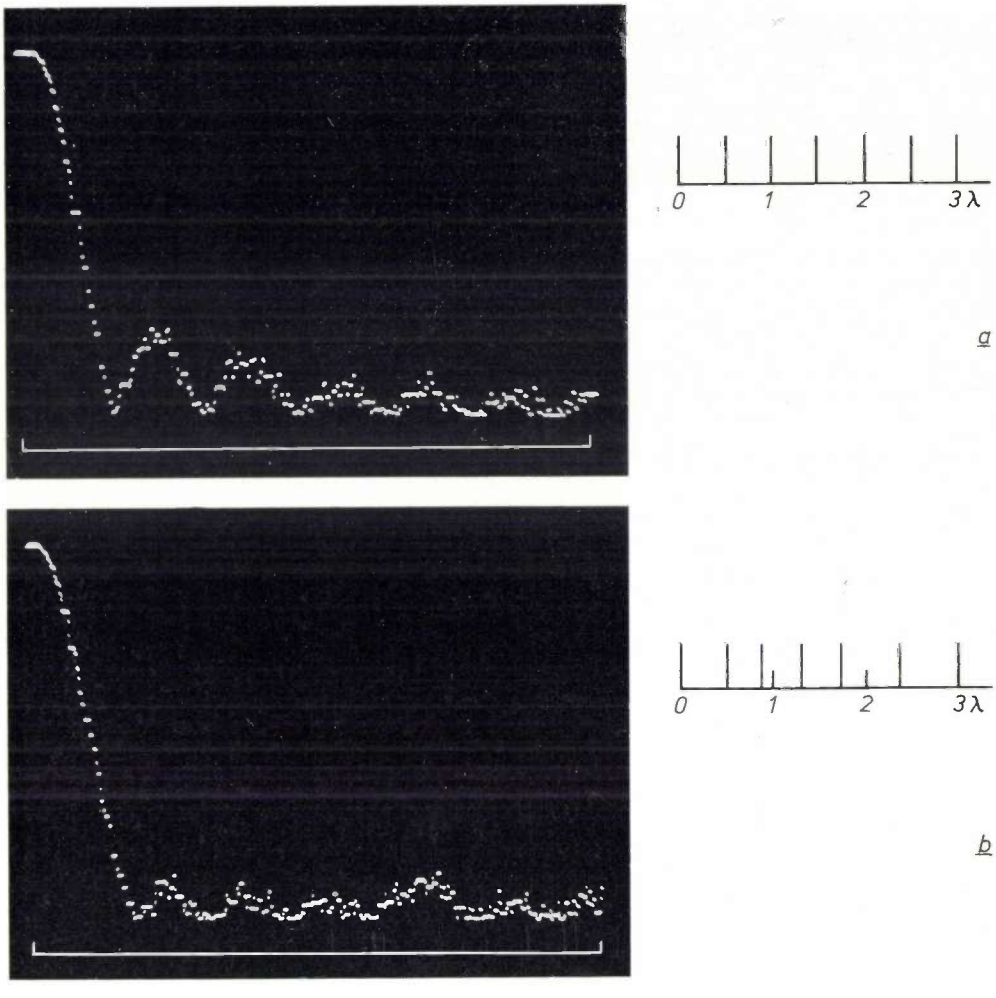


Fig. 17. Reduction of the side-lobe level by making the spacing of the elements unequal. The elements are fed with signals of the same amplitude and in phase. *a*) Radiation pattern of an aerial array with 13 elements, all spaced at $\frac{1}{2}\lambda$. *b*) The same system after changing the spacing between the elements (see the position diagram).

factor $e(\omega,\tau) = (\tau/T_p)(\sin\omega)/\omega\tau$ then coincides exactly with the third main lobe of the array factor, so that this lobe is not to be found in the radiation pattern (fig. 19*b*).

As can be seen from the earlier figures, the radiation patterns obtained by using the space-time analogy have the disadvantage that the picture looks as if it has noise superimposed on it. This "noise" appears because the

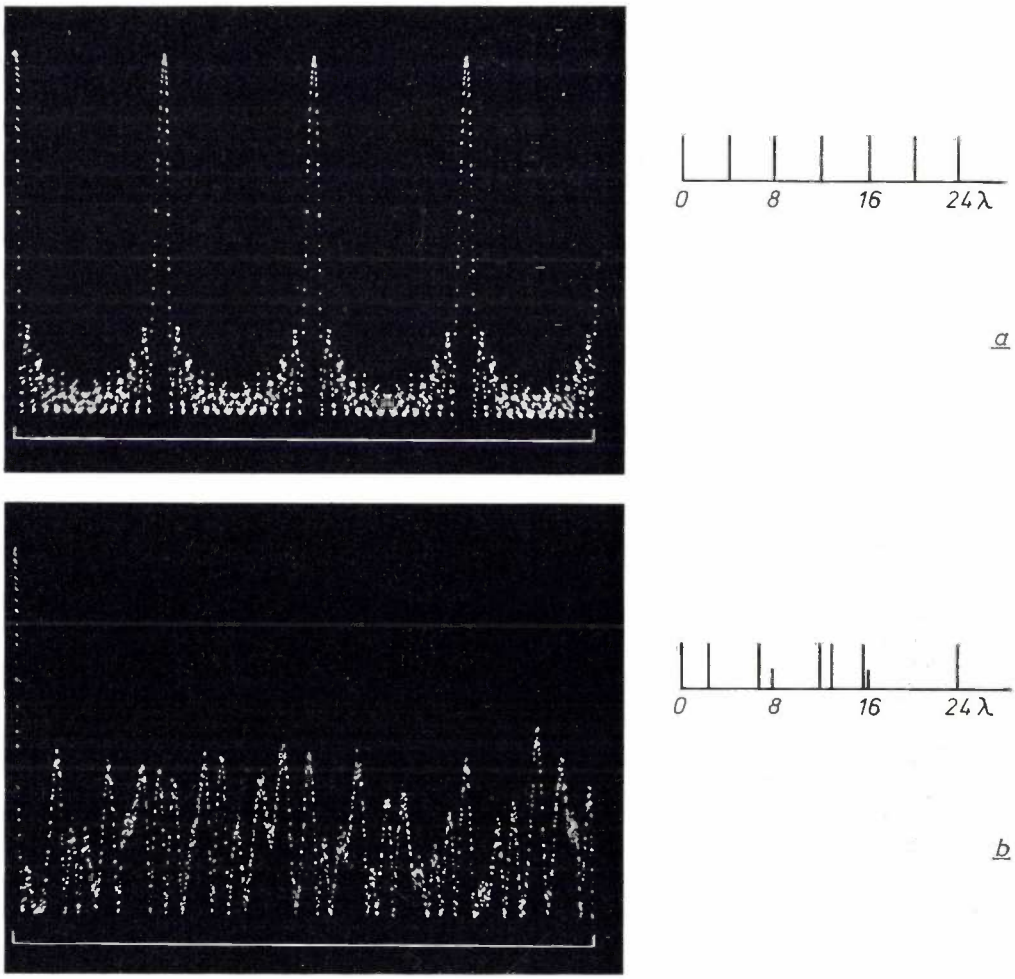
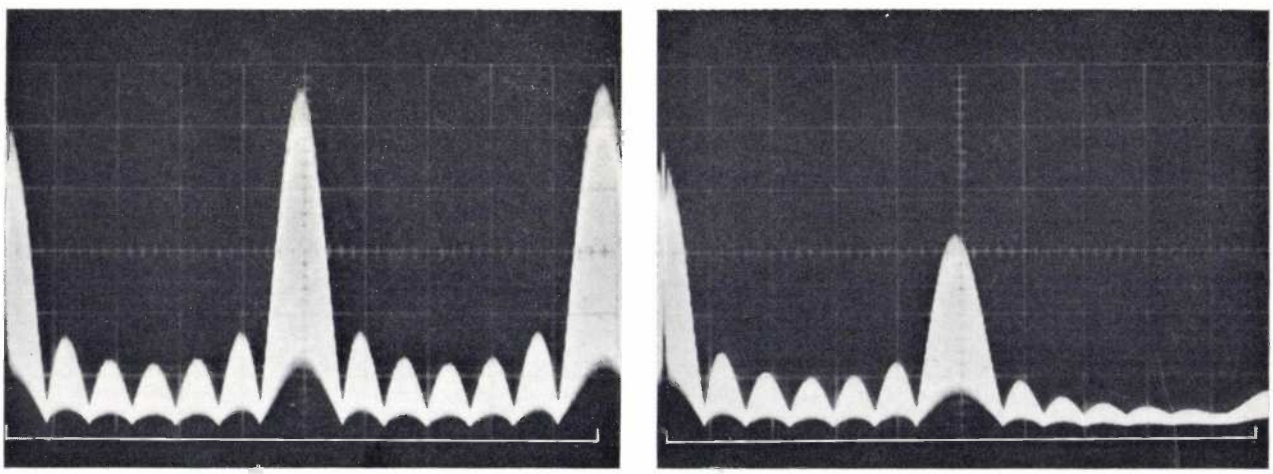


Fig. 18. Suppression of higher-order maxima by changing the spacing of the elements. *a*) Radiation pattern of an aerial array with 13 equally spaced elements (spacing 4λ). *b*) As a result of changing the spacing the three higher-order maxima have been suppressed, though at the expense of a higher side-lobe level.



a *b*

Fig. 19. *a*) Radiation pattern (space-frequency analogy) of an equally spaced aerial system with seven elements. *b*) The same aerial array, but now the element factor is included in the simulation.

cosine functions are generated in a stepped pattern. A particular example is to be seen in *fig. 20a*. A comparable pattern obtained by making use of the space-frequency analogy is much clearer (fig. 20*b*).

This disadvantage of the analogue computer can be removed by not approximating each cosine function to be generated in an equal number of steps per period, but by generating the values of these functions sampled at the same times. This requires a much more elaborate circuit, which we shall not describe here. It is worth noting, incidently, the good agreement between the patterns in fig. 20*a* and *b*.

Although the analogue computer described in this article should not be considered any more than a prototype, still capable of several improvements, it has already been found a useful aid in designing aerial arrays.

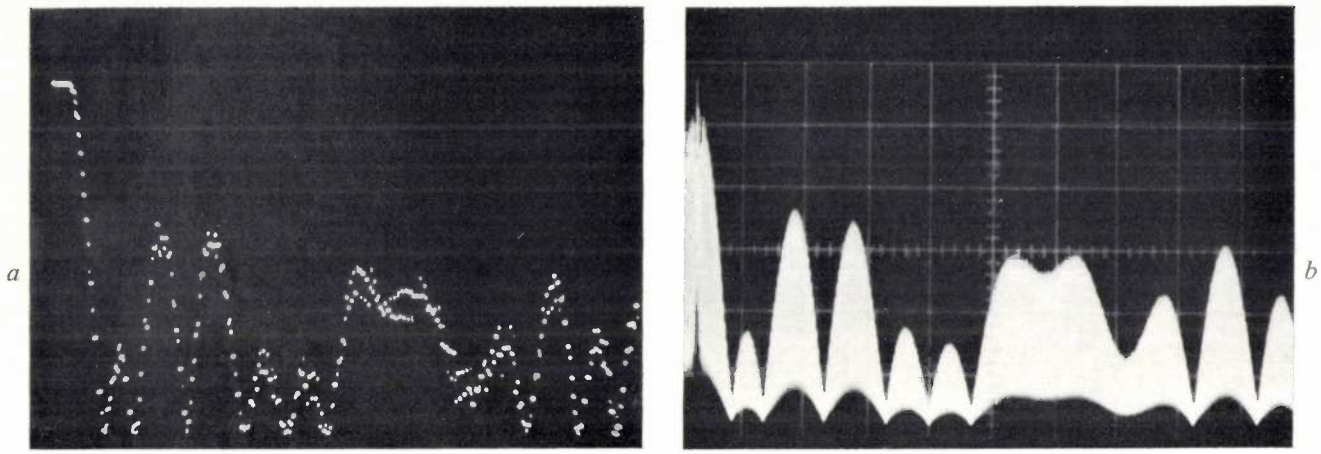


Fig. 20. Comparison between the space-time analogy (*a*) and the space-frequency analogy (*b*) for an unequally spaced system with seven elements. Features that are not very clear in (*a*), e.g. the broad side-lobe, are seen more clearly in (*b*).

Summary. If it is assumed that there is no interaction between the elements of an aerial, the array factors of symmetrical, unequally spaced linear aerial arrays can be found in two ways by means of an electronic analogue system:
1) by using a space-time analogy, in which the array is represented by a frequency spectrum such that the related time function is analogous to the array factor (i.e. a space function). An analogue computer has been made for this simulation;
2) by using a space-frequency analogy. Here the aerial elements are represented by pulse-shaped signals that are generated at times such that the frequency spectrum of this time function is analogous to the array factor.

A description is given of the electronic circuits of the computer based on the space-time analogy, with particular attention to the way in which the cosine functions are generated. The cosine curves are approximated in steps by means of pulse trains. Examples of radiation patterns obtained with this analogue computer are shown, together with some patterns obtained by means of the space-frequency analogy. The two methods each have their own advantages and limitations, and are capable of further extension and improvement. An advantage of the space-time analogy is that this principle can be used to simulate systems in which the direction of radiation is varied by varying the phases of the signals fed to the elements.

Traffic-flow analysis by radar

K. L. Fuller and A. J. Lambell

Since motorway construction is very expensive, and in many countries there is little space for new roads in or round the towns, it is essential that the existing roads should be used in the most efficient way. Now that inexpensive solid-state microwave devices are available, the closed-circuit television equipment at present in use as an aid to traffic control can be supplemented by Doppler radar, which has several advantages. The work described in the article below suggests that the Doppler radar alone may give adequate information about traffic flow.

In 1967 there were over 14 million motor vehicles licensed in the United Kingdom, which represents one vehicle for every 25 metres of public road. The increase in traffic since 1945 is illustrated in *fig. 1*, which shows the number of licensed vehicles in the United Kingdom for the period 1912-1967. The number of vehicles on the roads is doubling every ten years. An official estimate of the number of licensed vehicles in 2010 is 40 million, though this would appear to be a conservative figure. The road building programme is scarcely keeping pace with this increase and motorways are being built that are congested as soon as they are opened. Similar conditions prevail in many countries. Clearly it is essential on economic and social grounds that the existing roads should be used as efficiently as possible. Methods of increasing the traffic flow in situations where further road building is not possible are being used. For example in the West London traffic control scheme 70 sets of traffic lights in an area of 17 km^2 are controlled by computer to give a minimum total journey time [1]. The cost of the equipment for achieving such an increase in the traffic flow can be offset against the cost of the extra road that would be necessary to achieve the same result. As motorways cost £ 600 000 per km or more, it can be seen that a scheme which produces only a small increase in traffic flow may well be economically justifiable.

In order to achieve a better use of the available roads motorists will have to accept more external control over their actions. For example it is quite reasonable to impose speed restrictions on motorways in foggy weather. However, if drivers know that these restrictions are sometimes applied unnecessarily there is a danger that the restrictions will be ignored. It is therefore essential for the police to have adequate information on traffic conditions and weather, and the conversion of this information into directions to the motorist must be arranged to achieve maximum efficiency of traffic flow and maximum safety, while not giving unnecessary warnings and diversions.

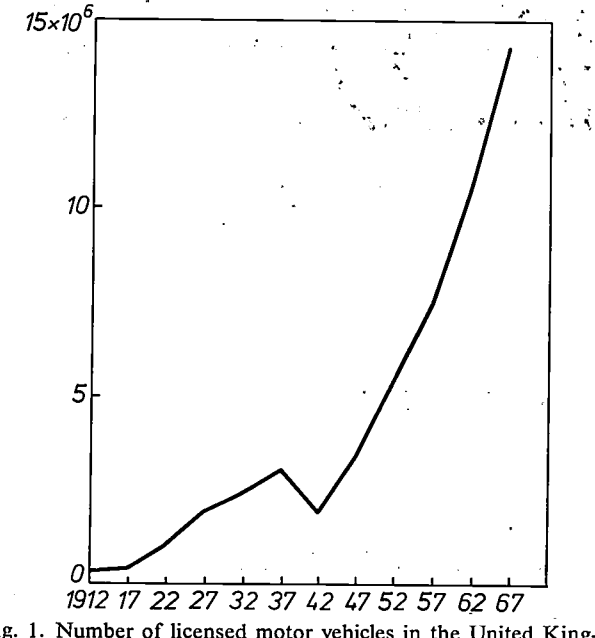


Fig. 1. Number of licensed motor vehicles in the United Kingdom: 1912-1967. This curve indicates the rapid and continuing increase in traffic in the last 20 years.

Radar as compared with closed-circuit television

One method of checking on traffic conditions which is in current use is to survey heavily used sections of motorway with closed-circuit television. The monitor

K. L. Fuller, B.Sc. (Eng.) and A. J. Lambell, M.A., are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] B. M. Cobbe, Traffic control for West London, Electronics and Power **13**, 118-121, 1967.

screens are at a control centre, with the controller watching for accidents or delays, and taking appropriate action. With a long section of motorway the problem arises that with so many screens to watch the controller may not notice an incident quickly enough. There is therefore a need for some device to direct the controller's attention to the appropriate screen or to allow the use of a single switched display. A suggested solution is that each television camera should have associated with it a Doppler radar (similar in principle to the police speed-measuring radar). This radar would produce a signal indicating whether or not there was slow-moving traffic in the section of road under surveillance. This would trigger a device (e.g. a lamp) in the control room to draw the controller's attention to the display.

Consideration of this proposal suggests that it would be possible to dispense with the closed-circuit television and retain the radar. The radar has the advantage that it works in darkness, fog and rain, whereas under these conditions television no longer gives a clear indication of traffic flow. It is at such times, of course, that accidents are most likely to occur and a speedy response is necessary. Moreover the radar measures the important parameter, speed, directly. The proposal to use radar has become realistic because of the availability of low-power solid-state microwave sources and microwave stripline components. A system based on klystron oscillators, with high voltage supplies and limited life, together with relatively expensive or bulky co-axial or waveguide circuits, and requiring many equipments in a single stretch of motorway would not be economic. As only low power is required for the motorway surveillance application, the new devices are ideal for such a system.

To assess the feasibility of the proposed motorway surveillance system an experimental radar with display equipment was built and trials were carried out on various sites (bridges, tunnels, etc.) and particularly on the overhead section of the M4 motorway, which runs westwards from the outskirts of London towards London Airport.

Experimental equipment

A block diagram of the system is shown in *fig. 2*. The solid-state source *S* generates a 10 GHz microwave signal which is formed into a beam by the aerial *Ae* to illuminate the required section of road. It is mounted at a suitable height above the motorway (say 2-7 m) and the beam directed at a point 400 m distant. Because the beam makes a small angle with the road, almost the whole of the carriageway up to 400 m is "illuminated". The signals reflected by objects in the beam are received by the same aerial. The circulator *C* separates the transmitted and received signals. The frequency of the signal reflected by stationary objects is the same as the transmitted frequency, but, owing to the Doppler effect, signals from moving objects are shifted up in frequency if motion is towards the aerial, and down if motion is away. The shift in frequency Δf is related to the transmitted frequency f_0 by the relation $\Delta f = 2vf_0/c$, where v is the velocity of the target (taken as positive for motion towards the aerial, and negative for motion away), and c is the velocity of light. For a vehicle travelling at 60 km/h the Doppler shift is 1100 Hz if $f_0 = 10$ GHz, i.e. an audio frequency.

With a perfect circulator the transmitted signal would all appear at port *1*, and none at port *2*, but in practice a small amount of signal does leak through. This provides a reference frequency to which the frequency of the received signal is compared. The trans-

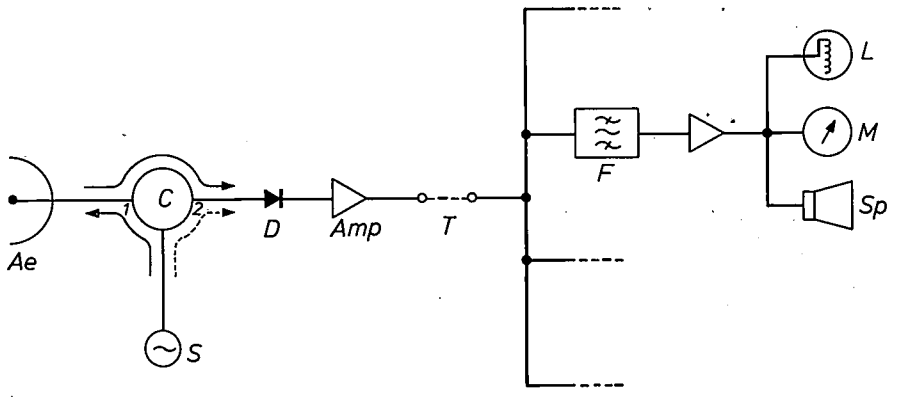


Fig. 2. Simplified block diagram of the Doppler radar for traffic-flow analysis. *S* solid-state microwave source (10 GHz). *C* circulator. *Ae* aerial. *D* detector. *Amp* head amplifier. *T* telephone line. *F* one of four filters. The output of each filter is amplified and displayed by a lamp *L* and a meter *M*; it can also be heard on a loudspeaker *Sp*.

mitted and received signals are mixed by the diode *D*, which gives an output signal of frequency equal to Δf. The information about the direction of motion is lost with this method of detection.

The signals from the detector are amplified and transmitted by telephone line to the control station. The frequencies are filtered into four speed ranges: 5-20 km/h, 20-40 km/h, 40-60 km/h, and over 60 km/h. After further amplification the output from each filter is fed to a loudspeaker *Sp*, a meter *M* and a lamp *L*.

The radar head that we used initially is shown in *fig. 3*. The aerial was a paraboloid reflector of 0.7 m

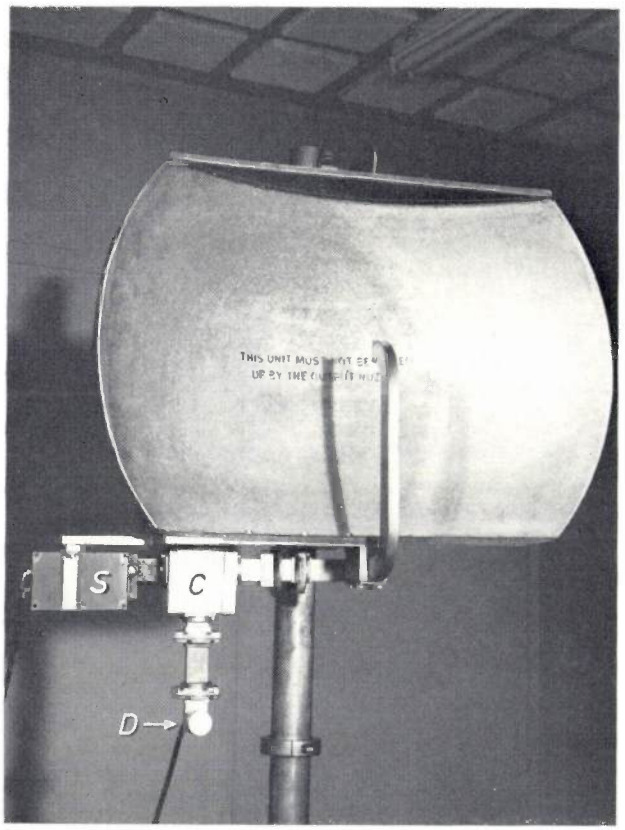


Fig. 3. Radar head comprising solid state source *S*, circulator *C*, detector *D* and an aerial (diameter 70 cm). This is mounted above the road so as to illuminate about 400 metres of carriageway.

diameter with a front feed, which had a beamwidth of about 4°. The solid-state source was a varactor multiplier chain which produced 15 mW of microwave power at a frequency of 10 GHz. At the low frequencies corresponding to slow-moving vehicles conventional point-contact diodes give poor sensitivity as the flicker noise is high. In this equipment a backward diode was used which gives lower flicker noise and a consequent 10 dB increase in sensitivity compared to the conventional diode.

[2] The circuits in the display were constructed by C. D. McEwen and J. F. Oakley of Mullard Research Laboratories.

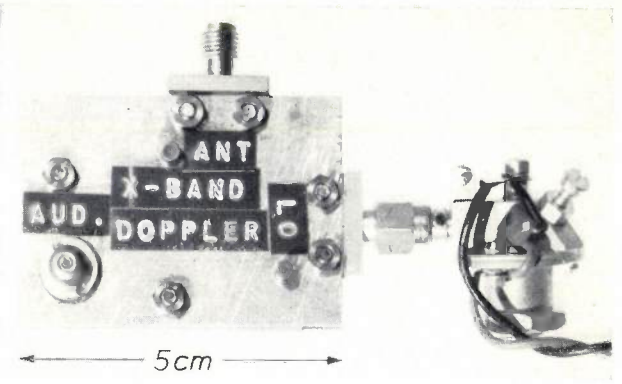


Fig. 4. Miniature radar equipment consisting of a Gunn oscillator, stripline coupler and detector.

A more recently developed head is shown in *fig. 4*, which illustrates the rapid advances that have been made in miniaturizing microwave components. The solid-state source is a Gunn oscillator, which gives almost the same output as the varactor multiplier chain but is much smaller and cheaper. The circulator and detector of fig. 3 have been replaced by a component which combines a 3 dB stripline coupler and detector in one unit.

The display unit [2] is shown in *fig. 5*. Because in this experimental equipment the signals are fed to a loudspeaker, to meters, and to lamps these three methods of presentation could be compared. On trials,

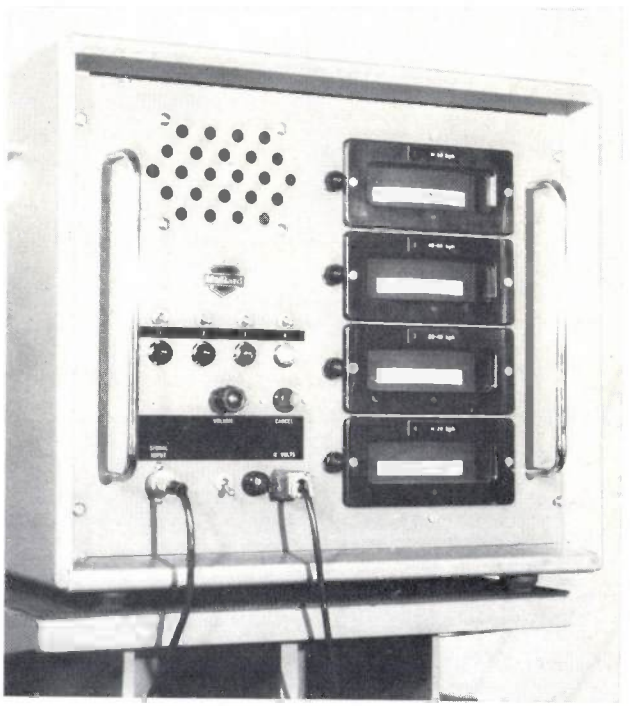


Fig. 5. The display unit, which is placed in the control centre, showing the loudspeaker (above left), meters (on the right) and lamps (lower left) which indicate the traffic speeds.

the loudspeaker is very useful for checking that the equipment is operating correctly, as the ear is a sensitive discriminator of frequency. The lamps are the most useful indication of the presence and speed of traffic. The two lamps representing the upper speed ranges *1* and *2* show green, while the lower speed ranges *3* and *4* show red. The radar cannot detect stationary vehicles, as there is no special feature of the signal reflected from them that distinguishes them from these conditions the ultimate Doppler shift may have a wide range of values which depends on the relative speed of the vehicles and the angle of the surfaces from which reflection has occurred. As any particular configuration of surfaces does not last for long the associated spurious signals are also short-lived. To eliminate these signals the circuits associated with the red lamps were modified so that the lamp only lights when the signal is present for two seconds. This time was

Fig. 6. The radar surveys the three left-hand lanes from the gantry to the slip-road (lower left). The positions of the radar and the television camera are marked *Ae* and *TV*.

reflections from the road. A vehicle that slows down and stops will show up as lamps *3* and *4* light briefly, but when it is stationary there is no further indication. The controller may not see that lamps *3* and *4* have lit, so it was arranged that these lamps should remain on until reset by the operator.

Preliminary trials showed that occasionally the red lamps lit when only fast-moving traffic was present. This occurred when a line of closely spaced vehicles moving with similar speeds was in the beam. The radar signal may then be reflected from one car to another before being finally reflected back to the aerial. Under chosen to be long enough to eliminate the "false alarms" but not so long that a slowing vehicle might be undetected.

Trials

The radar was installed on the overhead section of the M4 motorway to check its effectiveness in indicating traffic-flow conditions. The radar was mounted on an existing gantry about 5 metres above the carriageway. The aerial was positioned so that the beam covered the left-hand three lanes shown in *fig. 6*. The radar could detect vehicles as far as the slip-road which can

be seen on the left-hand side of the photograph. This section of the motorway is surveyed by a television camera mounted on the roof of a tall building by the side of the motorway. The display unit was installed at the motorway control room close to the television monitor screen. The traffic conditions indicated by the radar were compared with the conditions shown on the monitor.

Vehicles slowing down and stopping on the section of motorway surveyed showed up immediately. Moreover, stationary vehicles could also be detected indirectly since they slowed the traffic, causing the lamp representing band *3* to light up. This effect was so striking that the reset facility for the "slow" bands was not necessary. If the controller failed to notice the signal given by a vehicle slowing down and stopping, the presence of the vehicle was still revealed by the changed pattern of traffic flow.

The trial was not complete in the sense that the motorway controller only compared the existing closed-circuit television method with the Doppler radar over a short length of motorway. Ideally more equipments would be used with the controller relying on these alone for information. Nevertheless the limited trials carried out revealed that the equipment was potentially very useful, and achieved the expected results.

When the system was first installed the red lamp corresponding to the slowest speed band was lit almost continuously. This was caused by vibration of the gantry on which the radar was mounted, which produced a low-frequency noise output from the radar head. This spurious signal was eliminated by adding a high-pass filter to the system, which effectively removed any frequency corresponding to 5 km/h or lower. Once this modification had been made, no further "false alarms" were observed during the period of the trial.

A suggested installation

In the experimental system no attempt was made to distinguish the direction of traffic flow. One method of doing this is to use a narrow-beam aerial positioned so as to view one carriageway only. It is probable that this method would not give perfect isolation between carriageways, and slow vehicles in one carriageway might occasionally trigger the warning lamps corresponding to the adjacent carriage way. A better method is to modify the radar head as shown in *fig. 7*. A fraction of the transmitted power is fed to a rotating-field microwave single-sideband modulator. This is a device which changes the frequency of the microwave signal by an amount corresponding to the frequency of an applied electrical signal. For example the frequency of the transmitted signal could be shifted down in frequency by 5 kHz. This signal is now mixed with the reflected signal received by the aerial. Signals from stationary targets give rise to a detected output frequency of 5 kHz. Objects moving towards the radar give signals above 5 kHz, and those moving away give signals below 5 kHz. A filter rejects the signals at 5 kHz, and further filters analyse the other signals into frequency bands of particular speed and direction. An added advantage of this system is that a single broad-beam aerial can be used to cover all carriageways rather than having separate aerials for each. Slightly higher microwave power may then be necessary, but this is a feasible proposition and the cost is more than offset by the decrease in aerial costs.

Experience with the existing equipment shows that only three speed ranges are necessary. Very little traffic on motorways travels at less than 60 km/h. Traffic in the range 30-60 km/h represents either heavily-laden

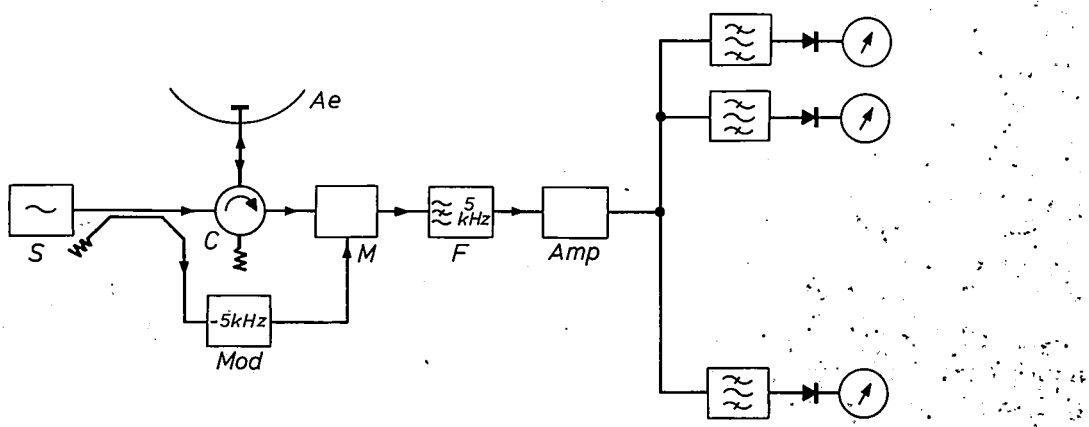


Fig. 7. Block diagram of a Doppler radar which measures speed and direction. Apart from the components shown in the arrangement of fig. 2, the system also includes a single-sideband modulator *Mod*, a balanced mixer *M*, and a filter *F* that rejects 5 kHz signals. The modulator shifts the direct signal to the detector downwards by 5 kHz, i.e. an amount greater than the highest expected Doppler shift. Consequently objects moving towards the radar head give signals above 5 kHz and those moving away give signals below 5 kHz.

vehicles, or cars travelling at reduced speed because of some minor obstruction. Anything travelling at less than 30 km/h almost always indicates a major obstruction. These speed ranges would be indicated by green, yellow and red lights respectively. The lights would be positioned on a mimic diagram of the road with two sets of three lights corresponding to each radar installation (one set for each direction). *Fig. 8* shows how such a mimic diagram appears in different traffic-flow conditions. Fig. 8*a* shows the normal traffic flow with all traffic travelling at more than 60 km/h. In fig. 8*b* there is either a slow-moving vehicle or a minor obstruction at point *4*. The obstruction might be a vehicle pulled in at one side of the road and blocking one lane. The nature of the situation will be clarified in a short time for, if the lamp represents a slow vehicle, the yellow lamps will light up in sequence as it moves along the road. In fig. 8*c* the pattern of lights indicates a major blockage at point *4*. The road is completely obstructed as no traffic appears at point *5*. Traffic is effectively stationary at point *4*, and traffic at *2* and *3* is slowing. With experience a controller should be able to decide quite quickly the nature of any delays and take appropriate action to investigate further and divert traffic where necessary.

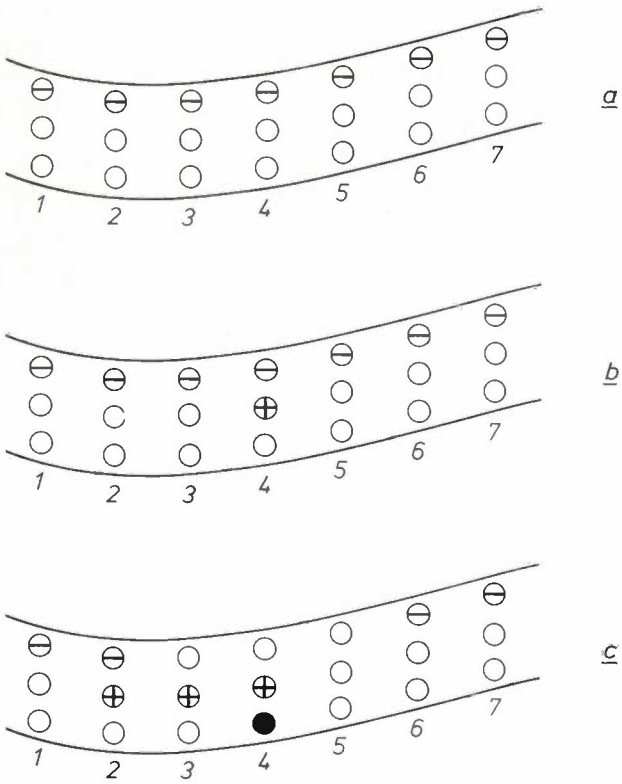


Fig. 8. A mimic diagram of one carriageway with lights to represent the speed of the traffic. This illustrates a possible method of displaying the information. *a*) Normal flow: all traffic moving at more than 60 km/h. *b*) Minor obstruction at *4*; indicated by yellow light. *c*) Major obstruction between *4* and *5*; red and yellow lights indicate a major change in traffic flow, all the lights at *5* are unlit.

The basic Doppler-radar head shown in fig. 4 can be used in many ways. For example, used with a suitable flush-mounted aerial it could be buried in the road to give a point measurement of traffic speed. It can also be used as a very sensitive burglar alarm, and the use of a similar device for the speed control of an automatically driven tractor has been reported. The traffic-control aid described above is just one example of the possible applications of low-power, low-cost radar systems.

The experimental equipment was built under contract to the United Kingdom Ministry of Transport and their permission to publish this article is acknowledged.

Summary. The introduction of solid-state microwave sources and stripline microwave components has made low-power radar systems economic for civilian applications. An experimental solid-state Doppler radar system has been used to investigate the possibilities of using radar to measure traffic speed as an aid to traffic control. A 15 mW solid-state 10 GHz source feeds a 0.7 metre diameter paraboloidal reflector which beams the signal so as to illuminate a 400 metre length of carriageway. The signals reflected from moving vehicles are shifted in frequency by an amount which is proportional to the velocity. These signals are mixed with the transmitted signal and give signals that are separated into frequency bands corresponding to four speed ranges. The signals in the four bands are displayed with the aid of lamps and meters and can also be heard on loudspeakers. The radar has been used on the overhead section of the M4 motorway in West London. Comparison of the speed indications of the radar with the television coverage of the same stretch of motorway shows that the radar gives a reliable indication of obstructions and delays with a negligible number of "false alarms".

"Gems" of lithium niobate

Lithium niobate ($LiNbO_3$) has for some years attracted much attention because of its remarkable ferroelectric and optical properties. It has a high refractive index ($n_o = 2.30$ and $n_e = 2.21$ for sodium light) and a high dispersive power (0.13 for the visible region of the spectrum). With properties such as these, lithium niobate would appear to be of interest as a gem material. Its hardness, however, (Mohs 5) leaves something to be desired. Lithium niobate can be given a variety of colours by adding oxides of the transition metals, in particular Cr_2O_3 (green), Fe_2O_3 (red) and Co_2O_3 (blue), in concentrations of about 0.1 wt. %. In the Philips laboratories at Aachen many single crystals have been made by the Czochralski method, with and without doping [1]. The crystals were pulled from an inductively heated platinum crucible, which had a cross-section of 40 mm and a height of 45 mm. The pulling rate was between 10 and 20 mm/h, and the rate of rotation about 30 rev/min. The crystals obtained in this way were of good quality with diameters up to 25 mm and lengths up to 80 mm (mass up to 120 g). The photograph shows an uncut single crystal of $LiNbO_3$ doped with Fe_2O_3, and a number of gems cut by the firm Gebr. Bank of Idar-Oberstein from variously doped crystals.

[1] See also J. Liebertz, Z. Dtsch. Gemmolog. Ges. 64, 15-16, 1968.

Thermogravimetric analysis applied to ferrites

P. J. L. Reijnen

Thermogravimetric analysis is a relatively simple but highly effective method of studying chemical reactions in which both solids and gases are involved at the same time. The accuracy of the method is high, and in the analysis described in this article as an example it was even possible to obtain information about the defect structure. This information is of great use in the control of sintering processes.

A solid in equilibrium with an ambient gas will take up or give off gas upon a change in the temperature or gas pressure, and its mass therefore changes until a new state of equilibrium is reached. The determination of these changes in mass as a function of the temperature and partial pressure of the gas with which the solid reacts therefore yields information on the solid-gas equilibrium and on the reaction producing that equilibrium. This is the principle of thermogravimetric analysis [1].

We have used this method to investigate ferrites in the temperature range from 1000-1400 °C, with such accuracy that it was even possible to obtain information about the defect structure which the material exhibits in this temperature range.

One indication will be sufficient to make it clear why this information is important. In the last few years great advances have been made in the control of sintering processes. Sintering reactions take place at high temperatures, often higher than 1000 °C. In the material transport that occurs at these temperatures, lattice defects play an important part, particularly vacancies. The proper control of sintering behaviour therefore requires knowledge of the defect structure of the material at the temperatures at which the sintering processes take place.

As an example of how ferrites, and other solids, can be investigated with the aid of a thermobalance, we shall describe in this article a thermogravimetric investigation of the ternary system $MgO\text{-}FeO\text{-}Fe_2O_3$ [2].

The thermobalance

In its simplest form a thermobalance consists of a balance and an open furnace of cylindrical shape, as illustrated schematically in *fig. 1*. The sample holder is attached by a refractory wire to one end of a balance arm, and is suspended in the middle of the furnace. The temperature of the furnace is measured and controlled by means of a thermocouple and a coiled heating element. As the interior of the furnace is in free communication with the outside air, the pressure is constant.

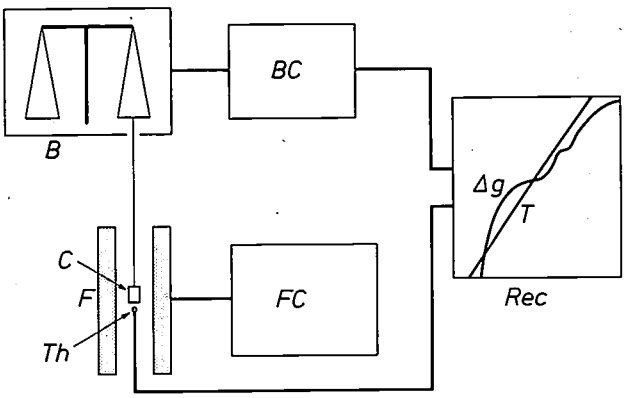


Fig. 1. Diagram of a thermobalance arrangement. *B* balance. *C* sample holder. *F* furnace. *Th* thermocouple. *BC* circuit that keeps the balance automatically in equilibrium. *FC* furnace-temperature stabilizer. The temperature *T* and the change in mass *Δg* are both registered by the recorder *Rec*.

In investigating equilibria, the temperature should be raised in steps and kept constant for a time after each step. In our investigation we raised the temperature in steps of 20 °C and kept it constant for 20 minutes between steps.

During the analysis *apparent* changes in mass also occur as a result of changes in the density of the air in the furnace and of air currents in the furnace. Although these apparent changes in mass are not large, it is often necessary to make a correction for them. The amount of the correction can for example be determined by applying the same thermogravimetric analysis to a "baked-out" sample of Al_2O_3, which should theoretically remain constant in mass.

The accuracy of the thermobalance which we used was 0.1 mg. The samples had a mass of about 5 g.

Dr. P. J. L. Reijnen, formerly with Philips Research Laboratories, Eindhoven, is now with La Radiotechnique Compelec, Evreux, France.

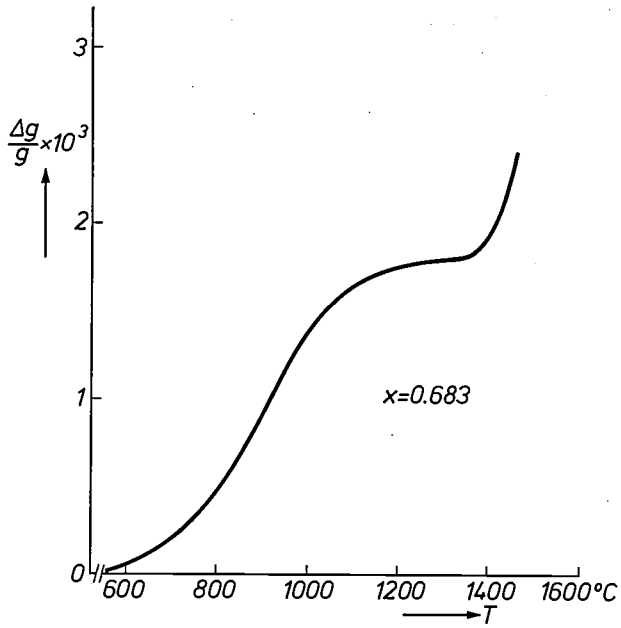


Fig. 2. Example of a thermogram for the system MgO-FeO-Fe_2O_3 with composition $x = 0.683$. The release of oxygen starts at about 550 °C. As the temperature rises, oxygen is initially given off in accordance with the reaction $3\ Fe_2O_3 \rightleftarrows 2\ Fe_3O_4 + \frac{1}{2}\ O_2$ (see also fig. 4). The horizontal part of the curve corresponds to the composition of stoichiometric spinel. With a further increase of temperature the curve rises much more steeply. The oxygen is then given off in accordance with the reaction: $Fe_3O_4 \rightleftarrows 3\ FeO + \frac{1}{2}\ O_2$, during which the wüstite phase is formed.

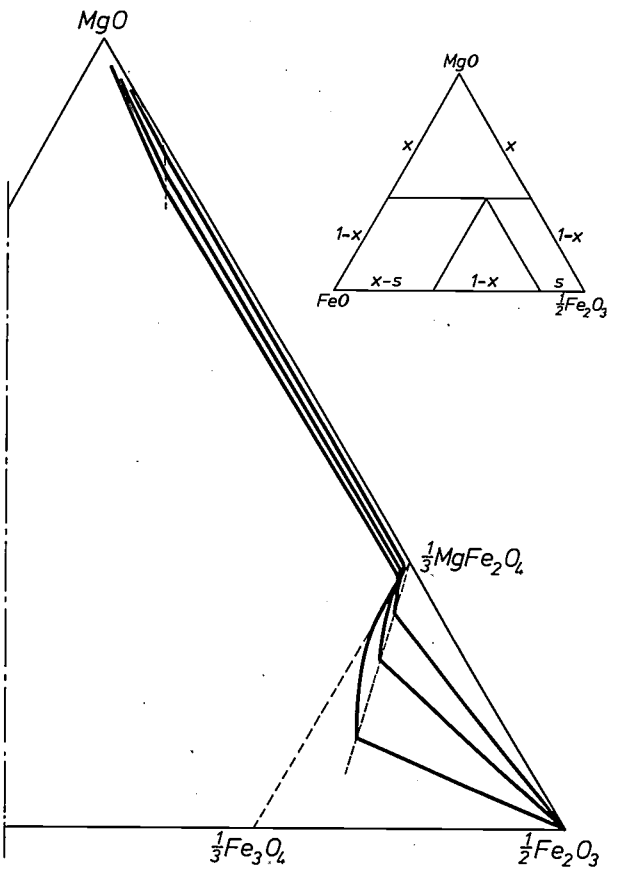


Fig. 3. The result of a thermogravimetric analysis carried out in atmospheric air on MgO-FeO-Fe_2O_3 for different values of the composition parameter x, presented in the form of isotherms in a phase diagram.

For further analysis a more advanced type of thermobalance has been developed, whose accuracy is about 10 times greater. It is also possible to vary the gas pressure in this more accurate instrument.

Determination of the phase diagram

To determine the phase diagram [3] of the system MgO-FeO-Fe_2O_3, samples consisting of MgO and Fe_2O_3 are prepared, each in a different mass ratio. The compounds MgO and Fe_2O_3 do not react with each other at low temperatures, so that their state of equilibrium as required for determining phase diagrams is not reached. For this reason the samples have to be subjected to a pretreatment, which mainly consists in grinding the starting material to a very fine powder and keeping it for some time at about 1000 °C. This treatment also ensures that the starting material is completely oxidized. The samples thus treated are subjected to a stepped increase in temperature in the thermobalance and the resultant decrease in mass is measured. These decreases in mass indicate the amount of oxygen which is released as a result of the reduction of trivalent to divalent iron:

$$Fe^{3+} + \tfrac{1}{2}\ O^{2-} \rightarrow Fe^{2+} + \tfrac{1}{4}\ O_2\uparrow \qquad . . \quad (1)$$

From the measured decrease in mass we can calculate the proportionate amounts of divalent and trivalent iron in the sample. It is usual to express this proportionality in the ratio of the quantity of Fe^{2+} to the total quantity of iron ($Fe^{2+} + Fe^{3+}$). We call this composition variable s; it can assume all values between 0 and 1. Correspondingly, x indicates the ratio of the quantity of iron to the quantity of magnesium and iron together, also expressed in gramme-equivalents. The value of x can be calculated from the weighed quantities of MgO and Fe_2O_3, but since we must make allowance for the fact that the starting materials may have been contaminated with iron during the grinding, x must again be accurately determined by chemical analysis. *Fig. 2* shows an example of a thermogram, obtained from a material with the composition $x = 0.683$.

The information given by the thermogravimetric analysis is collected in a diagram as shown in *fig. 3*. The composition variables x and s are plotted along two sides of an equilateral triangle, as is usual for

[1] A detailed treatment is given in: P. D. Garn, Thermoanalytical methods of investigation, Academic Press, New York 1965, chapters IX and X, and C. Duval, Inorganic thermogravimetric analysis, 2nd edn., Elsevier, Amsterdam 1963, chapters 1, 2 and 3.

[2] See also P. J. L. Reijnen, Philips Res. Repts. **23**, 151, 1968.

[3] The phase diagram of ternary systems is treated by J. L. Meijering in: Phase theory, III. Ternary systems, Philips tech. Rev. **27**, 213-227, 1966.

ternary phase diagrams. The curves in the diagram are isotherms. The diagram applies for a partial oxygen pressure of air of 1 atmosphere.

The shape of the isotherms in the diagram provides valuable indications. It can be seen that the isotherms consist partly of linear sections. From the phase rule it can be shown that *in the regions of composition where the isotherms are straight lines there are two solid phases in equilibrium with each other.*

The phase rule gives the relation between the number of phases P, the number of components C and the number of independent variables F of a system in thermodynamic equilibrium: $F = C - P + 2$. The number of components of the system we are considering is three. If in addition to the gas phase there is only one solid phase in the system, then the number of independent variables is three, if there are two solid phases then there are two independent variables, and so on. When the partial oxygen pressure is kept constant, there is only one independent variable if two solid phases are present. This means that at a given temperature the composition of each of the two phases is fixed. These constant compositions are indicated by the points marking the transition between the straight parts of the isotherms and the curved parts. The straight parts of the isotherms indicate the amounts in which the two phases with their constant composition occur. In other words, they correspond to the *connodes* of two-phase regions. The composition regions in which the curved parts of the isotherms occur are regions with only one solid phase.

By connecting up the points at which the straight parts and the curved parts of the isotherms join we can mark out the boundaries of the regions with a single solid phase (*fig. 4*). We find that there are two composition regions in which two solid phases are in equilibrium with one another, and three composition regions in which a single solid phase occurs. The three separate solid phases are: 1) the wüstite phase *Wü*, which is a solid solution of MgO, FeO and Fe_2O_3, 2) the spinel phase *Sp*, which is a solid solution of $MgFe_2^{III}O_4$, $Fe^{II}Fe_2^{III}O_4$ and Fe_2O_3, and 3) the haematite phase *He*, which consists mainly of Fe_2O_3.

One or two other points should be noted in connection with our interest in the defect structure of the system under investigation. We call the compositions on the line *PQ*, which consist exclusively of mixed crystals of MgO and FeO, *stoichiometric* wüstite. In this stoichiometric composition the ratio of cations to anions is exactly 1. In the composition region outside the line *PQ* we find that Fe_2O_3 is also dissolved in the wüstite phase. This gives rise to cation vacancies, for reasons which will be explained later.

The compositions on the line *RS*, which consist of pure mixed crystals of $MgFe_2^{III}O_4$ and $Fe^{II}Fe_2^{III}O_4$, are called *stoichiometric* spinel. The ratio of the number of cations to the number of anions here is 3:4. In this case also there is a composition region where Fe_2O_3 occurs in solution in the spinel phase, and here again we may assume that this phase is characterized by the presence of cation vacancies. We shall see below, incidentally, that FeO and MgO also dissolve in the spinel phase, though to a much smaller extent, so that the left-hand boundary of the spinel region does not exactly coincide with the line *RS*. In that region anion vacancies will occur.

The defect structure of the spinel phase

We have assumed above that in thermogravimetric analysis oxygen is released or taken up in accordance with the reaction:

$$Fe^{3+} + \tfrac{1}{2}\,O^{2-} \rightleftarrows Fe^{2+} + \tfrac{1}{4}\,O_2. \quad . \; . \quad (1)$$

Whether or not this assumption is correct may be checked as follows. An equilibrium constant K always has the form $A \exp(-\Delta G/RT)$, where ΔG is the Gibbs free energy of the equilibrium reaction, A a proportionality constant, R the gas constant and T the absolute temperature. This means that the logarithm of K must depend linearly on $1/T$. In the following we shall be concerned most particularly with the spinel phase.

Putting the concentration of the oxygen ions equal to 1, we can write for the equilibrium constant:

$$K = \frac{[Fe^{2+}]\,p(O_2)^{\frac{1}{4}}}{[Fe^{3+}]\,[O^{2-}]^{\frac{1}{2}}} = \frac{s}{x-s}\,p(O_2)^{\frac{1}{4}}. \quad . \; . \quad (2)$$

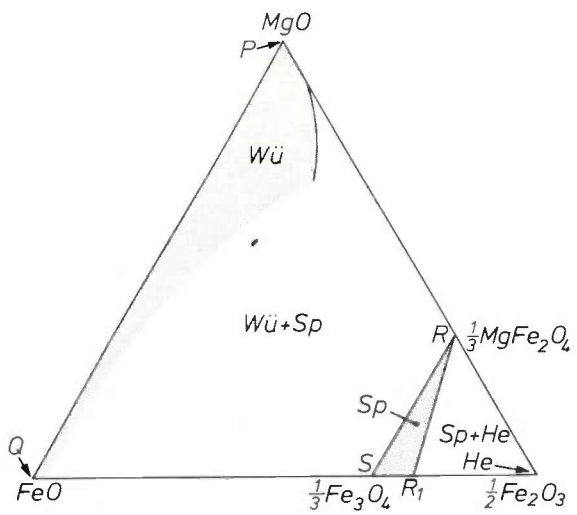


Fig. 4. From the shape of the isotherms in fig. 3 the phase boundaries can be constructed. The straight parts of the isotherms lie in regions with two solid phases, the curved parts in single solid-phase regions. One of these, the haematite phase *He*, is situated at the right-hand corner and has no significant extent. The shaded triangle *Sp* is the composition region of the spinel phase. The line *RS* indicates the compositions of stoichiometric spinel: mixed crystals of $Fe^{II}Fe_2^{III}O_4$ and $MgFe_2^{III}O_4$. The other part of the triangle corresponds to the spinel phase in which Fe_2O_3 also occurs in solution. The region *Wü* at the top of the diagram is the wüstite phase (only partially indicated), where the line *PQ* corresponds to the stoichiometric compositions.

In the analysis that we performed the partial oxygen pressure was constant. In *fig. 5* $\log_{10} K$, calculated from equation (2) with data obtained from a sample with the composition parameter $x = 0.702$, is plotted against $1/T$. As can be seen, the curve is not straight. This indicates that the above equilibrium (1) does not properly describe the reaction.

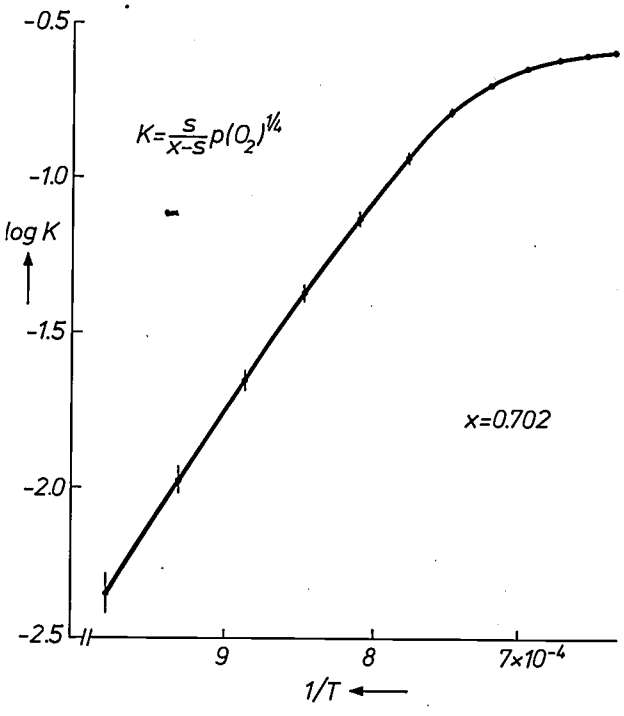


Fig. 5. Logarithm of the equilibrium constant K, calculated from equation (2), plotted as a function of the reciprocal of the absolute temperature T. The relation is not linear, indicating that the reaction equation (1), from which equation (2) was derived, cannot be correct. (In this and some later figures "log" indicates *logarithm to the base 10.*)

The main reason for this is that point defects, such as vacancies and interstitial ions, are concerned in the equilibrium. Vacancies can occur in various ways, for example as a result of the diffusion inwards of a vacant lattice site at the surface (Schottky disorder), or owing to the formation of interstitial ions (Frenkel disorder). In addition we must take account of the fact that the spinel phase contains two sub-lattices for the cations: the A sub-lattice, in which the cations are surrounded by four oxygen ions (tetrahedral sites), and the B sub-lattice, in which the cations are surrounded by eight oxygen ions (octahedral sites). A third of the cations are located on A sites and two-thirds on B sites. The distribution of the various types of cations over the available A and B sites depends on the temperature, and has a contributory effect on the equilibrium between solid and oxygen.

We shall now give some examples of the symbols used for the point defects; complex defects, formed by association of point defects and ions, will be discussed later.

Mg_A^{2+} = Mg^{2+} ion on A site,
Mg_B^{2+} = Mg^{2+} ion on B site,
Mg_i^{2+} = Mg^{2+} ion on interstitial site,
V_A = empty A site (vacancy),
V_A^+ = empty A site which has lost an electron,
V_O = oxygen vacancy,
V_O^- = oxygen vacancy which has gained an electron [4].

Thus, the formation of Schottky vacancies is described by:

$$\text{vacuum} \rightleftarrows V_A + 2\,V_B + 4\,V_O.$$

The formation of interstitial cations proceeds in accordance with the Frenkel equilibria:

$$Mg_A^{2+} \rightleftarrows Mg_i^{2+} + V_A$$

and

$$Fe_B^{3+} \rightleftarrows Fe_i^{3+} + V_B.$$

The distribution of a type of ion over A and B sites is represented by reaction equations such as:

$$Mg_A^{2+} + Fe_B^{3+} \rightleftarrows Fe_A^{3+} + Mg_B^{2+}, \text{ etc.}$$

All these and the following reaction equations satisfy the following conditions:

a) The electrical charge is preserved.
b) The ratio of the number of cation sites to the number of anion sites remains equal to 3 : 4 (this applies particularly to the spinel phase) and the ratio of the number of A sites to B sites remains equal to 1 : 2.
c) The number of atoms and ions of a particular element remains unchanged.

In an ionic lattice, point defects also occur if an ionic compound dissolves with a different cation-anion ratio. For example, if Al_2O_3 is dissolved in $MgAl_2O_4$, then for every twelve O^{2-} ions only eight Al^{3+} ions are added to the spinel lattice. However, for every twelve oxygen sites in the spinel lattice there are nine cation sites available. If we assume that the oxygen is not interstitially dissolved, an assumption which is justified on energy grounds, then for every four Al_2O_3 molecules in the spinel lattice, one cation site will remain vacant. The solution of Al_2O_3 in $MgAl_2O_4$ thus gives rise to cation vacancies. This also affects the concentrations

[4] In semiconductor theory it is usual to call an atomic vacancy an *uncharged vacancy* and an ionic vacancy a *charged vacancy*. A notation is therefore used in which, for example, the symbol for a cation located at its normal lattice site gives no indication of charge. This semiconductor notation is not so suitable for chemical reactions, since it is not compatible with the notation for ionic reactions in aqueous solutions. The notation used in the text is similar to the ionic notation used by Kröger and Vink; see F. A. Kröger, F. H. Stieltjes and H. J. Vink, Philips Res. Repts. **14**, 557, 1959.

of other point defects, since the equilibria between point defects and ions given earlier remain valid.

The number of defects is also affected by oxidation and reduction processes. Thus, the reduction of ferric ions leads in the first place to the formation of oxygen vacancies:

$$Fe_B^{3+} + \tfrac{1}{2}\,O^{2-} \rightleftarrows Fe_B^{2+} + \tfrac{1}{4}\,O_2 + \tfrac{1}{2}\,V_O. \quad . \quad (3)$$

Since we know that Fe_2O_3 is dissolved in the spinel phase, we must therefore also take into account the presence of cation vacancies, on the same grounds as with Al_2O_3. The cation vacancies participate in the Schottky equilibrium:

$$\text{vacuum} \rightleftarrows V_A + 2\,V_B + 4\,V_O,$$

with the equilibrium constant $K_S = [V_A]\,[V_B]^2\,[V_O]^4$.

From this it may be shown that even where the amount of dissolved Fe_2O_3 is low the number of anion vacancies is very small. Moreover, since we know that cation vacancies are mainly found in the B sub-lattice, our obvious course is to combine the above reaction (3) with the following equilibria:

$$\tfrac{2}{8}\,V_B + \tfrac{1}{8}\,V_A + \tfrac{1}{2}\,V_O \rightleftarrows \text{vacuum} \quad . \quad . \quad . \quad . \quad (4)$$

and

$$\tfrac{1}{8}\,V_B + \tfrac{1}{8}\,Fe_A^{3+} \rightleftarrows \tfrac{1}{8}\,Fe_B^{3+} + \tfrac{1}{8}\,V_A. \quad . \quad (5)$$

The result is then:

$$\tfrac{1}{8}\,Fe_A^{3+} + \tfrac{7}{8}\,Fe_B^{3+} + \tfrac{3}{8}\,V_B + \tfrac{1}{2}\,O^{2-} \rightleftarrows Fe_B^{2+} + \tfrac{1}{4}\,O_2. \quad . \quad . \quad . \quad (6)$$

The ions and defects taking part in this equilibrium all occur in appreciable amounts in the spinel considered here. The reaction equation (6) is more suitable for describing the equilibrium between solid and oxygen than reaction equation (1): it describes the take-up or release of oxygen, taking into account the defect structure of the solid.

To calculate the equilibrium constant K for this equation from the observed losses in mass we must also know the equilibrium constants of the distribution equilibria:

$$Fe_A^{3+} + Mg_B^{2+} \overset{K_1}{\rightleftarrows} Fe_B^{3+} + Mg_A^{2+},$$

$$Fe_A^{3+} + Fe_B^{2+} \overset{K_2}{\rightleftarrows} Fe_B^{3+} + Fe_A^{2+},$$

and

$$Fe_A^{3+} + V_B \overset{K_3}{\rightleftarrows} Fe_B^{3+} + V_A.$$

The equilibrium is also affected by any complex defects and charged vacancies and ions that may be present. A quantitative calculation of the equilibrium constant from the experimental data is therefore only possible if we make additional assumptions regarding the other equilibria, that is if we idealize the system.

The special case where the ions and vacancies have no preference for either A or B sites ($K_1 = K_2 = K_3 = 1$), so that in fact the distinction between an A and B sub-lattice ceases to exist, will now be worked out in more detail. The redox equilibrium can be simplified to:

$$Fe^{3+} + \tfrac{3}{8}\,V_k + \tfrac{1}{2}\,O^{2-} \rightleftarrows Fe^{2+} + \tfrac{1}{4}\,O_2, \quad . \quad . \quad (7)$$

where V_k is a vacant cation site.

To calculate the concentrations of the ions taking part in the reaction we take as a starting point a spinel phase with the composition

$$(1-x)\,MgO + (x-s)\,FeO_{1.5} + s\,FeO.$$

We assume that only two types of point defects are present, i.e. cation vacancies V_k and oxygen vacancies V_O, in other words we assume that the concentrations of interstitial cations, charged defects and complex defects are negligibly small. The number of ions of each type per mole amounts to:

$$\begin{aligned} n(Mg^{2+}) &= (1-x)N, \\ n(Fe^{3+}) &= (x-s)N, \\ n(Fe^{2+}) &= sN, \\ n(O^{2-}) &= (1+\tfrac{1}{2}x-\tfrac{1}{2}s)N, \end{aligned}$$

where N is Avogadro's number. Let the number of vacant oxygen sites be δN. The total number of cation sites is $\frac{3}{4}$ of the number of anion sites, and is thus $\frac{3}{4}(1+\frac{1}{2}x-\frac{1}{2}s)N + \frac{3}{4}\delta N$. For the number of vacant cation sites we thus find: $(\frac{3}{8}x-\frac{3}{8}s-\frac{1}{4})N + \frac{3}{4}\delta N$.

We have already noted that when a small quantity of Fe_2O_3 is dissolved in the spinel phase the number of cation vacancies is much greater than the number of anion vacancies. In the cation-deficient region we may therefore write $(\frac{3}{8}x-\frac{3}{8}s-\frac{1}{4})N$ for the number of cation vacancies. The concentrations of the cations are now expressed as follows:

$$[M] = \frac{\text{number of ions M}}{\text{total number of cations}}.$$

The concentration of the oxygen ions is set equal to 1 because nearly all the oxygen sites are occupied. We then obtain:

$$\begin{aligned} K &= \frac{[Fe^{2+}]\,p(O_2)^{\frac{1}{4}}}{[Fe^{3+}]\,[V_k]^{\frac{3}{8}}\,[O^{2-}]^{\frac{1}{2}}} \\ &= \frac{s(\frac{3}{4}+\frac{3}{8}x-\frac{3}{8}s)^{\frac{3}{8}}}{(x-s)\,(\frac{3}{8}x-\frac{3}{8}s-\frac{1}{4})^{\frac{3}{8}}} \times p(O_2)^{\frac{1}{4}} \\ &= \frac{s(6+3x-3s)^{\frac{3}{8}}}{(x-s)\,(3x-3s-2)^{\frac{3}{8}}} \times p(O_2)^{\frac{1}{4}}. \quad . \quad . \quad . \quad (8) \end{aligned}$$

If we now plot $\log_{10} K$ against $1/T$ for the composi-

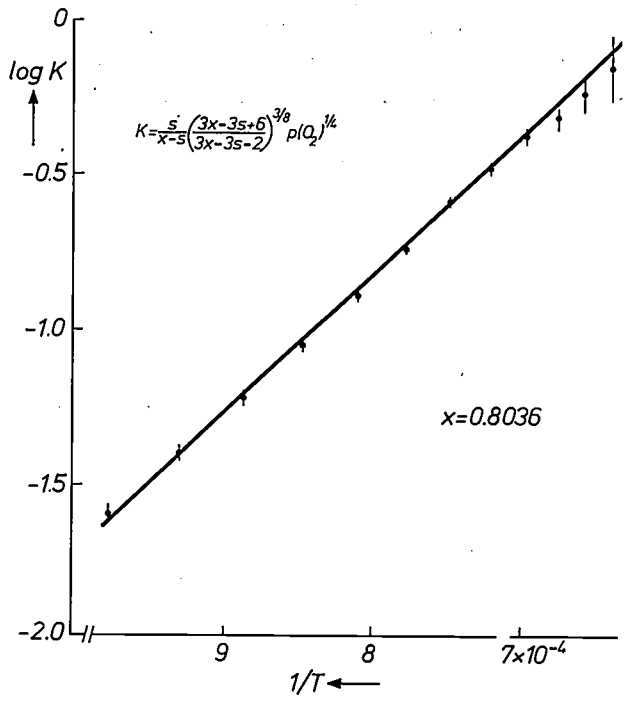


Fig. 6. $\log_{10} K$, calculated from equation (8), plotted against $1/T$. Within the error of measurement the relationship is linear.

tion with $x = 0.8036$, we do in fact find a linear relationship (*fig. 6*).

We have attempted to gain further understanding of the defect structure by testing the correctness of more complicated models in the manner described. Since a cation vacancy is surrounded by four or six oxygen ions, and there is thus a surplus negative charge present at the site of the vacancy, one might assume that the vacancies have a positive charge. Taking this as a starting assumption, and taking the equilibrium $Fe^{3+} + V_k \rightleftarrows V_k^+ + Fe^{2+}$ into account in the calculations, we obtain the following expression for the equilibrium constant:

$$K = \left(\frac{3x + 5s - 2}{5x - 5s + 2}\right)^{\frac{5}{8}} \times \left(\frac{3x - 3s + 6}{3x - 3s - 2}\right)^{\frac{3}{8}} \times p(O_2)^{\frac{1}{4}}. \quad . \; . \; . \quad (9)$$

In this case, however, the relation between $\log_{10} K$ and $1/T$ is *not* linear, from which it follows that our assumption that the vacancies are charged is not correct.

From a similar investigation of the wüstite phase we have been able to establish that the charge compensation in this phase is due to the neighbouring lattice sites of the vacancy being occupied by ferric ions: associates of vacancies and ions, i.e. complex point defects, are therefore formed. Assuming that something similar takes place in the spinel phase, we then obtain for the equilibrium constant:

$$K = \frac{4s}{(x - s + 2)^{\frac{5}{8}} \times (3x - 3s - 2)^{\frac{3}{8}}} \times p(O_2)^{\frac{1}{4}}. \quad . \quad (10)$$

Here too the graph of $\log_{10} K$ against $1/T$ appears to be a straight line within the measurement error.

To distinguish between the model of equation (8) and that of equation (10) we need a thermobalance even more accurate than the balance with which the results used above were obtained. In the meantime we have developed a balance that is ten times more accurate and makes an automatic correction for the apparent changes in mass of the sample [5]. With this thermobalance it should certainly be possible to determine further detailed information about the defect structure.

The new balance is a highly sensitive microtorsion balance. The automatic correction for apparent changes in mass is effected by providing the balance with two identical furnaces, one beneath each balance arm. The total gas pressure can be varied between 10^{-6} bar and 1 bar, and the partial oxygen pressure between 10^{-20} bar and 1 bar. The partial oxygen pressure is measured and regulated with the aid of a ZrO_2 cell [6].

More accurate determination of the phase diagram of the spinel phase

The expressions for K used above are only valid for the composition range in which one solid phase occurs. At the temperature at which a second phase appears the curve of $\log_{10} K$ as a function of $1/T$ therefore shows a discontinuity, as can be seen in *fig. 7*; on

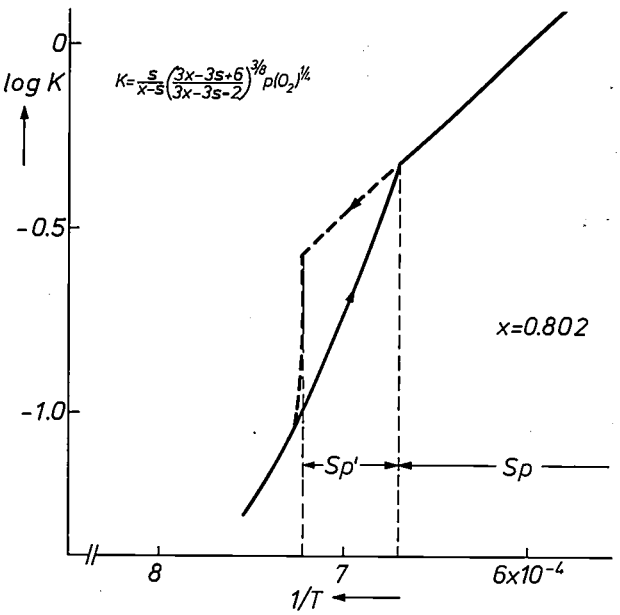


Fig. 7. As fig. 6, but now for a case where a second solid phase appears at a particular temperature. At this temperature the curve shows a discontinuity and is not straight in the two-phase region. On cooling, the discontinuity appears at a considerably lower temperature than on heating up, which indicates that in addition to the stable spinel phase *Sp* there is also a metastable spinel phase *Sp'*.

[5] See P. J. L. Reijnen and P. Roksnoer, A thermo-microbalance with automatic compensation for apparent weight changes and control of oxygen partial pressure, in: Thermal Analysis, Proc. 2nd Int. Conf., Worcester, Mass., 1968, Vol. 1, pp. 289-294.

[6] L. Heijne, Philips tech. Rev., to be published.

cooling this discontinuity is obtained at a lower temperature indicating that there is a metastable spinel region. If we determine a series of curves for various values of x, we can then establish the boundaries of the two-phase region with great accuracy from the location of these discontinuities (*fig. 8*). The boundary of the spinel region on the anion-deficient side cannot be determined with sufficient precision, but it appears from the curvature of the isotherms that FeO and MgO also dissolve in the spinel phase, though only to a limited extent. In the related composition range *anion vacancies* will occur.

As we noted in the introduction, knowledge of the defect structure of the material is of great importance in the control of sintering processes. In these processes, where mass transport takes place by diffusion, there must be a movement of both cations and anions. The rate of reaction is determined by the rate of diffusion of the slowest ion, which is the oxygen ion in the system discussed above. In the presence of anion vacancies oxygen ions can move much faster and therefore the material transport during sintering takes place more quickly. The prediction that this will result in a less porous sintered product has been largely confirmed by experiment [7].

[7] P. J. L. Reijnen, Science of Ceramics 4, 169, 1968, and Reactivity of Solids, Proc. 6th Int. Symp., Schenectady 1968, p. 99.

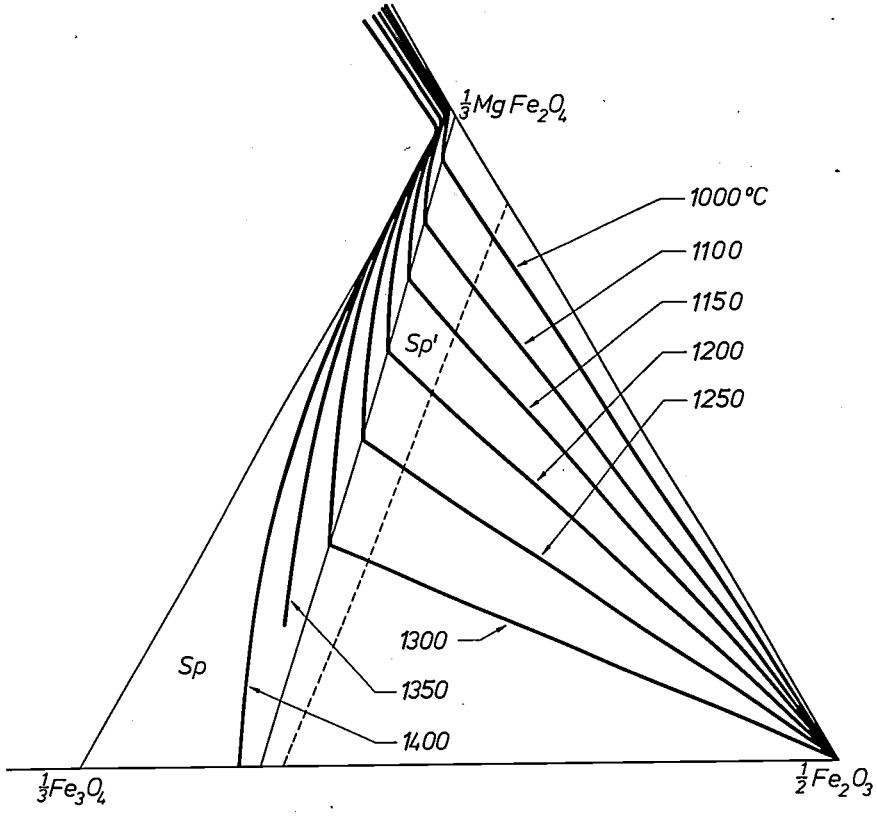


Fig. 8. Detail of the phase diagram of the system MgO-FeO-Fe_2O_3 in the spinel region. The isotherms have been derived from thermograms like those of fig. 2. The boundaries of the stable and metastable spinel phase (regions *Sp* and *Sp'*) have been derived from the discontinuities in curves of the type shown in fig. 7.

Summary. Reactions and equilibria between solids and gases can be investigated with a thermobalance, which is used to determine the release or take-up of the gas as a function of temperature (and pressure). Great accuracy can be achieved with a thermobalance even at high temperature. As an example a thermogravimetric investigation of the system MgO-FeO-Fe_2O_3 is described, carried out between 1000 and 1400 °C at atmospheric pressure on samples of mass about 5 g; the change in mass could be determined with an accuracy of 0.1 mg. After a discussion of the determination of the phase diagram, it is explained how a knowledge of the defect structure of the spinel phase can be obtained by testing models of the equilibrium between the spinel phase and oxygen against experiment. It is shown that the equilibrium equation should include cation vacancies, which are not charged. It is likely that neighbouring ferric ions compensate the negative charge of the oxygen ions surrounding the cation vacancies. In addition a small composition region containing anion vacancies has been brought to light. An improved microtorsion balance has been developed, whose accuracy is ten times better.

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.

D. Andrew: The role of energetic neutrals in a magnetron sputter ion pump.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **2**, 1609-1615, 1969 (No. 11). *M*

P. M. van den Avoort: Traitement des signaux dans le tube index.
Onde électr. **48**, 921-924, 1968 (No. 499). *E*

R. Bleekrode & J. W. van der Laarse: Optical determination of Cs ground-state depletion in Cs-Ar low-pressure discharges, II. Radial and axial Cs-atom distributions.
J. appl. Phys. **40**, 2401-2403, 1969 (No. 6). *E*

P. J. C. Bogers & A. J. G. Op het Veld: A procedure for making electric contact with embedded very thin wires or other very small specimens for electrolytic polishing or etching.
Pract. Metallogr. **6**, 503-504, 1969 (No. 8). *E*

B. Boittiaux, E. Constant, B. Kramer, M. Lefebvre, G. Vaesken (all with Faculté des Sciences de Lille) **& A. Semichon:** Propriétés générales des diodes semiconductrices en régime d'avalanche.
Acta electronica **12**, 157-200, 1969 (No. 2). *L*

P. B. Braun, J. Hornstra & J. I. Leenhouts: The crystal-structure determination of seven $9\beta,10\alpha$-steroids.
Philips Res. Repts. **24**, 427-474, 1969 (No. 5). *E*

J. C. Brice: The effect of kinetics on the stability of crystal interfaces during growth.
J. Crystal Growth **6**, 9-12, 1969 (No. 1). *M*

K. H. J. Buschow & A. S. van der Goot: Phase relations, crystal structures, and magnetic properties of erbium-iron compounds.
Phys. Stat. sol. **35**, 515-522, 1969 (No. 1). *E*

F. J. du Chatenier & J. van den Broek: Electrical properties of vapour-deposited layers of red lead monoxide.
Philips Res. Repts. **24**, 392-406, 1969 (No. 5). *E*

T. **Chisholm:** Some comments on the paper "An opto-electronic cold cathode for cathode-ray tubes".
Solid-State Electronics **12**, 925-926, 1969 (No. 11). *M*

H. Dammann: Computer generated quaternary phase-only holograms.
Physics Letters **29A**, 301-302, 1969 (No. 6). *H*

M. Davio & J. P. Deschamps: Classes of solutions of Boolean equations.
Philips Res. Repts. **24**, 373-378, 1969 (No. 5). *B*

S. Garbe: Factors affecting the photoemission from caesium oxide covered GaAs.
Solid-State Electronics **12**, 893-901, 1969 (No. 11). *A*

Y. Genin & A. Thayse: Analysis of satellite-orbit perturbations due to forces deriving from a potential.
Philips Res. Repts. **24**, 477-558, 1969 (No. 6). *B*

E. E. Havinga, J. H. N. van Vucht & K. H. J. Buschow: Relative stability of various stacking orders in close-packed metal structures.
Philips Res. Repts. **24**, 407-426, 1969 (No. 5). *E*

K. R. Hofmann: Some aspects of Gunn oscillations in thin dielectric-loaded samples.
Electronics Letters **5**, 227-228, 1969 (No. 11). *E*

K. R. Hofmann: Gunn oscillations in thin samples with capacitive surface loading.
Electronics Letters **5**, 289-290, 1969 (No. 13). *E*

M. G. Hulyer: Still picture television.
Roy. Telev. Soc. J. **12**, 170-174, 1969 (No. 8). *M*

Y. Kamp & J. Neirynck: Image parameter theory for noncommensurate transmission lines.
Archiv elektr. Übertr. **23,** 129-136, 1969 (No. 3). *B*

F. M. Klaassen: A computation of the high-frequency noise quantities of a MOS-FET.
Philips Res. Repts. **24,** 559-571, 1969 (No. 6). *E*

J. E. Knowles: A sensitive apparatus for the measurement of induced magneto-crystalline anisotropy.
J. sci. Instr. (J. Physics E), ser. 2, **2,** 917-920, 1969 (No. 11). *M*

W. G. Koster & J. B. Peacock (Institute for Perception Research, Eindhoven): The influence of intensity of visual stimuli on the psychological refractory phase.
Acta psychol. **30,** 232-253, 1969.

J. Liebertz & P. Quadflieg: Einkristallzüchtung und einige physikalische Eigenschaften von Langbeinit ($K_2Mg_2(SO_4)_3$).
J. Crystal Growth **6,** 109-110, 1969 (No. 1). *A*

J. Liebertz & G. Rosenstein: Untersuchungen im System $LiNbO_3$-$MgTiO_3$.
Ber. Dtsch. Keram. Ges. **46,** 548-550, 1969 (No. 10). *A*

G. J. Lubben: Le tube index.
Onde électr. **48,** 918-920, 1968 (No. 499). *E*

R. J. Meijer: The Philips Stirling engine.
Ingenieur **81,** W 69-79, W 81-93, 1969 (Nos. 18, 19). *E*

C. van Opdorp & J. Vrakking: Avalanche breakdown in epitaxial SiC *p-n* junctions.
J. appl. Phys. **40,** 2320-2322, 1969 (No. 5). *E*

J. Robillard: Application de la physique de l'état solide aux systèmes d'affichage d'information.
Rev. fr. Inform. Recherche opér. **2,** No. 15, 53-84, 1968. *L*

P. J. Rommers & J. Visser: Spectrophotometric determination of micro amounts of nitrogen as indophenol.
Analyst **94,** 653-658, 1969 (No. 1121). *E*

W. Schilz: Helicon wave propagation in a two-component solid state plasma.
Phys. Stat. sol. **34,** 213-220, 1969 (No. 1). *H*

M. Valton: Mécanismes physiques de l'avalanche dans les semiconducteurs.
Acta electronica **12,** 131-155, 1969 (No. 2). *L*

M. T. Vlaardingerbroek, P. M. Boers & G. A. Acket: High-frequency conductivity and energy relaxation of hot electrons in GaAs.
Philips Res. Repts. **24,** 379-391, 1969 (No. 5). *E*

J. H. Waszink & J. Polman: Cesium depletion in the positive column of Cs-Ar discharges.
J. appl. Phys. **40,** 2403-2408, 1969 (No. 6). *E*

Contents of Philips Telecommunication Review **28,** No. 4, 1969:

C. J. Krayenbrink: AIRLORD Mk2, pre-departure handling system for airports (pp. 161-173).
H. J. Spoon: The telegraph input-output multiplexer for the DS 714 message switching system (pp. 175-183).
E. A. Limberopoulos: Two new message switching centres for the SITA network (pp. 184-187).

Contents of Mullard Technical Communications **11,** No. 103, 1970:

A. Lindell: "Electronic parts of assessed quality": an outline of the BS9000 scheme (pp. 50-57).
M. C. Gander: Integrated video pre-amplifier, sync, a.g.c. and noise protection system for monochrome receivers (pp. 58-65).
D. J. Beakhust: Decoding circuits with colour-difference output (pp. 66-73).
J. Merrett: Dynamic stability of thyristor single-phase motor speed-control systems for d.c. shunt machines (pp. 75-80).

Contents of Valvo Berichte **15,** No. 3, 1969:

P. G. J. Barten: Theorie des Moiré bei Schattenmasken-Farbbildröhren (pp. 79-92).
C. J. Boers: Hochspannungserzeugung aus dem Zeilenrücklauf-Impuls (pp. 93-108).

Volume 31, 1970, No. 1 pages 1-32 Published 19th Aug. 1970

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No. 2

Photomagnetic effects

U. Enz and R. W. Teale

In recent years some interesting effects have been found, in a number of magnetic substances, that can be considered as a counterpart to the magneto-optical phenomena discovered by Faraday in the 19th century. Until recently research on these photomagnetic effects has been mainly concerned with their basic physics. The article below gives a survey of this work. The applications envisaged for photomagnetics are mainly in the field of information storage.

For some years substances have been known whose magnetic properties can be influenced by infra-red or visible light irradiation. These photomagnetic effects, in which photons bring about changes in the magnetic properties of the material, must be distinguished from magneto-optical effects (such as Faraday rotation), where the material's optical properties depend on its state of magnetization. In the cases investigated so far, the photomagnetic effects have only been found at temperature below ambient. At a sufficiently low temperature the change is permanent but at somewhat higher temperatures the initial state returns after the irradation has stopped.

The first observation of a photomagnetic effect was the discovery that the extra magnetic anisotropy of silicon-doped yttrium iron garnet, caused by cooling in a magnetic field, could be changed by illumination with an incandescent lamp [1] ("photomagnetic anneal"). Soon afterwards we observed associated effects, such as changes of initial permeability and coercive force induced by illumination, in experiments both on yttrium iron garnet [2] [4] and on gallium-doped $CdCr_2Se_4$ [3] [4].

All these phenomena may be explained in terms of an internal photoeffect; an electron bound to a particular lattice site at low temperature is able to move under illumination. This electron movement is equivalent to the displacement of a polyvalent ion. If the structure of the lattice is such that ions of a particular valency influence the magnetic properties of the substance differently depending on their site in the lattice, then the direct consequence of the apparent displacement of the ions as a result of illumination is a change in the value of the magnetic quantities.

We shall first give some observations of the photomagnetic anneal in yttrium iron garnet, and then describe measurements of the effect of illumination on the initial permeability and coercive force, both for yttrium iron garnet and $CdCr_2Se_4$. In both cases the observations will be followed by an explanation given in terms of a model. A comparison of these two models gives an understanding of the conditions under which the effects occur.

In conclusion some possible applications of the photomagnetic effects are mentioned.

Photomagnetic anneal

Yttrium iron garnet (usually abbreviated to YIG), the first substance whose magnetic anisotropy was found to be modified by light [1], owes its name to the correspondence in structure with the semi-precious stone garnet. The pure substance has the formula $Y_3Fe_5O_{12}$ and is known only in synthetic form. The silicon-doped material used for our experiments had the nominal chemical composition $Y_3Fe^{3+}_{5-2x}(Fe^{2+}Si^{4+})_xO_{12}$, where

Dr. U. Enz is with Philips Research Laboratories, Eindhoven. Dr. R. W. Teale is a Senior Lecturer in Physics at the University of Sheffield.

[1] R. W. Teale and D. W. Temple, Phys. Rev. Letters **19**, 904, 1967.
[2] U. Enz and H. van der Heide, Solid State Comm. **6**, 347, 1968.
[3] W. Lems, P. J. Rijnierse, P. F. Bongers and U. Enz, Phys. Rev. Letters **21**, 1643, 1968.
[4] U. Enz, W. Lems, R. Metselaar, P. J. Rijnierse and R. W. Teale, IEEE Trans. MAG-5, 467, 1969 (No. 3).

$0.1 < x < 0.3$ at synthesis. Subsequent analysis of the material yielded lower values. So tetravalent silicon has been substituted for part of the trivalent iron. To compensate the resultant surplus charge an equal number of ferric ions has been reduced to ferrous ions.

The photomagnetic-anneal effect has been observed in single-crystal samples, of the given composition, by magnetic resonance experiments. In a single crystal of magnetic material, the magnetization shows a preference for one or more directions in the crystal structure, and the magnitude of this preference can be expressed in terms of an anisotropy field. In YIG, the preferred directions are the body diagonals of the cubic unit cell, the crystallographic ⟨111⟩ directions. Our experiments were magnetic resonance experiments, but we shall not describe them in detail here. In the experiments the sum of the applied external field and the anisotropy field in the same direction is kept constant, at a known value. Hence the anisotropy field may be derived by measuring the required value of the applied field strength.

A single-crystal sphere of this material with a diameter of 0.4 mm is cooled to 20 °K in a strong magnetic field in the dark. The magnetic field is applied parallel to one of the body diagonals of the cubic unit cell of the substance (the crystallographic [111] direction). If we keep the sample in the dark but now re-apply the external field in the direction of one of the other body diagonals of the unit cell (in this case the $[11\bar{1}]$ direction), we find that a stronger magnetic field is now needed to produce resonance. This means that the anisotropy field, which assists the applied field, is weaker in the $[11\bar{1}]$ than in the [111] direction. This implies that there is now a frozen-in preference for magnetization in the [111] direction, the original direction of the applied field.

In contrast to other materials which can be observed to have a frozen-in state of magnetization, the direction of the frozen-in anisotropy of silicon-doped YIG can be changed by illuminating the sample with an incandescent lamp (*fig. 1*). This is the effect we have called the photomagnetic anneal. As can be seen in the figure, some relaxation of the frozen-in anisotropy occurs in the dark as well. The change in the anisotropy which is caused by the illumination depends on the temperature to which the sample has been cooled. Thus, in fig. 1 the difference ΔH between the final levels of the curves *a* and *b* is 140 Oe for a sample temperature of 20 °K. The same sample gives $\Delta H =$ 200 Oe at 4.2 °K; ΔH only reaches 21 Oe at 66 °K.

The structure of yttrium iron garnet

Before outlining the model used to explain the observed effects, we shall describe the structure of YIG. The undoped material has a cubic structure in which all Fe ions are trivalent. These ions, however, occupy two different types of sites in the lattice: sites surrounded tetrahedrally and sites surrounded octahedrally by oxygen ions. The octahedrally surrounded Fe^{3+} ions form a body-centred cubic structure. A three-dimensional representation of the structure is shown in *fig. 2*.

In the immediate environment of an octahedral Fe^{3+} ion the Y^{3+} ions, which can be distinguished in two groups of three, define a local threefold axis. This local threefold axis coincides with one of the body diagonals of the cubes of the body-centred lattice, the crystallographic ⟨111⟩ directions. Neighbouring Fe^{3+} ions in octahedral coordination always have a different one of the four possible ⟨111⟩ directions as local threefold axis.

If part of the Fe^{3+} in $Y_3Fe_5O_{12}$ is replaced by Si^{4+}, these silicon ions will always occupy sites tetrahedrally surrounded by oxygen.

Explanation of the observations

The observations may be explained as follows. There are four possible types of sites for the Fe^{2+} ion needed for charge compensation, that is to say there

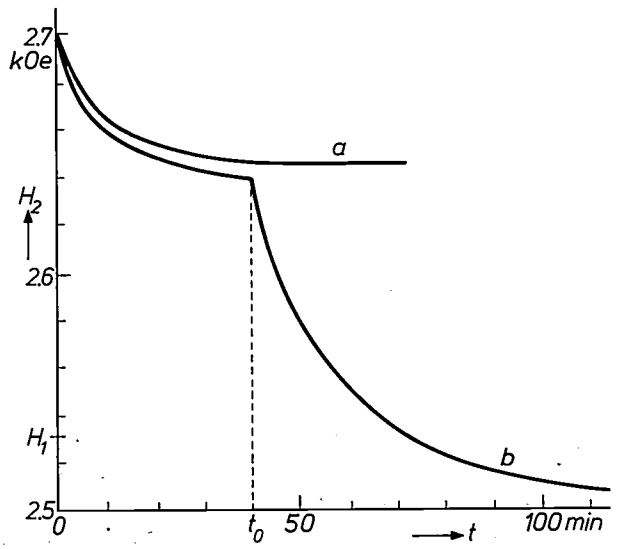


Fig. 1. Photomagnetic anneal of single crystal $Y_3Fe_{4.9}Si_{0.1}O_{12}$. The strength H_2 of the external field needed to keep the ferromagnetic resonance frequency constant with magnetization in the $[11\bar{1}]$ direction gives a measure of the anisotropy. The material is cooled in the dark to 20 °K in an external field in the [111] direction. The magnetic anisotropy then corresponds to the level of applied field indicated as H_1. At the time $t = 0$ the field is rotated to the $[11\bar{1}]$ direction.

Curve *a* relates to a sample that was not illuminated; some relaxation of the anisotropy is observed. Curve *b* relates to a sample that was illuminated with an incandescent lamp from the time $t = t_0$. In this case there is complete relaxation of the anisotropy, i.e. after exposure to light there is an anisotropy field in the $[1\bar{1}1]$ direction of approximately the same magnitude as the original anisotropy field in the [111] direction.

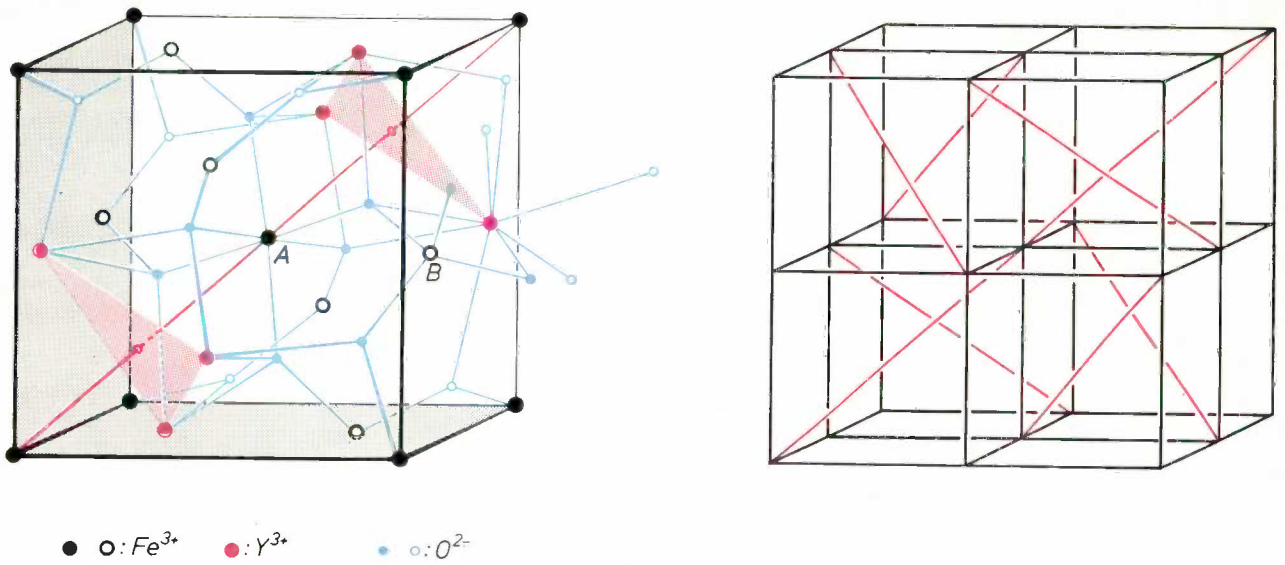


Fig. 2. Crystal structure of $Y_3Fe_5O_{12}$ [5] (yttrium iron garnet or YIG). The right-hand picture gives the complete unit cell; the left-hand picture shows the details of one octant of this cell. The positions of the ions in the complete cell are obtained when the octant given is brought into line with the octants of the complete cell in such a way that the diagonals indicated in red coincide.

The Fe^{3+} ions shown as full dots are surrounded octahedrally by six oxygen ions on octahedral sites and form a body-centred cubic lattice. The Fe^{3+} ions shown as black rings are surrounded by four oxygen ions on tetrahedral sites. Two Fe^{3+} ions, (*A* and *B* in the diagram) are shown surrounded by six and four oxygen ions respectively as full blue dots. The other oxygen ions are indicated by blue rings. The two groups of three Y^{3+} ions that define a local threefold axis at the site of the Fe^{3+} ion *A* are connected by red triangles. The local threefold axis coincides with the diagonal given in red.

are four types of sites where the extra electron supplied by Si^{4+} can be localized. These are the octahedral sites, with the four different directions for the local threefold axis. When an external field is applied, that position whose local threefold axis makes the smallest angle with the direction of the field is energetically somewhat more favourable than the sites with the axis along the other three $\langle 111 \rangle$ directions. The energy difference is small compared with thermal energy at room temperature. However, at 20 °K it is sufficient to cause an observable over-occupation of the energetically more favourable positions, and thus make the anisotropy field stronger in the direction of this particular threefold axis.

At 20 °K the electrons are not sufficiently mobile to allow the occupation of the available sites to adjust spontaneously to the change in the potential-well topography caused by a change in direction of the magnetic field. It is only when energy is supplied by the photons injected upon illumination that the electrons are able to move, as a result of which the electron distribution, and hence the magnetic anisotropy, can adjust to the new direction of magnetization.

The changes are certainly not due to the heating of the crystal by the light source; in our experiments the power radiated into the crystal was too low to cause any appreciable rise in the temperature of the sample.

What we have said above implies that the light is unpolarized. In experiments [6] in which the sample was illuminated with polarized light, it was found that the final state reached did show a small variation, which was dependent on the position of the plane of polarization relative to the crystal axes.

The effect of illumination on initial permeability and coercive force

At much smaller silicon concentrations than those for which the photomagnetic anneal is observable, the illumination can affect the initial permeability and the coercive force of yttrium iron garnet. We observed this effect for garnets having the composition $Y_3Fe_{5-x}Si_xO_{12}$, with $x < 0.05$ [2][4] at synthesis; at this composition the photomagnetic-anneal effect is very weak.

For the measurement of permeability and coercive force, we used small frames cut from single crystals and rings of polycrystalline material made from pressed and sintered powder (*fig. 3*).

[5] Adapted from P. P. Ewald and C. Hermann, Strukturbericht 1913-1926, page 364.

[6] R. W. Teale, D. W. Temple, U. Enz and R. F. Pearson, J. appl. Phys. **40**, 1435, 1969 (No. 3).

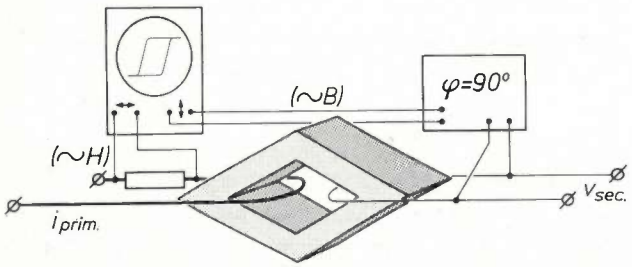


Fig. 3. Measurement of initial permeability and coercive force (schematic). The primary winding carries a sinusoidal current of 10 KHz frequency. The relation between B and H, the hysteresis loop (except for a phase shift equal to the relation between V_{sec} and i_{prim}) is displayed on a cathode-ray oscilloscope. When the magnitude of i_{prim} is small the B-H curve is a line whose slope is equal to the initial permeability; when the amplitude of i_{prim} is sufficiently large, an open hysteresis loop is displayed.
In measurements on single-crystal material diamond-shaped frames are used whose sides coincide with the ⟨111⟩ directions of the crystal; for polycrystalline material small rings are used.

The sample is cooled in the dark to 77 °K and demagnetized by means of a monotonically decreasing alternating magnetic field. When the sample is illuminated with light from an incandescent lamp, the initial permeability decreases. This decrease is larger, and occurs in a much shorter time, in the polycrystalline rings than in the single-crystal frames (*fig. 4*).

By using intermittent light we found that the change in the initial permeability depends on the quantity of absorbed radiation. When the same average power as in the continuous illumination experiments is supplied in a rapid succession of flashes, the decrease of permeability follows exactly the same course as under continuous irradiation. When a sample is illuminated with flashes at such a slow rate that the effect of each individual flash can be seen, a stepwise decrease of permeability is observed (*fig. 5*).

The maximum permeability change $\Delta\mu_r$ that can be brought about by illumination depends on the temperature to which the material has been cooled. At about 120 °K the value of $\Delta\mu_r$ decreases sharply as the temperature increases, and above 200 °K the effect has entirely disappeared.

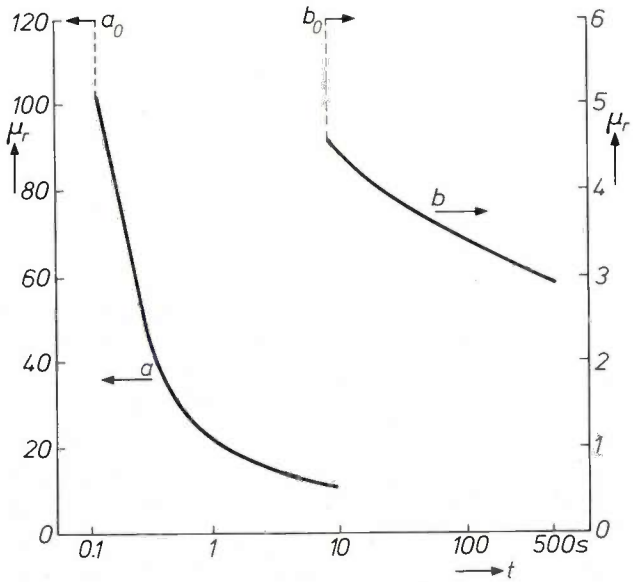


Fig. 4. Variation of initial permeability μ_r of yttrium iron garnet as a function of the time t elapsed since the beginning of the illumination of the sample, at a temperature of 77 °K. Curve a relates to polycrystalline material of composition $Y_3Fe_{5-x}Si_xO_{12}$ with $x = 0.006$, curve b to single-crystal material of composition $Y_3Fe_{5-x}Si_xO_{12}$ with $x = 0.05$.
In order to present the widely different behaviour of both samples in one figure, a logarithmic time scale is used. The permeability values at the moment $t = 0$ are indicated for both samples with horizontal arrows, denoted a_0 and b_0.

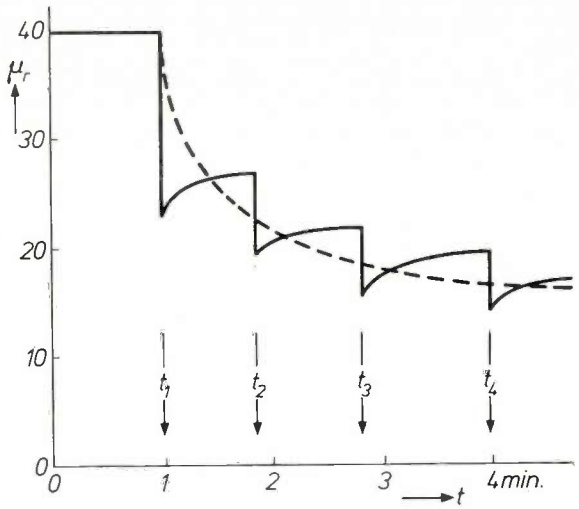


Fig. 5. Decrease of the initial permeability μ_r of single-crystal yttrium iron garnet of composition $Y_3Fe_{5-x}Si_xO_{12}$ ($x = 0.025$) as a result of illumination at 77 °K. The dashed curve relates to continuous irradiation with an intensity of 10^{-2} W/cm^2, the full curve to illumination in flashes (with the same average intensity) at the times t_1, t_2, t_3 and t_4. In the latter case the decrease always takes place within 10^{-2} seconds; the subsequent slight increase following each decrease of the permeability is presumably due to temperature effects. The integrating character of the effect can clearly be seen here.

As long as the temperature of the illuminated sample is kept below 150 °K, the permeability remains low. After the illumination is removed the permeability only returns to its original value when the temperature has risen to a value at which the electrons have become sufficiently mobile. This change in the initial permeability of yttrium iron garnet is only caused by radiation with a wavelength less than 1.3 μm.

The coercive force of yttrium iron garnet is also affected by electromagnetic radiation at low temperature. *Fig. 6* shows there is a marked change in the form of the hysteresis loop when the sample is illuminated. Again, this effect is much stronger in polycrystalline than in single-crystal samples (*Table I*). An interesting

Table I. Coercive force of yttrium iron garnet of the composition $Y_3Fe_{5-x}Si_xO_{12}$ at 77 °K.

	x	unilluminated	illuminated
Single crystal	0.025	0.5 Oe	0.8 Oe
Polycrystalline material	0.006	0.6 Oe	2.0 Oe

demonstration can be given with the set-up shown in fig. 3. The amplitude of the primary alternating current is set at a value such that the maximum primary field strength is only just higher than the coercive force of the unilluminated material. The hysteresis loop of the material is then described once in each current cycle. When the ring is illuminated, the coercive force of the material increases and the primary field cannot reverse the direction of magnetization. The hysteresis loop is then no longer described. Only when the amplitude of the primary alternating field is increased does the hysteresis loop reappear, but in a considerably different shape.

Explanation of the observed effects

Previously in our explanation of the photomagnetic anneal effect, we characterized the octahedral sites in the YIG lattice by the different orientations of the local threefold axes. We now need to make a second distinction: between sites in the immediate environment of an Si^{4+} ion and sites at a greater distance from it. The octahedral sites near an Si^{4+} ion are referred to as type I sites, those at a greater distance are called type II sites. Owing to Coulomb forces between Si^{4+} and Fe^{2+} ions, the Fe^{2+} ions at type I sites have a smaller potential energy than those at the type II sites.

When the material has been cooled, the Fe^{2+} ions will be localized at type I sites. Upon irradiation, electrons, and therefore effective Fe^{2+} ions, can move to type II sites (dissociation of the Si^{4+}-Fe^{2+} pair). Now if the temperature is low enough, the electrons will be relatively immobile and will remain localized at the type II sites. At a somewhat higher temperature a spontaneous return to type I sites will occur. We must now assume that a magnetic domain wall is more strongly bound to Fe^{2+} ions which occupy type II sites than to those in type I sites. This stronger binding results in a greater resistance of the material to changes in magnetization [7], which is reflected in a lower initial permeability and a higher coercive force.

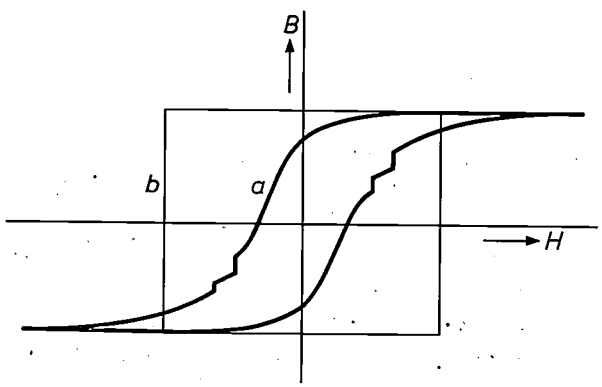


Fig. 6. The hysteresis loop of polycrystalline yttrium iron garnet, of composition $Y_3Fe_{5-x}Si_xO_{12}$, with $x = 0.006$, measured at 77 °K. Loop *a* was measured on unilluminated material, loop *b* on illuminated material. The irregularities in curve *a* are due to the Barkhausen effect, i.e. the discontinuous displacement of domain walls in a continuously changing external field.

Observations on $CdCr_2Se_4$

As well as the photomagnetic changes in silicon-doped YIG, we also found that illumination affects the initial permeability of $CdCr_2Se_4$ in which trivalent gallium is substituted for part of the divalent cadmium [3]. At temperatures below 130 °K this material is ferromagnetic. The samples investigated had the composition $Cd_{1-y}Ga_yCr_2Se_4$ with $y = 0.015$.

At 4.2 °K, illumination causes a virtually permanent change in the initial permeability, whereas at 77 °K the original state returns when the illumination is removed (*fig. 7*). As in the case of YIG, the effect can be explained by assuming two types of site.

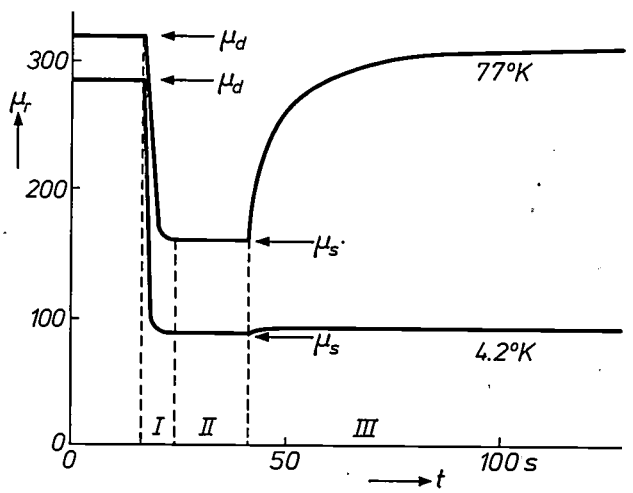


Fig. 7. Change of initial permeability upon the illumination of $Cd_{1-y}Ga_yCr_2Se_4$, with $y = 0.015$, as a function of time, at two temperatures. The permeability values in the unilluminated state (μ_d) and the saturation value of the permeability during illumination (μ_s) are indicated. Three stages are clearly distinguishable: *I*, the permeability decreases at the beginning of illumination; *II*, the permeability reaches a stable final value (μ_s) whose level is determined by the illumination intensity; *III*, the permeability recovers after the illumination has stopped (virtually no recovery at 4.2 °K owing to the very low mobility of the electrons at that temperature).

[7] With this type of material the permeability is mainly governed by the motion of domain walls, which separate homogeneously magnetized domains with different directions of magnetization.

Again illumination converts a type I site into a type II site and hence the domain walls are more strongly bound to their instantaneous position in the crystal lattice.

For a particular material, at a fixed temperature, the value μ_s, to which the initial permeability decreases upon illumination, is solely a function of the intensity of illumination. From the experimentally determined shape of this function and the measured behaviour of μ_r during the recovery following the illumination (phase *III* in fig. 7) we can calculate the time dependence of the permeability at the beginning of the illumination (phase *I*). Good agreement is found between the calculated curve for the change in the magnetic stiffness $1/\mu_r$ and the experimental data (*fig. 8*).

The assumptions underlying the calculation are the following.
1) The change in the magnetic stiffness (the reciprocal of the permeability) as a consequence of the occupation of type II sites caused by illumination is at every instant proportional to the number of occupied type II sites n [8]:

$$\Delta\ \mu^{-1} = Cn\,, \qquad \ldots\ldots\ldots \quad (1)$$

where n is a function of the time t and intensity I of the illumination. (μ is written without subscript in the text below.)
2) Both during and after illumination, recombination of the Cr^{2+}-Ga^{3+} pairs occurs at a rate which is proportional both to the number of occupied type II sites and to the number of vacant type I sites. The number of vacant type I sites is also equal to n, and so the rate of recombination is proportional to n^2.
3) The dissociation rate of the Cr^{2+}-Ga^{3+} pairs (occupation rate of type II sites) is proportional to the illumination intensity I and to the number of type I sites occupied at any given instant. This number amounts to n_0-n, where n_0 is the number of Ga^{3+} ions and hence also the maximum available number of type I sites.

The net rate of occupation of the type II sites, that is, the difference between the dissociation rate and the recombination rate of Cr^{2+}-Ga^{3+} pairs, may be obtained from the results of assumptions 2) and 3). Thus,

$$dn/dt = \beta I(n_0-n)-\alpha n^2\,, \qquad \ldots\ldots \quad (2)$$

where α and β are proportionality constants.

With the boundary condition $n = 0$ for $t = 0$, i.e. all type II sites are vacant before the beginning of illumination, the solution of the differential equation (2) is:

$$n(t) = \frac{2n_0}{1 + z\coth(\frac{1}{2}\beta I z t)}\,, \qquad \ldots\ldots \quad (3)$$

where $z^2 = 1 + 4\alpha n_0/\beta I$. After a sufficient time of illumination, n reaches the stationary value n_s:

$$n_s = 2n_0/(1 + z) \qquad \ldots\ldots\ldots \quad (4)$$

and consequently the change in the magnetic stiffness reaches a stationary value which is a function of the illumination intensity:

$$(\Delta\ \mu^{-1})_s = Cn_s\,. \qquad \ldots\ldots\ldots \quad (5)$$

Equation (4) and (5) give a relation between the saturation value of the change of stiffness $(\Delta\ \mu^{-1})_s$ and the illumination intensity. By fitting this curve to the experimentally determined curves we can obtain numerical values for α/β and C.

The recovery of the magnetic stiffness after the illumination has been switched off at the moment $t = 0$ is found from the differential equation (2) with $I = 0$ and taking as the boundary condition $n = n_s$ for $t = 0$. The solution in this case is:

$$\frac{1}{n(t)} = \frac{1}{n_s} + \alpha t\,, \qquad \ldots\ldots\ldots \quad (6)$$

which, after being divided by C, becomes

$$\frac{1}{\Delta\ \mu^{-1}} = \frac{1}{Cn_s} + \frac{\alpha}{C}\,t.$$

The experiments have confirmed that the reciprocal of the change in magnetic stiffness after illumination has ended is indeed a linear function of time. The slope of the straight line found by experiment yields α/C and hence α, so that all parameters are now established and the time dependence of the magnetic stiffness after switching on the illumination can be calculated.

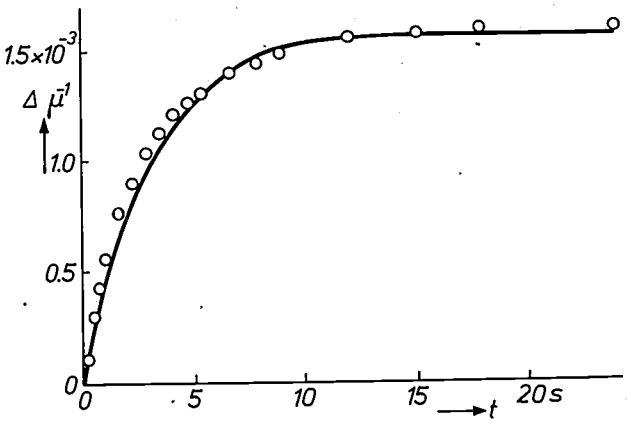


Fig. 8. Change of magnetic stiffness $\Delta\mu^{-1}$ (the change of the reciprocal of permeability) of $Cd_{1-y}Ga_yCr_2Se_4$ ($y = 0.015$), as a function of the time after the beginning of illumination with an incandescent lamp (illumination intensity 9×10^{-5} W/cm², temperature of sample 77 °K). The points give the experimental results. The solid line indicates the expected behaviour as calculated from relation (3).

Compatibility of the two models

We have described two photomagnetic effects and explained them in terms of two different models. The question is, do these two models give rise to contradictions? We can get some insight into this by counting the number of sites available for the ferrous ions in both cases. Such a count also indicates the limits of the dope concentration regions in which the two effects occur.

The formula for yttrium iron garnet can be written as $Y_3Q_2R_3O_{12}$, where Q are the metal ions that are surrounded octahedrally by oxygen and R metal ions surrounded tetrahedrally. For material doped with x atoms of silicon per molecule this results in $Y_3(Fe^{3+}_{2-x}\,Fe^{2+}_x)\,(Fe^{3+}_{3-x}\,Si^{4+}_x)O_{12}$. We see that a fraction $x/3$ of the tetrahedral sites is occupied by silicon, and further that there are 2/3 octahedral sites per

tetrahedral site. The number of octahedral sites l available per Si ion for the associated Fe^{2+} ion is $l = (2/3)\,(3/x) = 2/x$. The photomagnetic anneal has been observed for $0.1 < x < 0.3$, in which case $6.6 < l < 20$. For an Si ion the number of nearest-neighbour sites surrounded octahedrally is four and the number of next-nearest-neighbour octahedral sites is eight. With a maximum of twenty available sites for an Fe^{2+} ion there can therefore be no question of the existence of type II sites at a considerable distance from an Si ion, which we used to explain the permeability effect. It is also plausible that for $x > 0.3$ the photomagnetic anneal will get smaller because in this case many Si ions will have less than four sites available for the Fe^{2+} ion.

For $x < 0.05$—the condition for the occurence of the permeability effect—we have $l > 40$, so that the type II sites necessary for this effect will be numerous. The concentration of Si ions, and hence of Fe^{2+} ions, has now become so small however, that the photomagnetic anneal is no longer observable. Thus it is apparent that the two models do not exclude one another but they are, in fact, complementary.

Unfortunately, there are not sufficient experimental data available to allow us to construct a single model which explains all the observed effects.

At present no practical applications of the photomagnetic effects are known to exist. One difficulty is that the effects have, so far, only been found at very low temperatures. The search for applications can be usefully complemented by a search for materials that show these effects at higher temperatures [9].

Conceivable applications are photomagnetic radiation detectors, whether or not of the integrating type, and optically controlled inductors. Other possibilities are magnetic xerography and the photomagnetic storage of information. For this application a slice of garnet is uniformly magnetized and then made magnetically hard at appropriate places by illumination. An external field is now applied of such orientation and magnitude as to reverse the direction of magnetization of the unexposed parts but not that of the exposed parts, so that the light pattern projected on to the slice is converted into a pattern of different directions of magnetization. In a ferrite-core memory certain cores could be illuminated in order to prevent reversal of their magnetization upon the arrival of the next control pulse.

[8] The basic physics of this relationship has been discussed by W. Lems, R. Metselaar, P. J. Rijnierse and U. Enz in Z. angew. Phys. **29**, 87, 1970 (No. 1) and in J. appl. Phys. **41**, 1248, 1970 (No. 3).

[9] Recently materials have been found that show the phenomena reported here up to temperatures of 200 °K: Th. Holtwijk, W. Lems, A. G. H. Verhulst and U. Enz, IEEE Trans. on Magnetics (to be published).

Summary. The direct influence of electromagnetic radiation on magnetic properties has been detected at low temperatures ($T < 150$ °K) in Si-doped YIG ($Y_3Fe_{5-x}Si_xO_{12}$) and in Ga-doped $CdCr_2Se_4$ ($Cd_{1-y}Ga_yCr_2Se_4$). Its manifestations are 1) a change in magnetocrystalline anisotropy observed in ferromagnetic resonance experiments at high doping levels ($0.1 < x < 0.3$) and 2) a reduction in initial permeability and an increase in coercive force occurring for low doping levels ($x < 0.05$, $y = 0.015$). These changes are produced by irradiation with light of wavelength < 1.3 μm and are attributed to light-induced electron transfer resulting in a redistribution of Fe^{2+} (or Cr^{2+}) ions. At low temperature the redistribution and its manifestations are permanent; at higher temperature there is a relaxation due to thermal electron motion. The magnitude of the effects depends upon the product of light intensity and irradiation time. Quantitative results of a simple two-centre model are shown to agree well with experiment. Potential applications are in the field of information storage and processing.

A vibrating-reed magnetometer for microscopic particles

H. Zijlstra

The search for still better materials for permanent magnets has long been hampered by a succession of problems. A new and highly sensitive magnetometer has now been developed that eases many of the difficulties and opens the way to many new studies. The high sensitivity, achieved by ingenious use of a mechanical resonance, is sufficient to enable the hysteresis loops of individual microscopic particles to be determined with ease. Changes in magnetic moment of less than 5×10^{-18} Wb m have been measured. The results achieved with these new measurements have already had a considerable impact on the theoretical approach to magnetic processes involving only a few domains.

Investigation of microscopic particles

To improve powder materials for permanent magnets it is necessary to determine the hysteresis loops of particles whose dimensions lie between 0.1 and 5 μm. This is the size of the crystallites in the familiar ferroxdure material and in the new magnetic material $SmCo_5$, an intermetallic compound of samarium and cobalt [1]. The magnetic moments [2] of such microscopic particles are about 10^{-16} Wb m. The magnetic processes in particles of this size extend over only a few domains. The main object of the research into these materials is to establish the influence of the crystal structure — with any defects — and of the impurities on the coercivity. The new magnetometer [3] that we shall describe here plays a useful part in this research. Measurements with this instrument have already confirmed that the present idea behind the choice of grain size in magnetic powders — that there is a critical size — is not very relevant and should make way for the idea of critical fields [4]. The results prove, moreover, that theoretical considerations limited to only one domain have little practical significance. We shall return to the critical fields in the last section of this article, in which we discuss a measurement made with the new instrument.

Most magnetometers are inadequate for such investigations. Even for magnetic moments of about 10^{-15} Wb m — which could be typical for particles 100 μm across — the output signal is too small to be usable. The new instrument meets all the requirements imposed by the investigation of magnetic powders. It is about a hundred times more sensitive than other magnetometers, yet is no more susceptible to interfering signals.

Dr. Ir. H. Zijlstra is with Philips Research Laboratories, Eindhoven.

The new magnetometer

The new magnetometer is a simple instrument which is easy to use and works by the method due to Curie (and Faraday). This method is based on the force experienced by a sample particle in a non-uniform magnetic field. This force is equal to $m\,dH/dz$, the product of the magnetic moment m to be determined and the known gradient of the magnetic field strength H. The instrument measures the effect of the force, the displacement of the particle in the z-direction.

We have achieved the high sensitivity by using a periodic displacing force whose frequency is equal to the mechanical resonance frequency of a short wire which carries the sample particle at one end [5]. Instead of a single very small deflection (the static case) there is a periodic deflection which is about a hundred times larger because of the resonance and therefore readily observable. The arrangement is shown in *fig. 1*. The test particle is attached by adhesive [6] to the free end of a thin gold wire (the "reed"), which is horizontal when not in vibration. The other end of the wire is clamped in a supporting block. The wire is made short enough for the resonant frequency to be

[1] K. H. J. Buschow, W. Luiten, P. A. Naastepad and F. F. Westendorp, Philips tech. Rev. 29, 336, 1968.

[2] The SI unit of magnetic moment is the weber m; the electromagnetic unit of magnetic moment (the erg/oersted), which corresponds to $4\pi \times 10^{-10}$ weber m, is still frequently encountered.

[3] See also H. Zijlstra, A vibrating-reed magnetometer for microscopic particles, Rev. sci. Instr. **41**, 1241-1243, 1970 (No. 8).

[4] H. Zijlstra, Coercivity and wall motion, paper presented at Second European Conference on Hard Magnetic Materials, Milan, September 1969, in IEEE Trans. MAG-6, 179-181, 1970 (No. 2).

[5] The method of measuring magnetic susceptibility proposed by Y. L. Yousef, H. Mikhail and R. K. Girgis, Rev. sci. Instr. **22**, 342, 1951, is also based on resonance enhancement.

[6] Details of the use of adhesive are given in [3].

above the frequency of interfering mechanical vibration from the surroundings. The position of the wire can be adjusted with a micrometer screw. Bending of the wire under its own weight can be neglected. Gold is a suitable material for the wire: it can easily be drawn thin and since it is only weakly diamagnetic, it is not subject to magnetic interference. The gradient coils C_1 and C_2 — shown in the figure — each have 350 turns, with an outside diameter of 6 mm and an inside diameter of 3 mm. These coils are connected in series opposition. The magnetic field of these coils is therefore zero at the centre of the 2 mm air gap, where the sample is located, and at this point the vertical field gradient dH/dz has a finite magnitude. The coils are fed by an oscillator, tuned to the mechanical resonance frequency of the gold wire (about 35 Hz). The wire and the gradient coils are mounted between the poles of an electromagnet K. This gives a uniform vertical magnetic field, which can be varied to change the magnetic moment of the sample.

The magnetic hysteresis effects to be investigated are produced by varying the field strength of K cyclically. A measurement is then made of the way in which the amplitude of the vibration of the gold wire depends on the field strength. The vibrating wire is illuminated stroboscopically by flashes that are synchronized with the oscillator and can be observed from the side through a microscope. The phase is adjusted so that the lamp flashes each time the wire reaches the maximum deflection on one side. The calibrated eyepiece of the microscope then gives a stationary and sharp picture of the free end of the wire,

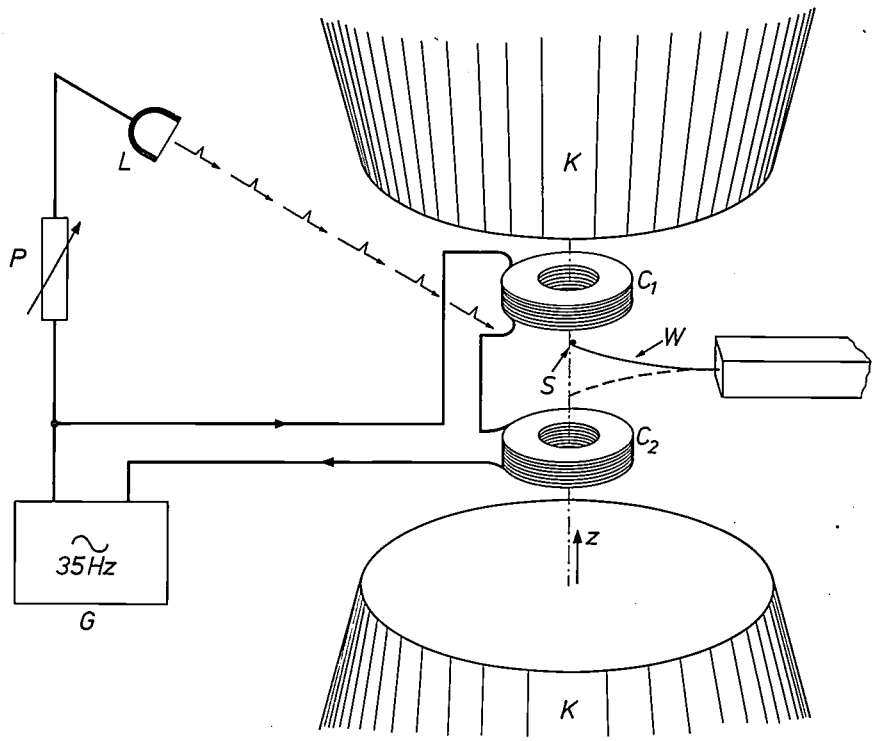


Fig. 1. A magnetometer based on the Curie (or Faraday) method, with improved sensitivity. The sample particle S is fixed to the free end of the gold wire W (length 20 mm, diameter 38 μm), which vibrates in the vertical direction (the z-axis). The amplitude of the vibration is determined with a calibrated microscope (not shown). C_1, C_2 gradient coils. K poles of variable electromagnet. L stroboscope. P phase-shifter. G oscillator, tuned to the resonant frequency of W.

whose deflection from its stationary position is the amplitude to be measured (*fig. 2*). The stroboscopic illumination also shows whether the magnetic moment being measured is directed upwards or downwards. In the first case for example the wire could appear above the stationary position and would therefore be below it in the second case.

In practice it is preferable to observe the microscope image via a television camera rather than by eye. The wire then appears magnified about 1000× on the monitor screen. Amplitudes of a few μm can be measured in this way.

We shall now add a few comments about finding the best values for the quantities that determine the vi-

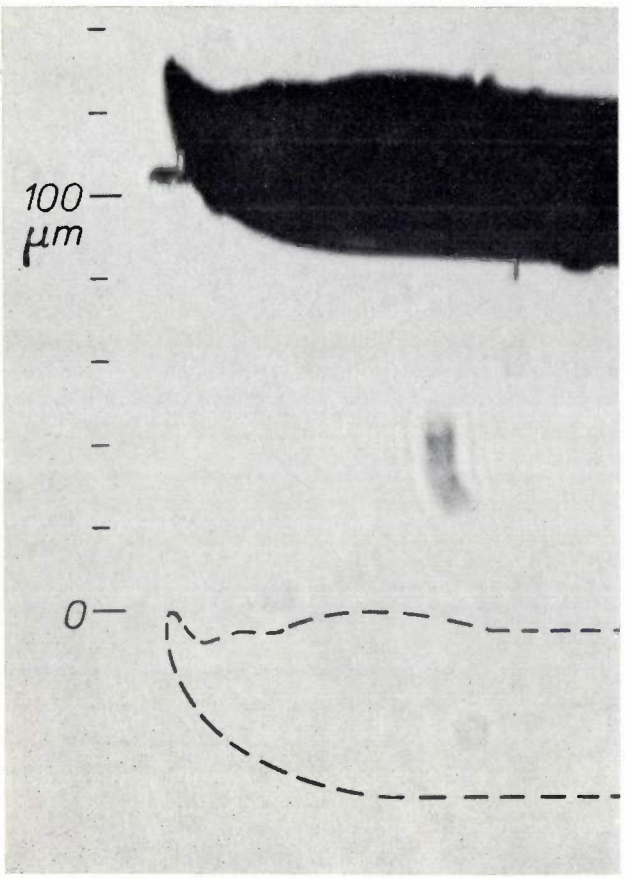


Fig. 2. Displacement of the vibrating wire due to a magnetic moment of about 10^{-16} Wb m, arising from a particle of $SmCo_5$ of mass about 10^{-9} gramme; a stroboscopic image was observed via a calibrated microscope. The sharp projection at the end of the wire is the drop of adhesive containing the test particle, which is itself invisible. The dashed line indicates the equilibrium position. The scale markings in the figure correspond to those of the calibrated eyepiece, which was removed before making the photograph.

bration of the wire. This will also give some idea of the theoretical sensitivity.

The wire is made to vibrate (see *fig. 3*) by the periodic vertical force $F \sin 2\pi f_r t$, where F is the peak value of the force and f_r the frequency. The deflection of the wire, i.e. the amplitude a to be measured, is given by:

$$a = a_s 2\pi E/E_l.$$

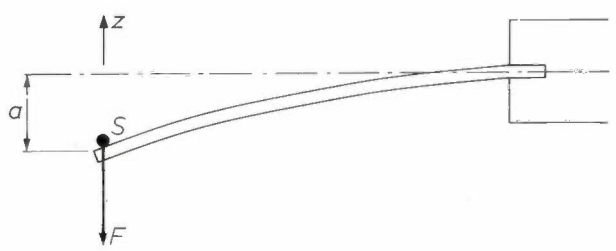


Fig. 3. The vibrating wire clamped at one end. The magnetic sample S at the free end is subjected to a periodic force, of amplitude F, in the z-direction. The amplitude a of the periodic deflection is greatest when the frequency of the force is equal to the mechanical resonance frequency of the vibrating wire.

In this expression a_s represents the static deflection that would be obtained if the force were constant and equal to F; E is the total vibrational energy (kinetic + potential energy of the vibrating system) and E_l the energy loss due to damping in one period. The improvement a/a_s in the sensitivity is equal to $2\pi E/E_l$, the Q (or quality) factor. Thus, when E_l decreases, the sensitivity increases. If the practical difficulties could be overcome, it would be preferable to mount the system in a vacuum to minimize air damping. Further study [3] of the Q factor leads to the approximate equation:

$$\frac{m}{a} = C\eta \frac{d}{l}\left(\frac{\gamma}{\varrho}\right)^{\frac{1}{2}}\left(\frac{dH}{dz}\right)^{-1}.$$

The left-hand side is the ratio of the unknown magnetic moment m of the sample and the amplitude a being measured. The constant C on the right-hand side is approximately equal to 1/2. It can be seen how the viscosity η of the ambient gas affects the ratio m/a: η should be as small as possible. Small values should also be chosen for the ratio of diameter d and length l of the wire and for its modulus of elasticity γ; the density ϱ of the material, however, and the field gradient dH/dz should have high values. Insertion of the selected values in the equation [3] shows that an amplitude of 1 μm indicates a magnetic moment of about 10^{-17} Wb m. It should be appreciated that this result is no more than an approximation. A more accurate prediction of the behaviour of the vibrating wire can only be obtained from the complete solution of the appropriate non-linear differential equation.

Measurements for an $SmCo_5$ particle (10^{-9} gramme)

An example of the magnetic behaviour of an individual $SmCo_5$ particle measuring about 5 μm and weighing about 10^{-9} gramme is given in *fig. 4*. The saturation value of the magnetic moment of this particle is about 10^{-16} Wb m. The magnetic field strength produced at the location of the particle by the electromagnet K (see fig. 1) is plotted horizontally; this field strength determines the measured magnetic moment. The arrows indicate the sequence of changes in the principal field.

To clarify the physical background to the results of these measurements it will be useful to explain the idea of critical fields [4], which we referred to at the beginning of this article. In this concept, the coercivity of a particle is closely related to three forces deriving from its structure, and not to its physical dimensions. The three forces are a nucleation field for creating a domain wall, a field derived from a pinning potential, which pins the wall at certain positions, and a

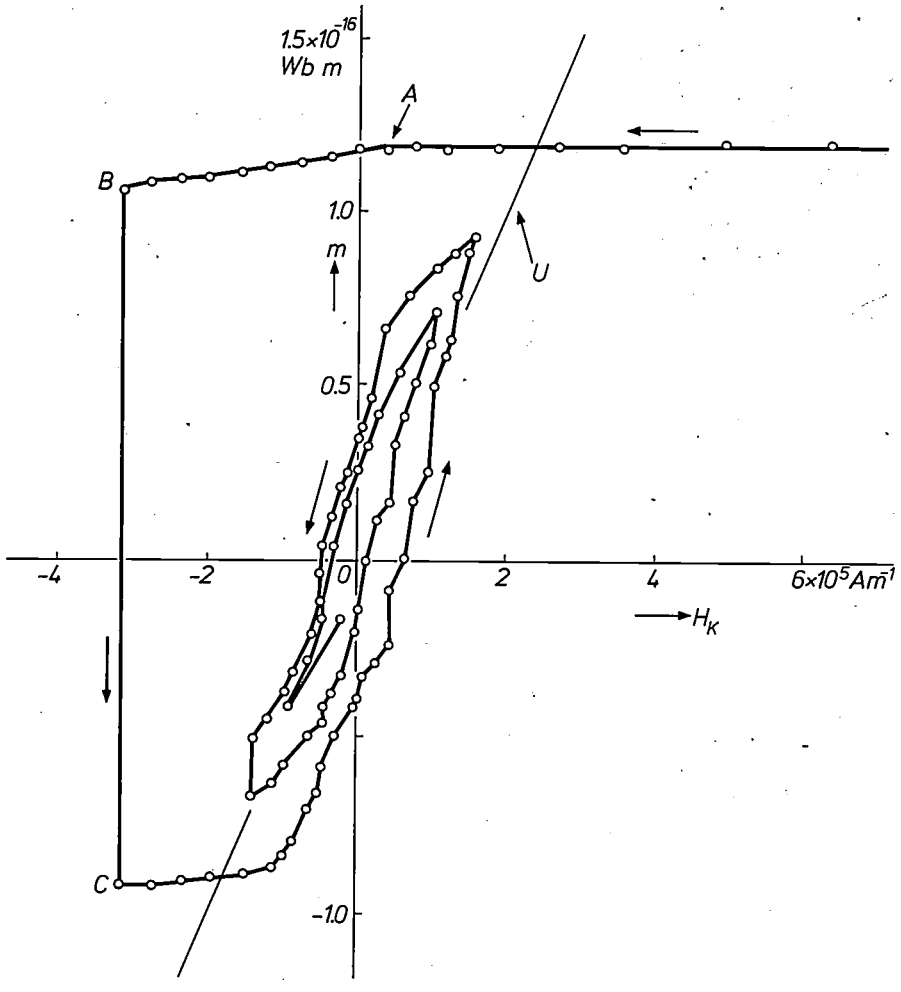


Fig. 4. Example of the magnetic behaviour of an $SmCo_5$ particle, about 5 μm across and with a mass of about 10^{-9} gramme. The measured magnetic moment m is plotted vertically, and the magnetic field strength H_K produced at the location of the particle by the electromagnet (K in fig. 1) is plotted horizontally; this field is varied in the direction indicated by the arrows. The lower limit of the detectable change in magnetic moment is about 4×10^{-18} Wb m. The points A, B, C and the line U are explained in the text.

propagation field that causes a domain wall to move. A high coercivity is made possible by either a strong nucleation field or a strong pinning potential. Which of these two predominates in a particular case can be found from measurements on individual particles.

In the example given in fig. 4 A indicates the creation of a domain wall, which at first remains pinned. At B the wall breaks away and drifts to the other side of the particle. At C the wall becomes pinned again. The displacement from B to C evidently takes place without much difficulty: the change required in the field produced by the magnet K is small. The propagation field is therefore practically zero. We conclude from these results that the coercivity of this particle is mainly determined by the strong pinning potential that occurs near the surface.

Inside the particle a domain wall can move fairly easily. The line U in fig. 4, which represents the behaviour of a ferromagnetic sphere with freely movable walls, provides evidence of this, since the slope of the line agrees fairly well with the slope of the minor hysteresis loop shown for the particle.

The small vertical "jumps" on the minor loop are due to the Barkhausen effect, in which the wall starts to move and then almost immediately stops again. These small changes of magnetic moment, a few times 10^{-18} Wb m, can be reproducibly demonstrated with the new magnetometer. In practice the sensitivity of the instrument therefore exceeds the value derived from the approximate equation.

Summary. A description is given of the construction, optimization and performance of a vibrating-reed magnetometer based on the Curie (or Faraday) method, but using a periodic displacement force. A sample particle, fastened by adhesive to the free end of a gold wire (length 20 mm, diameter 38 μm), is located between two gradient coils in series opposition. The frequency of the field gradient is equal to the resonant frequency of the vibrating wire. The amplitude of the deflection of the wire is measured with a calibrated microscope; a stationary image is obtained by stroboscopic illumination. The magnitude and direction of the magnetic moment are found from the measurement. There is no perceptible interfering vibration from the surroundings. The smallest measured change of magnetic moment, for a hysteresis loop of an $SmCo_5$ particle (mass about 10^{-9} gramme), is of the order of 10^{-18} Wb m. The measurements indicate which of the critical fields determine the coercivity.

Renaissance in ceramic technology

A. L. Stuijts

The article below is almost identical with the text of the address given by Prof. Stuijts upon his inauguration as Visiting Professor at the Technical University of Eindhoven. Since about 1950 many new methods have been employed in ceramic technology with considerable success, and these have led to better control of the microstructure of ceramic materials. Nowadays it is no longer so necessary as it once was to adapt practical applications to available materials: the material can now be "tailored" to suit the application. The notable features of the new methods include special additives, the use of very fine powders with high sintering reactivity, materials with a slight deviation from the stoichiometric composition to allow the introduction of certain desired lattice vacancies, and the simultaneous application of high pressure and high temperature. These new developments in ceramic technology are of such significance that we are planning to follow Prof. Stuijts's introduction to the subject with a second article in the near future.

Ceramic technology, past and present

Solids have played such an important part in the evolution of man that whole ages of man's history have been named after them: the Stone Age, the Bronze Age and the Iron Age. One might wonder what material our present age will be named after. Some have suggested that we might well be on the threshold of the Glass Age.

The availability of materials has in many cases determined the outward forms of a civilization. This applies equally to the important technical advances of the last century. The growth of the railways and the later motor traffic depended to a very great extent on the production of iron; this is clearly reflected in the French and German words for railways, "chemins de fer" and "Eisenbahnen". The development of the aircraft industry only really got into its stride when aluminium alloys became available. It was not until the advent of the tungsten filament that the incandescent lamp really came into general use as a light source.

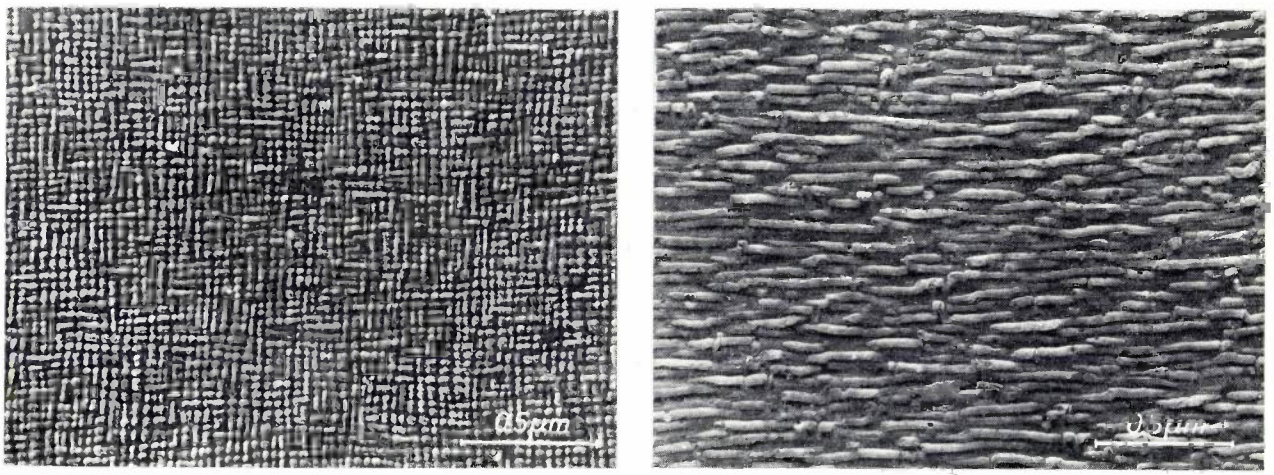
The examples mentioned make it clear that metals have had a dominant influence on the major technical developments. In about 1950 a new age in the science of materials was born, in which many new materials were developed by applying scientific principles. Until about 1950 alloys with special properties, such as soft and hard magnetic alloys, refractory and hard-wearing materials, were almost solely the result of empirical research. The manufacturer of tungsten filaments for incandescent lamps in about 1950 did of course possess a high degree of craftsmanship, but for that matter the bell-founder in ancient China was also a reasonably good craftsman. Generalizing, I think I may say that scientifically there is not so very much difference. You can see this, for example, from the doping of the tungsten for the purpose of preventing wire fracture due to sagging or shearing of the crystals. This was discovered entirely by chance [1], and wire fracture is controlled in almost as mystic a way as the beautiful tone of the bell produced by the Chinese bell-founder: in the poem which apparently served as a manufacturing recipe in old China, the sweet tones of the bell were called forth by the tragic sacrifice of the bell-founder's beautiful daughter.

Nevertheless, in the period before 1950 there was a great deal of scientific work in metallurgy which has provided the basis for important recent developments of materials. It might perhaps be said that practical metallurgy had reached too advanced a stage of technical development for it to be successfully explored by fundamental physical research. It was only a few years ago, for instance, that it became possible to understand the magnetic structure of a complex alloy like "Ticonal" (*fig. 1*).

In the development of modern materials an essential part has also been played by solid-state physics. The major advances here were made after the Second World War, and they have led to developments that include nuclear fuels, materials with special ferromagnetic and ferroelectric properties, and important

Prof. Ir. A. L. Stuijts is with Philips Research Laboratories, Eindhoven, and is also Visiting Professor in Technology of Inorganic Materials at the Technical University of Eindhoven.

semiconducting materials [2]. The most widely known amongst these is the application of the semiconducting materials in the manufacture of transistors. The development of solid-state physics has come about through close cooperation between physicists and chemists, who, with a better picture of the atom, were then able to understand many of the intrinsic properties of solids.



a *b*

Fig. 1. The submicroscopic structure of a material can have a great influence on its physical properties. The above photographs show the orientation of the crystallites of which the alloy "Ticonal" consists. During their precipitation the small needles were oriented in a magnetic field. Their beautifully regular arrangement gives a high value of $(BH)_{max}$ (11×10^6 gauss oersted). *a*) The material as seen in the longitudinal direction of the needles; *b*) in a plane parallel to the needles.

The potential properties of a material are determined by the chemical composition and crystal structure of the material. In other words, whether a material is semiconducting, magnetic, stable at high temperatures or transparent, depends in the first place on the chemical composition and crystal structure of the metal, alloy or compound.

Even in the early twenties it had become clear to metallurgists that the really important technical properties of a material depend on the state of aggregation and the distribution of the phases from which the material is built up. *Microstructure* is the term often used for this. The study of the microstructure of metals and alloys, and of its relation to their composition and their physical and mechanical properties is called metallography.

This conception that the important technical properties of materials depend to a very great extent on the microstructure is of general validity for modern materials. In the case of composites it seems almost a tautology.

"Tailoring" materials to meet practical requirements

Starting from this idea it becomes possible in principle to synthesize materials that meet the precise requirements of the application, provided that two conditions are satisfied. First of all, it is necessary to be able to formulate quantitatively the relationship between the desired properties and the microstructure required to achieve them. Secondly, the desired structure must in fact be reproducible and economically obtainable.

Solid-state physics, which studies the relationship between microstructure and material properties, and physical metallurgy, which is concerned with the origin of microstructures, have together laid the basis for the advances that have been made in the modern technology of materials. The most surprising aspect of this has been that research in these fields has in many cases shown sufficient basic knowledge to be available to allow certain specific requirements to be met. The ability to supply materials "made to measure" obviously represents an important step forward. For the user, the electronic engineer, the mechanical engineer, the aircraft designer, the nuclear physicist, it is an ideal situation, for it is no longer necessary for the designer to match his approach to the available materials. On the contrary, it is now becoming increasingly possible to obtain "custom-tailored" materials, which meet the specific requirements of the technical problem.

This development has of course far-reaching practical consequences. Industries that had the foresight to recognize the essential role of scientific

[1] See for example J. L. Meijering and G. D. Rieck, Philips tech. Rev. **19**, 109, 1957/58.
[2] See Scientific American **217**, No. 3, Sept. 1967.

research on materials have as a result gained a valuable lead. In aerospace techniques it has become common practice to treat materials research and materials development as an essential part of any new project. The same view is reflected in the structure of the nuclear-energy establishments. In many countries, however, so much manpower has been poured into these establishments that there is now surplus research capacity. Attempts are now being made to apply the knowledge that has been obtained about the control of material properties more widely in national industry, outside the specific domain of nuclear energy.

New ceramic materials

In the renaissance of materials to which I have referred, ceramics play an essential part. Most people think of ceramics as synonymous with earthenware or porcelain, products with a very ancient history. In this traditional sense, ceramics are still the product of an important branch of industry.

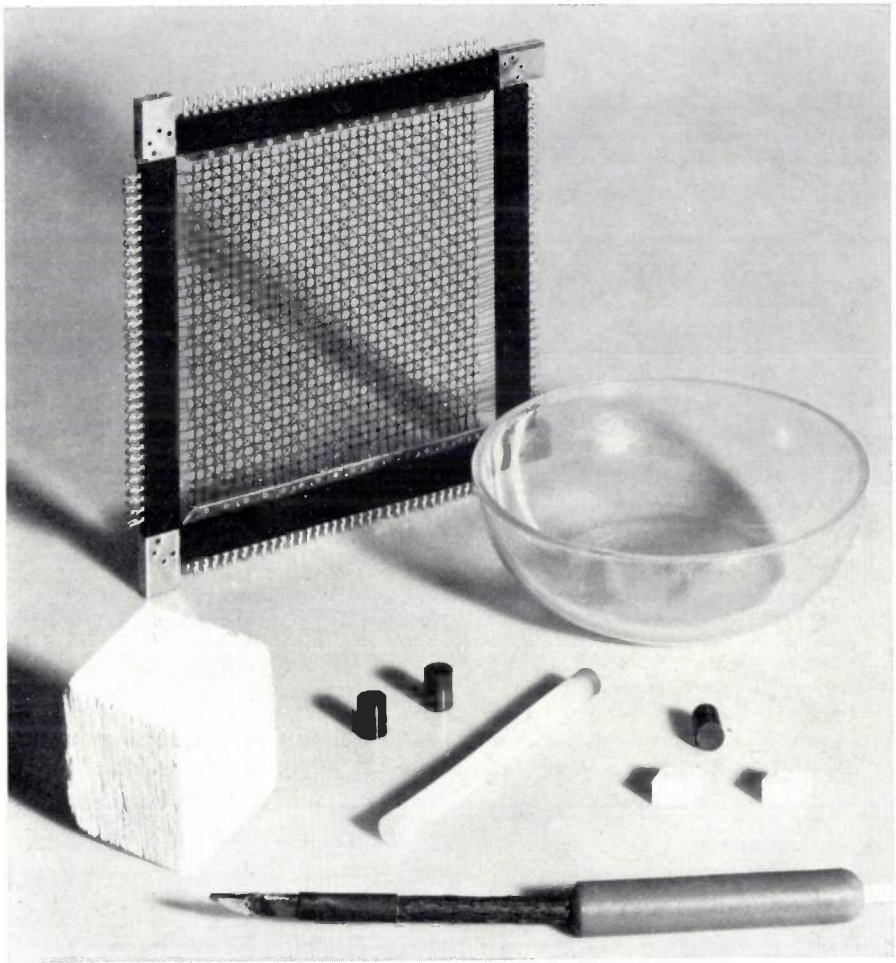


Fig. 2. Some modern ceramic products. A memory matrix with ferrite cores can be seen in the background. The white blocks on the right are made from fine-crystalline, densely sintered Al_2O_3; the material is so strong, resistant to wear and hard that the blocks can be used as cutting tools. The small black cylinders consist of uranium dioxide, fuel for nuclear reactors. The piece of tubing is made of coarse-crystalline non-porous Al_2O_3; it is resistant to high temperatures, to chemical attack at such temperatures, and is also (owing to the small quantity of pores translucent to visible light. The tubing is used as the envelope for high-pressure sodium lamps. The dish on the right and the large porous cube on the left are made of glass ceramic; their coefficient of expansion is extremely small and this makes the material resistant to temperatures up to about 700 °C; the porous material can be used, for example, for heat exchangers. Finally, in the foreground a soldering iron (laboratory model) can be seen in which the heating element consists of ceramic lanthanum-doped barium lead titanate [3].

Today, however, the term ceramics embraces more than the traditional ceramic materials, which all contain clay as a basic substance. Modern ceramic materials comprise a very wide range of inorganic compounds in which the oxides are predominant.

Graphite is also counted as a ceramic, and so are a number of carbides, nitrides, borides, silicides and other compounds. More and more use is being made of their unique properties. They are very hard, are highly resistant to wear or erosion, possess great strength at high temperatures, are chemically inert and many of them have remarkable electrical and magnetic properties. It is a very heterogeneous group of materials, with a very wide and varied range of applications (*fig. 2*). The American Ceramic Society did its utmost to find another name for ceramics. The best it could find was "inorganic, nonmetallic materials"; but the presence of two negatives in one designation did not seem too desirable. And so we refer merely to them as "ceramic" materials that have many features in common with traditional ceramics.

(*fig. 3*). During the forming process it is necessary to take the shrinkage that occurs during firing into account.

A characteristic difference between the modern and the traditional ceramics is perhaps the substitution of the concept tolerance for the old aesthetic criteria in the setting of standards.

In traditional ceramic products the formation of a liquid phase is almost invariably essential to allow the sintering process from the high melting primary phase to take place at reasonably low temperatures. In the new ceramic products (which may in fact be regarded as a class of materials and not so much as the products of a special branch of industry) liquid phases are not usually considered to favour the ready control of physical properties.

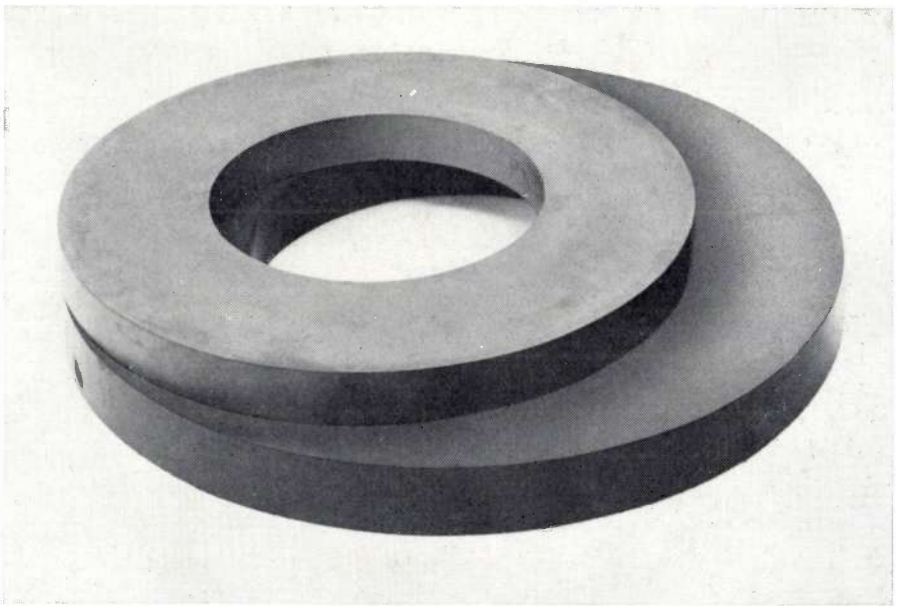


Fig. 3. An illustration of the shrinkage due to sintering. The photograph shows two ferrite rings for particle accelerators: the lower ring has been compacted but has not been sintered while the upper ring, which originally had the same diameter, has been sintered. The diameter of the lower ring is 55 cm, the diameter of the upper ring is 44 cm.

Most ceramic products are made from powders, and powder technology has always been the working terrain of the ceramist. Because of the brittleness of ceramic materials, grinding is the only way in which the fired products can be machined, and this is an expensive operation. A great advantage of the ceramic process, however, is that the products can be formed during manufacture. Starting from powdered raw materials, a product is formed that has a porosity of about 50% and is as yet not particularly strong. After the product has been heated for a time at high temperature the volume of pores decreases to less than 10% and the product is very much stronger

The first group of modern ceramic materials are the ferrites, materials derived from the mineral magnetite. The combination of useful magnetic and electrical properties makes these ferrites suitable for use at high frequencies, where problems arise with metallic magnetic materials. Many scientists at Philips Research Laboratories have had a considerable share in the development of these materials [4]. The development

[3] See for example E. Andrich, Philips tech. Rev. **30**, 170, 1969 (No. 6/7).

[4] See J. Smit and H. P. J. Wijn, Ferrites, Philips Technical Library, Eindhoven 1959.

of ferrites has been an excellent example of team work between solid-state physicists and chemists.

The second group are materials with a high dielectric constant, a typical example of which is barium titanate.

A third group are the nuclear ceramics, i.e. materials used as reactor fuels. The development of nuclear energy imposes specific requirements on the accurate control of the properties and composition of materials, such as high-melting uranium compounds, beryllium oxide, and also graphite. Uranium dioxide is without question the reactor fuel of today and for the near future.

For the electronics industry, particularly for the development of computers, for aerospace techniques, for the optical industry, and also for more everyday applications, as in the service-goods industry, there has grown up in the last ten years a vast need for knowledge of the behaviour and the manufacturing process of a large number of inorganic non-metallic materials. Prominent among these are the pure oxides and a number of special glasses [5].

Microstructure

It is still common practice to use polycrystalline materials, produced by a ceramic process, even though the synthesis of single crystals is also a practical proposition. For industrial purposes, however, the synthesis of single crystals has its attractions in only a few cases. The chemical compositions needed to obtain the exact properties required are often highly complicated. In this case a ceramic process is much simpler in principle and is also more economical for producing sufficiently homogeneous products on a large scale.

In many cases it is also necessary to use a ceramic technique because problems connected with dissociation reactions can arise in a melting process owing to the high melting point of the compounds.

The choice of a ceramic process implies the introduction of a number of microstructure parameters. The principal elements of the microstructure are: the size of the constituent crystallites, their size distribution, the presence of grain boundaries between the crystallites, the possible presence of crystallographic preferred orientations of the crystals, as found in the permanently magnetic barium hexaferrite $BaFe_{12}O_{19}$ (*fig. 4*), and the presence of a certain residual porosity.

The nature of the porosity may vary considerably. The pores may be very small in relation to the crystallites, but also relatively large. The pores may only be present at grain boundaries, or they may be entirely inside the crystallites (*fig. 5*).

One of the simplest examples of the way in which a physical property can be influenced through the microstructure is the scattering of light rays by pores. In aluminium oxide, which has a relatively high refractive index, the light transmission decreases to 0.01 % at a residual porosity of 3 %. Even when the porosity is 0.3 % the material transmits only 10 % of the light transmitted by a completely solid material.

The extent to which the various elements of the microstructure of a ceramic material influence the physical and mechanical properties is the subject of research in the material sciences [6]. Sometimes the relationship is extremely complex or not readily accessible to experiment. Of particular interest are the modern ceramic materials where the choice of microstructure is essential to the achievement of a desired physical property. It may be necessary, for example, to have a large number of grain boundaries, or a particular texture of crystal orientations.

Control of the microstructure

I shall now consider some basic aspects of the sintering process that determine the microstructure of ceramic materials. The microstructure is brought

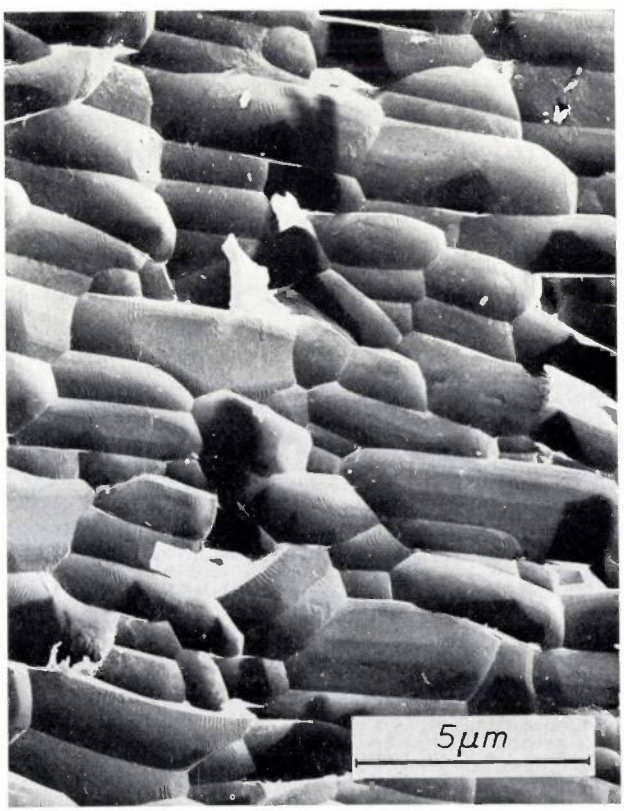
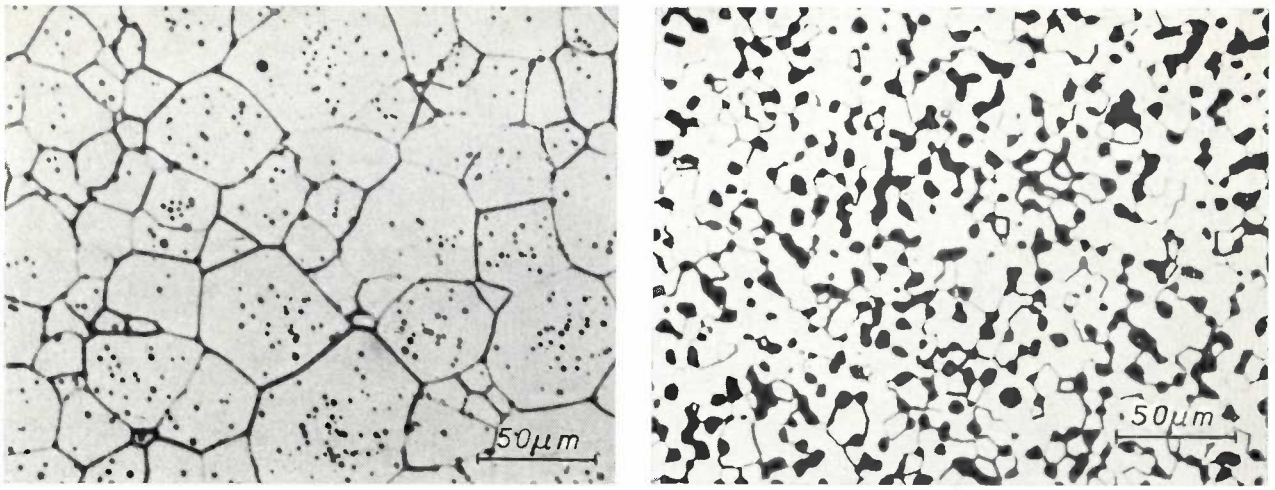


Fig. 4. Crystallites of barium hexaferrite oriented along the *c*-axis. The orientation is brought about by introducing a suspension of finely ground powder particles into a magnetic field, filtering them and then compacting them. The $(BH)_{max}$ achieved is more than 4×10^6 gauss oersted.



a *b*

Fig. 5. *a*) Pores in crystallites in Ni-Zn ferrite. *b*) Pores at the crystallite boundaries in Mn-Zn ferrite.

about during the sintering of an aggregate of small powder particles, which may or may not be compacted to a particular shape. The free surface of a substance represents an extra amount of energy, called surface energy. The difference in surface energy between a powder and a single crystal supplies the driving force for a number of thermally activated processes, principal among which are sintering and grain growth.

In the sintering process the powder particles in a compacted product are able to grow together and the open spaces, the pores, are filled up with material. Because of this the product exhibits macroscopic shrinkage. The amount of energy available for the sintering process depends mainly on the surface of the powder. Since this surface energy is very small, no more than about 1 calorie per gramme atom, a sufficiently high *sinter reactivity* can only be achieved with sufficiently small particles. Generally speaking, the particle size required is less than 1 micron.

This initial condition sets some difficult requirements on the actual fabrication process from the very start, because the handling of such fine powders presents considerable technical difficulties. In the sintering process the voids between the powder particles, the pores, must be filled with material in order to obtain a strong and solid product. Effective sintering thus means that the packing of the particles by compaction in the forming process must be such that a product is obtained in which the small pores are distributed as homogeneously as possible. The poor flow properties of fine powders complicate the forming process to a considerable extent.

In order to close up the pores a material-transport mechanism is necessary. In its simplest form, sintering can be compared with the flowing together of drops of liquid when brought into contact with each other. For drops of liquid this process is easy to understand, since the liquid flow is brought about by the effect of surface tension. The only resistance to this material transport is the viscosity of the liquid. But how does this work out in the case of an aggregate of powder particles?

In its original significance the phenomenon of sintering has always been discussed in connection with the occurrence of liquid phases upon heating, which causes a strengthening of the powder aggregate. The name sintering is derived from this process. It has been found, however, that although liquid phases do occur, the densification is not caused by macroscopic liquid flows. The process appears to have more to do with the solution of the solid phase in the liquid phase, followed by precipitation. An important problem then remaining is to establish how the material transport takes place in the systems where no liquid phases are formed, in particular in the systems which are characteristic of modern ceramic products.

After many years of phenomenological studies the Russian researcher J. Frenkel laid the basis for a scientific approach to the sintering of solids in 1945 [7]. He presented a theory for the formation and growth

[5] See G. Bayer, Oxyde als Werkstoffe der modernen Technik, Neue Zürcher Zeitung 27 Jan. 1970, Fernausgabe No. 26, pp. 17-20 and 25.

[6] This subject is discussed in detail in an article by G. H. Jonker and A. L. Stuijts, shortly to appear in this journal.

[7] For the work of J. Frenkel and the other investigators mentioned here, see the following review articles:
H. Fischmeister and E. Exner, Theorien des Sinterns, Metall **18**, 932-940, 1964 and **19**, 113-119 and 941-946, 1965;
R. L. Coble and J. E. Burke, Sintering in ceramics, Progress in Ceramic Science **3**, 197-251, 1963;
F. Thümmler and W. Thomma, The sintering process, Metallurg. Revs. **12**, 69-108, 1967.

of the contact surface between two spherical particles, under the influence of the surface tension, by means of a mechanism of viscous flow. A year later B. Ya. Pines extended this theory with an analysis of the concentrations of point defects in solids with variously curved surfaces. In the subsequent years much fruitful theoretical and experimental work was done on the mechanisms underlying material transport in sintering. This work, started by G. C. Kuczynski with metal powders and followed by many elegant investigations undertaken by other American researchers, for example W. D. Kingery and his associates, was based on model experiments performed on different groups of materials: metals, oxides, glass, and ionic compounds such as sodium chloride. As a result of this work the basic mechanisms of material transport have become well known.

It also emerged from these experiments that the principal mechanism that can explain the shrinkage of crystalline materials during sintering is based on diffusion of the components of the compound through the bulk of the material. The possibility that sintering is determined by bulk diffusion, and thus that point defects in the lattice play an important part, had already been recognized, but it took a relatively long time before it could be properly interpreted. Although it was clear where the flow of material was going to, it was not clear where it came from. The assumption that the material was transported from the outside surface of the object into the bulk proved to be untenable. In that case the sintering process would be slower than was actually observed, and moreover no dependence was found between the sintering rate and the size of the object.

It was in the same period that the Nabarro–Herring theory became known. This theory gives an explanation of the deformation of a polycrystalline metal under an external pressure at high temperature. This microcreep, as it is called, gives rise to diffusion flows in the crystals, in which the material is transported from grain boundaries that are under pressure to others that are under a tensile stress (*fig. 6*). These grain boundaries, being the interfaces between neighbouring crystallites, play an essential role in the course of the process. A small grain size is very important to this mechanism.

With this theory, the shrinkage during the sintering of crystalline materials could be explained. It fell to J. E. Burke to establish a clear relationship between the behaviour of the grain boundaries during the sintering process and the results of the process. It is known that there can be grain growth at the temperatures at which sintering is carried out, the driving force behind this grain growth being the interfacial energy between neighbouring crystallites. Because of the grain growth the diffusion has to extend over greater distances.

Although normal growth processes, including grain coarsening always follow a course in which the large grains grow at the expense of the small ones, the processes take place in a controlled manner as long as all the crystallites have about the same chance of growing.

From microscopic observations Burke concluded that a certain form of grain growth that is fatal to the further densification of the material can occur during sintering. It appears that some crystals, for reasons that are not yet entirely clear, may grow fairly suddenly to relatively large dimensions. By enclosing the pores in the crystal during their rapid growth, they render the important Nabarro-Herring mechanism inoperative.

The effect described here, in which a few crystals grow to dimensions much greater than those of the crystals in the matrix, occurs very frequently in crystalline materials and has been given various

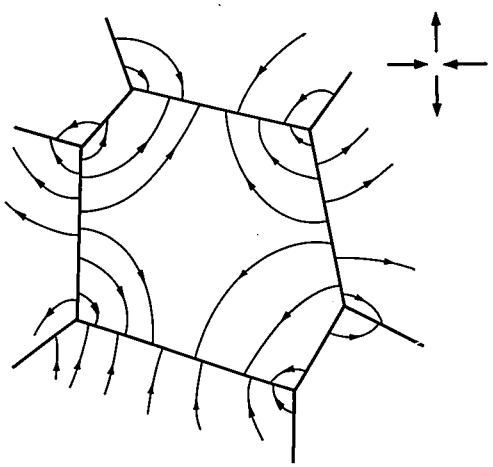


Fig. 6. The diffusion flow in a polycrystalline material: there is transport from grain boundaries that are under pressure to grain boundaries that are under a tensile stress. (Made available by C. Herring, J. appl. Phys. **21**, 437, 1950.)

names in the literature. One of the most characteristic names, which reflects the uncontrolled nature of the process, is cannibal grain growth (*fig. 7*).

There are some exceptional cases where this form of grain growth is desired. In general, however, Burke's studies of these processes have shown that control of grain growth is essential to effective sintering. On the basis of the knowledge thus gained R. L. Coble managed to sinter aluminium oxide completely solid by doping it with magnesium oxide. This was a real breakthrough in sintering practice.

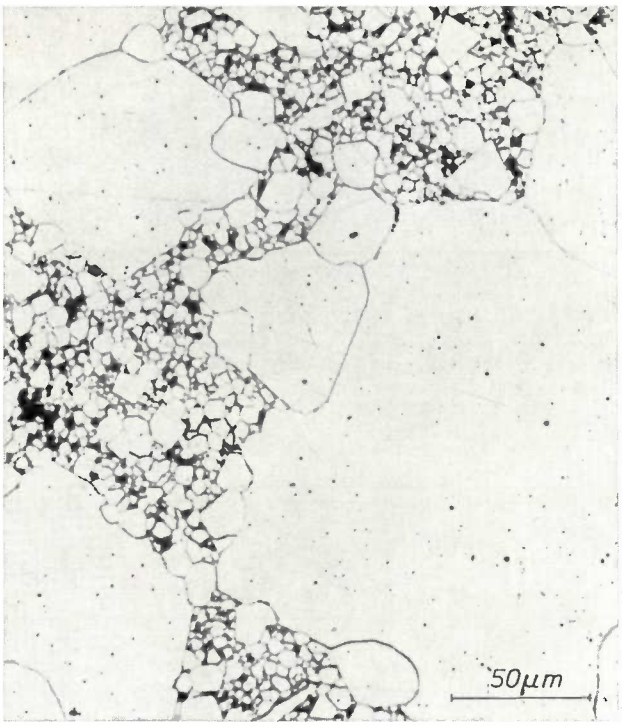


Fig. 7. An example of "cannibal" grain growth, in Ni-Zn ferrite.

The important technical result of Coble's work was that it now became possible to use aluminium oxide, long known for its useful properties, for the envelopes of high-pressure sodium lamps. In this application full advantage can be taken of the high melting point of the material, its great resistance to chemical attack, and its reasonably high strength at high temperatures. When the material is sintered completely free from pores, it can be made translucent to visible light. Lamps that used such a translucent aluminium oxide give a beautiful white light and a high luminous efficiency. It is typical of the further development in the sintering of many kinds of ceramic products that the transparency of the product has come to be used as a measure of the control of the process. In more senses than one, scientific research has made the subject clear!

After this, of course, it became all the rage for ceramists to study the effect of doping on sintering and grain-growth behaviour. In most cases the results were negative, in others they were not applicable owing to the introduction of the second phase. But our own results with the sintering of magnetic spinels had already shown that there were other possible ways of achieving control of sintering processes [8]. Since these materials are opaque, at least to light in the visible part of the spectrum, this result could not be represented in terms of transparency, which had become the standard in sintering practice. In addition to this instance, many other examples have meanwhile become known where densification sintering to a completely solid product is possible without the need for doping.

These results are probably to be interpreted in the following way. Sintering and grain growth are two processes that occur simultaneously, and densification can be hindered by the grain growth. To obtain a good result the conditions must therefore be chosen in such a way that the sintering process can take place rapidly. In the first place, very fine powders of high sinter reactivity should be used. Secondly, the compacted product must be as homogeneous as possible in order to rule out problems from local variations in sintering rate. Finally, as the sintering of crystalline substances is a diffusion process, there must also be empty lattice sites, or vacancies, present.

In the last few years it has become clear that in this last respect ionic compounds differ characteristically from metals. In metals the concentration of vacancies is usually dependent on the temperature alone, and is therefore difficult to control independently. In ionic crystals, on the other hand, there may often be marked deviations from the stoichiometric composition. This means that the concentrations of lattice defects in the crystal are not only determined by the temperature but also depend on the chosen chemical composition. Recent investigations by P. Reijnen have shown that since optimum sintering behaviour depends on the transport of both cations and anions, the choice of the lattice vacancies and their concentration is of the utmost importance [9]. It is for example rather remarkable that adding magnesium oxide to aluminium oxide gives the appropriate lattice vacancies for good sinterability.

The great value of the knowledge outlined above is that straightforward practical rules can be laid down which should ensure good sintering. On the other hand it cannot yet be clearly established whether an effect of grain-growth inhibition, by means of an appropriate doping agent, is in itself sufficient to give good sintering. In contrast to the control of the chemical composition, it is not possible to use the available knowledge of grain-growth effects to indicate practical rules for improved sintering.

In what I have said so far I have given a broad picture of some basic aspects of the sintering process that determine the ultimate microstructure. Owing to the wide variety of materials and applications I have necessarily had to leave out a large number of sintering phenomena, some of them highly complex.

[8] A. L. Stuijts, Proc. Brit. Ceramic Soc. **2**, 73, 1964.
[9] P. J. L. Reijnen in: Reactivity of solids, Proc. 6th Int. Symp., Schenectady 1968, p. 99, and also Philips tech. Rev. **31**, 24, 1970 (No. 1).

They form a fascinating study for the scientific worker who wishes to understand the microstructures of materials from the processes taking place in them. I shall now mention a few points which are typical of the *technology* of modern ceramic materials, paying special attention to the synthesis of materials with "tailored" properties.

In the first place the knowledge gained from research has led to a lively interest in the development of new processes for the synthesis of raw materials and compounds. The preparation of sinterable powders by the conventional processes of mixing, calcination and grinding, is of limited scope. Moreover, the understanding of the role of vacancies can only be applied if the chosen chemical composition does in fact determine their nature and concentration. Impurities can then have a marked disturbing effect, and "p.p.m." has now become a common term in this field too. The development of processes for the preparation of raw materials whose purity is maintained, or where purification takes place during the process, and in which the chemical composition can be well controlled and adjusted, is now in full progress. In the field of electronic materials there is considerable interest in the preparation of compounds by the spray-drying or freeze-drying of mixed salts. For nuclear-energy applications the development of the sol-gel process has opened up unique possibilities [10]. With all these processes powders of high purity and high sinter reactivity can be produced.

The powders obtained by these processes do not all lend themselves to the application of a normal forming method. Often it is not possible to compact them to the desired high density. The sol-gel process can produce beautifully rounded pellets with dimensions ranging from a few microns to a few millimetres (*fig. 8*). It has not yet been found possible, however, to retain the intrinsic advantages of this process for making objects of other shapes. A study of the forming process is therefore needed, with special reference to the way in which the densification is influenced by characteristic features of the powders used.

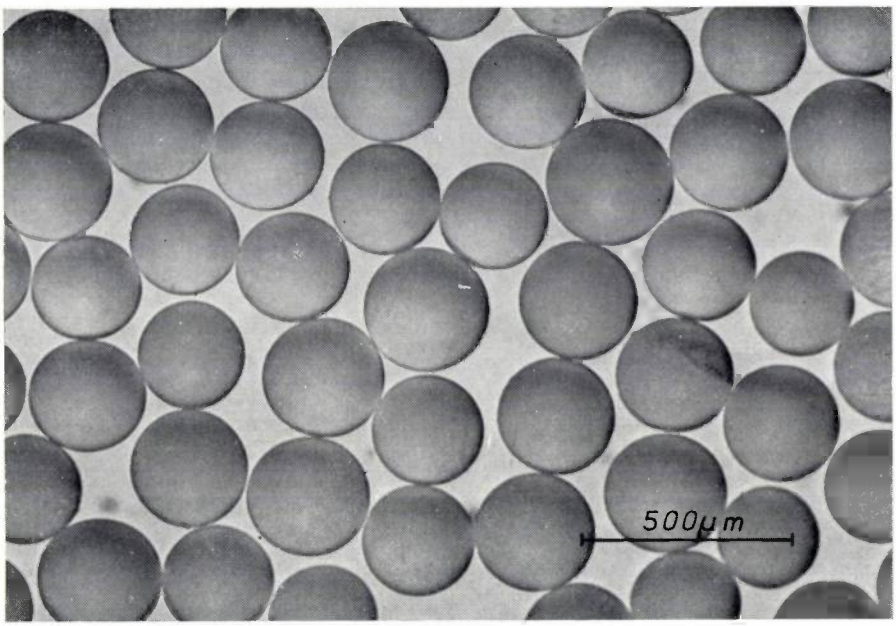


Fig. 8. Small spheres of Al_2O_3 obtained by sintering a highly homogeneously distributed fine powder made by the sol-gel process [10]. (By courtesy of the Atomic Energy Research Establishment, Harwell.)

Finally, the development of processes for forming microstructures that are difficult or impossible to obtain by the normal process is very important. I have in mind materials in which the constituent crystallites have a crystallographic texture, as required for optimum properties in hexagonal permanent-magnetic ferrites. Other examples are compounds with volatile components, or compounds in which the chemical composition cannot be freely chosen to obtain good sinterability. Materials are also often needed which have a low pore volume yet at the same time very small crystallites. In many of these cases the processes used combine high temperature with high pressure [11]. The external pressure applied

allows the sintering temperature to be substantially reduced, thus strongly suppressing the grain growth. Many materials with interesting properties have been obtained by such methods. They include microwave ferrites for use at high microwave power, magnesium oxide of very high mechanical strength, fine-crystalline aluminium oxide, and the piezoelectric material potassium sodium niobate, which cannot be sintered by the normal method.

The renaissance in ceramic technology which I have described is a typical example of the way in which many new materials are being created and used in modern engineering. A characteristic of the various material technologies is the control of physical and mechanical properties through the control of microstructures. To be able to synthesize materials it is essential to control the processes that determine the formation of these structures. This is at once product engineering and process engineering, and it is the main function of the technologist here to ensure that he knows the industrially most suitable process for giving a material the desired microstructure.

[10] A review article by A. L. Stuijts has been published in Science of Ceramics **5**, 335, 1970.

[11] See for example G. J. Oudemans, Philips tech. Rev. **29**, 45, 1968.

Summary. The article is almost identical with the text of the address delivered by the author on his inauguration as Visiting Professor at the Technical University of Eindhoven. In about 1950 a breakthrough occurred in ceramic technology. New understanding enabled the microstructure to be controlled with such accuracy that the materials could be "tailored" to the requirements of practical applications. The author discusses the use of special doping agents, the application of very fine and highly pure powders possessing high sintering reactivity, made for example by spray-drying or freeze-drying or by the sol-gel process, the use of materials which have a slight deviation from the stoichiometric composition, the simultaneous application of high pressure and high temperature, and other related subjects.

An electronic starter for long fluorescent lamps

J. C. Moerkens

The semiconductor techniques of today offer many new possibilities to designers of control gear for fluorescent lamps. Circuits that were once too unreliable, expensive or inefficient, or reduced the life of the control gear can now be employed with success.

Large areas can be lit more attractively by using long fluorescent lamps rather than short ones. The longer lamps also have the advantage that they are more efficient. The light loss that occurs at the electrodes because the cathode fall of a gas discharge is dark is independent of the length of the lamp, and is therefore less significant for longer lamps. The mounting and maintenance costs per unit length are also lower for longer lamps. For all these reasons fluorescent lamps about 2.5 metres long, popularly known as 8 ft lamps, have been developed along with the conventional types of 1 to 1.5 m length — the 4 ft and 5 ft lamps.

Because of their greater length these 8 ft lamps need a higher ignition voltage and a higher maintaining voltage than the types now in common use. The conventional glow starters (*fig. 1a*) cannot therefore be used for the 8 ft lamps. These starters were originally developed for igniting lamps with a maintaining voltage of 100 V at a supply voltage of 220 V; the maintaining voltage of the long lamps is about 180 V and is too close to the supply voltage for glow starters to operate reliably. Alternative solutions, such as starterless circuits or ignition by means of an auxiliary circuit, also have their drawbacks. For 8 ft lamps the starterless circuit (fig. 1*b*) requires a transformer with an open-circuit voltage of at least 350 V. A transformer of this type with the built-in leakage reactance for stabilizing the lamp would be expensive and the electrical losses would be high. Moreover, this kind of circuit usually requires special lamps — fitted with a metallized ignition strip — or an earthed metal fitting. The other alternative, starting the lamp with an auxiliary circuit (generally a resonant circuit), has the disadvantage that this circuit continues to take current after the lamp has lit. Although this loss can be avoided by switching off the auxiliary circuit after starting, extra switches or relays are required, making the installation more expensive to buy and to maintain.

Now that inexpensive and reliable semiconductor devices are readily available a lamp can be ignited by an auxiliary circuit that can be switched off without using switching contacts after the lamp has lit. The disadvantages of ignition by an auxiliary circuit are eliminated in this way.

A starter circuit will now be described that has been developed for 8 ft fluorescent lamps, with switching by semiconductor devices. The circuit is designed to operate from the regular supply mains (e.g. 220 V, 50 Hz, single phase).

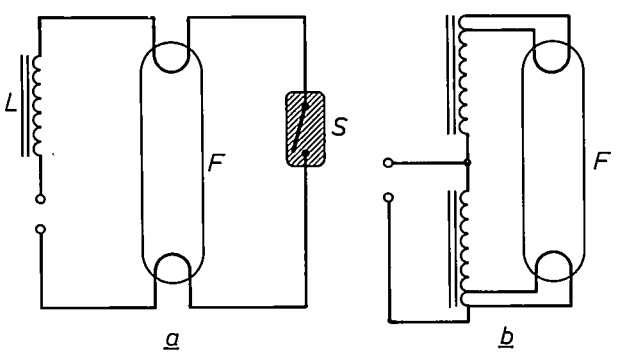


Fig. 1. *a*) Circuit for a fluorescent lamp with glow starter. *F* fluorescent lamp. *L* choke, whose impedance stabilizes the discharge. *S* bimetal glow switch in a small neon-filled bulb. Before the mains supply voltage is switched on *S* is open. When the mains voltage is switched on there is a gas discharge in the neon-filled bulb; this discharge heats the bimetal element and the switch closes. Current now flows, heating the electrodes. The neon discharge is short-circuited by the closed contacts and is therefore extinguished. Without the heat from the discharge the bimetal element cools down and the switch opens again. The resultant voltage surge caused by the inductance *L* ignites the lamp. The glow starter is designed so that the neon discharge is ignited at the lowest value of the mains voltage, but not by the maintaining voltage of the lamp.
b) Starterless circuit. Before the lamp has ignited the transformer supplies the voltage for heating the electrodes as well as the ignition voltage across the lamp. After ignition the voltage across the lamp falls because of the phase shift in the voltage across the upper windings of the transformer, which act as the ballast impedance. During operation the full filament voltage is applied to the filaments of the lamp. The lower windings of the transformer form an undesirable load on the mains while the lamp is operating.

J. C. Moerkens is with the Philips Lighting Division, Eindhoven.

Principle of the starter circuit

The function of the starter and ballast circuits of a fluorescent lamp is to ignite the lamp and keep it operating stably. For ignition it is necessary first of all to heat the electrodes, then to apply a voltage surge that is large enough to initiate the discharge in the lamp.

Our circuit (*fig. 2*) consists basically of two chokes, L_1 and L_2, a capacitor C, and a switch S. Together with the capacitor the chokes form two resonant circuits L_1C and L_2C, which can be excited in turn by opening and closing the switch. Choke L_1 also stabilizes the lamp current while the lamp is running (i.e. it is the *ballast*).

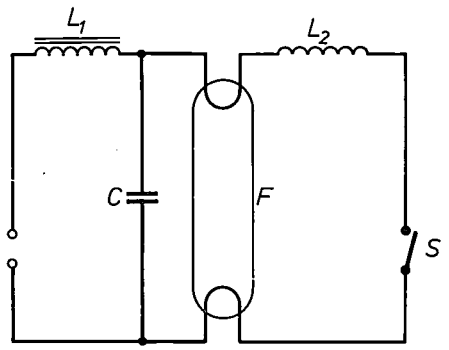


Fig. 2. Basic diagram of the starter for long fluorescent lamps. The capacitor C draws energy from the mains. When switch S closes, this energy is partly dissipated by the resonant circuit L_2C in the electrodes of the lamp. When S opens again at the instant that C is charged up, the voltage surge produced by the resonance of the circuit L_1C is sufficient to ignite the discharge across the lamp F. While the lamp is alight the choke L_1 keeps the discharge stable.

In describing the operation of the starter it is convenient to assume that the circuit is connected to the supply voltage at the instant at which this voltage passes through zero ($t = t_0$, see *fig. 3*). The switch S is then open. The values of L_1 and C are chosen so that the resonant frequency of L_1C is well above the mains frequency and the impedance of C is much greater than the impedance of L_1: the voltage across C therefore follows the mains voltage fairly closely in both phase and amplitude. If we now connect the switch S at the time t_1, the circuit L_2C is completed. When the switch is closed, a charge has built up on the capacitor, and consequently this circuit is now set into oscillation. The inductance of L_2 is made much smaller than that of L_1, so that the resonant frequency of L_2C is much higher than that of L_1C. The impedance of L_1 to these oscillations is sufficiently high to prevent the low impedance of the mains from damping the circuit L_2C. The electrodes of the lamp, which are in the form of filaments, give the only damping in the circuit. The current in the circuit heats the filaments, thus causing the ignition voltage of the lamp to fall. We now open the switch S again at the instant at which the current in the circuit L_2C has returned to zero. The circuit L_2C is then broken and circuit L_1C is triggered into oscillation by the opening of the switch. As a result the voltage across the capacitor increases to an amplitude of more than twice the instantaneous value of the mains voltage. If the electrodes have been pre-heated sufficiently, the lamp will ignite; if not, the whole cycle must be repeated until the lamp does ignite.

Practical model; characteristics

In the actual device (*fig. 4*) the function of the switch S is performed by a thyristor circuit. The switch closes at the instant that one of the two thyristors $T_{1,2}$ connected in parallel opposition starts to conduct. The instant at which a thyristor starts to conduct is determined by the R-C time constant of the network consisting of the resistors R_1, R_2 and R_3 and the capacitors C_1 and C_2. The capacitors are charged by the supply voltage via the resistance network. When the voltage across the capacitors exceeds a given value, one of the two diode pairs connected in parallel opposition, $D_{1,2}$, called "diacs" (*fig. 5*) switches on and part of the charge flows to the control electrode of the associated thyristor, which then starts to conduct; the switch is

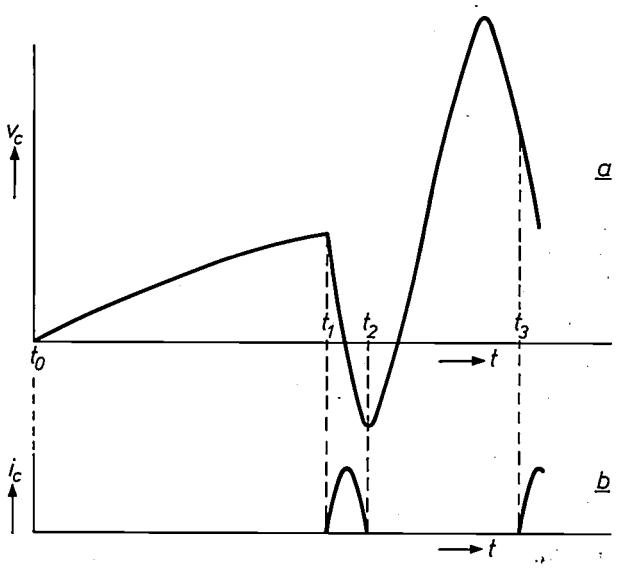


Fig. 3. *a*) Voltage v_c across the capacitor C in fig. 2 as a function of time t after switching on the supply voltage at time t_0. At t_1 the switch S closes and the circuit L_2C is set into oscillation. At t_2 the switch opens again and the circuit L_1C goes into oscillation. At t_3 the switch again closes, and the whole cycle begins again from $t = t_1$.
b) The current i_c in the circuit L_2C as a function of time.

then closed. When the current through the conducting thyristor becomes zero — which happens after half a period of the oscillation in the circuit L_2C — the conduction ceases: the switch is then open again. The voltage across capacitor C is then increased in amplitude by the oscillation in the circuit L_1C. After about half a period of this oscillation, C_1 and C_2 are again charged up to the breakdown voltage of the diacs, the thyristor can start to conduct again and the whole cycle is repeated. The breakdown voltage of the diacs is reached more quickly the second time than the first: the capacitors only need to be charged from the time at which the forward voltage of the diacs is reached.

During the first half period of the mains voltage the whole cycle is repeated a number of times. Once the supply voltage has dropped to a value at which it is no longer high enough to charge up C_1 and C_2 to the breakdown voltage of the diacs, the switch remains open and the oscillation of the circuit L_1C dies away. Usually the lamp will not yet have ignited, and the whole process repeats itself during the next half period of the mains voltage, the other thyristor now periodically opening and closing the switch. The waveform of the voltage across the lamp during ignition is shown in *fig. 6*.

The ignited lamp

Once the lamp has ignited, the voltage across it will be very nearly a square wave (*fig. 7*). The amplitude of this square-wave voltage is equal to the maintaining

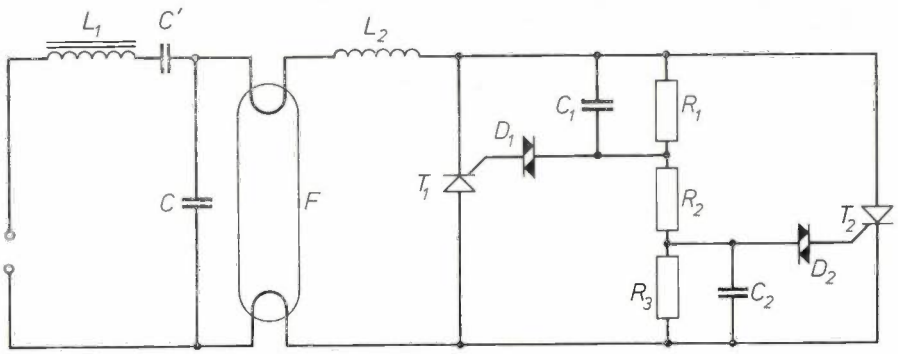


Fig. 4. Complete circuit diagram of the starter. L_1, L_2 and C as in *fig. 2*. The thyristors T_1 and T_2 form a switch; T_1 conducts during the negative half-cycle of the supply voltage, T_2 during the positive half-cycle. The time at which the switch closes is determined by the R-C network $R_1R_2R_3C_1C_2$. When the voltage across C_1 and C_2 exceeds the breakdown voltage of the diacs D_1 and D_2, one of the two delivers a voltage surge to the control electrode of the corresponding thyristor, causing it to conduct. The thyristor stops conducting (the switch opens) when the current flowing through it falls to zero. During the next half-cycle of the mains voltage the switching function is taken over by the other thyristor.

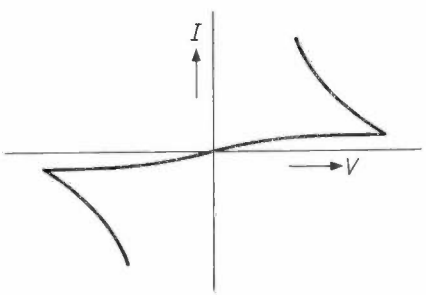


Fig. 5. Current-voltage characteristic of a "diac". A diac consists of two P-N-P-N diodes connected in parallel opposition. This combination gives a current-voltage characteristic rather like that of a gas discharge. A certain voltage has to be reached before the diac starts to conduct, but after this the voltage can fall quite a long way before the current cuts off.

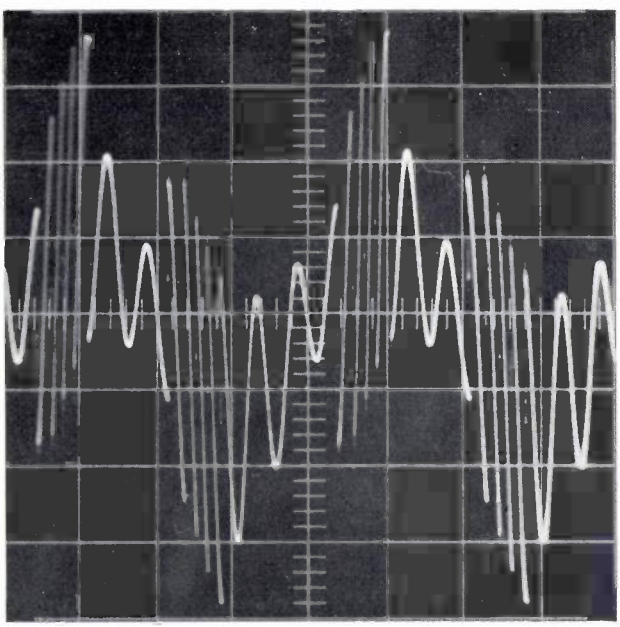


Fig. 6. Oscillogram of the voltage across the lamp during the ignition process. The groups of peaks following quickly upon one another which repeat themselves every half cycle are the excitation effects arising in the circuits L_1C and L_2C (see fig. 3).

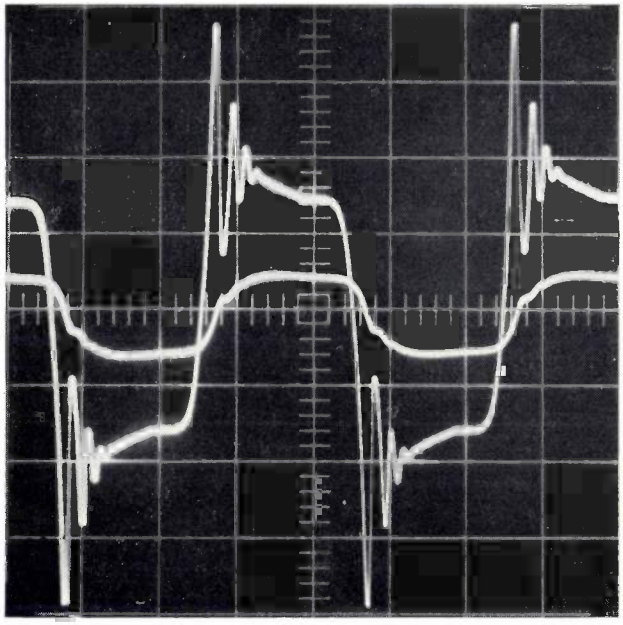


Fig. 7. Oscillogram of the voltage across the lamp while it is alight (curve with high amplitude) and of the voltage across the capacitors C_1 and C_2 in fig. 4 (low-amplitude curve). The reignition peaks with the oscillations that follow them can clearly be seen in the voltage across the lamp. These peaks do not appear in the voltage across C_1 and C_2, because of phase shift and smoothing.

voltage of the lamp, and is thus lower than that of the supply voltage. The voltage divider $R_1R_2R_3$ is designed so that the capacitors C_1 and C_2 can no longer be charged up by this square-wave voltage to the breakdown voltage of the diacs; the thyristors then receive no further triggering pulses and the switch now stays open the whole time. The starter circuit is then out of action.

While the lamp is operating the discharge is extinguished after every half-period of the supply voltage and then has to be reignited. In a short lamp, where maintaining voltage is about half the mains voltage, this presents no difficulties. After the discharge is extinguished the voltage across the ballast choke falls rapidly and the full instantaneous value of the mains voltage then appears across the lamp (*fig. 8a*).

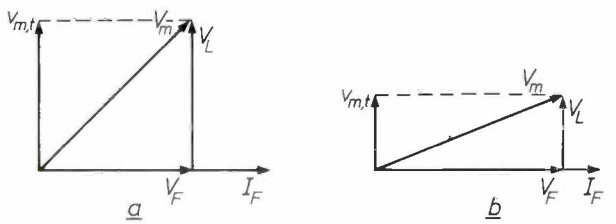


Fig. 8. *a*) Vector diagram for a fluorescent lamp whose operating voltage V_F is much smaller than the mains voltage V_m. *b*) The same but with V_F only slightly lower than the mains voltage. In this case the voltage $v_{m,t}$, which has to reignite the lamp, is too small. V_L is the voltage across the ballast.

This voltage is sufficient to reignite the discharge that has just been extinguished.

For a long lamp, where the maintaining voltage is only slightly lower than the mains voltage, the choke used for stabilizing the discharge has a much lower inductance. The voltage across the lamp after extinction of the discharge is therefore too low to permit reignition (fig. 8*b*). If we now connect a capacitor in series with the lamp, as well as the choke, we get the situation illustrated in *fig. 9*. The voltages across capacitor and choke differ in phase by 180°, and we can now make the inductance and the capacitance so large that the voltage that appears across the lamp after the discharge has been extinguished is high enough to reignite the discharge.

While the lamp remains alight the voltage across the auxiliary capacitors C_1 and C_2 continues to follow the voltage across the lamp. However, as fig. 7 shows, there is a slight delay. This means that when the lamp ignites again there will be a negative voltage on the control electrodes of the thyristors, so that they will be better able to stand the voltage surges that appear across them during reignition.

The negative voltage across the capacitors also

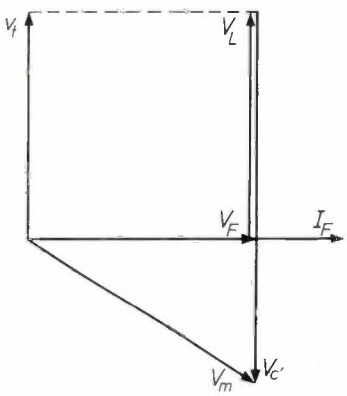


Fig. 9. Vector diagram for a lamp with a maintaining voltage of 180 V connected in series with a choke and a capacitor to a supply voltage of 220 V. With this arrangement the voltage v_t that appears across the lamp when it has extinguished and the voltage across the choke is zero is sufficient to reignite the lamp. $V_{C'}$ is the voltage across the capacitor C' (fig. 4), the other magnitudes are as in fig. 8.

prevents them from being charged up by the reignition peak to a value higher than the ignition voltage of the diacs, which would have the effect of bringing the starter circuit back into operation again.

Phase shift and mains voltage distortion

Since the maintaining voltage of an 8 ft lamp is close to the mains voltage, the phase difference between the mains voltage and the current through the lamp is small (fig. 9). Long lamps do not therefore need any

special measures for phase correction, such as those required for 4 ft and 5 ft types.

We have already seen that the voltage waveform across the terminals of the operating lamp is approximately a square wave. Fourier analysis of such a waveform shows that it has strong odd harmonics. These harmonics of the supply frequency must not of course enter the mains, nor should the high-frequency components due to the ignition surge, or they would cause interference elsewhere. In our circuit the capacitor C and the choke L_1 not only fulfil their normal starter and ballast functions, they also form a highly effective filter (see fig. 4), which stops the third, fifth and higher harmonics. Additional measures to suppress mains interference caused by the lamp are therefore superfluous.

Finally, a few practical details of our circuit should be given. The total electrical losses in an electronic starter and ballas designed for an 85 W lamp are only 12 W while the lamp is alight (including the losses at the lamp electrodes). The power factor ($\cos \varphi$) is 0.8 (capacitive). The use of this circuit does away with the need for an earthed metal fitting or for special lamps fitted with an ignition strip.

Summary. To ignite fluorescent lamps about 2.5 metres long, which are very effective for lighting large areas, normal glow starters cannot be used. This is because the maintaining voltage of such long lamps is so close to the mains voltage that glow starters would not operate reliably. With the electronic starter described in this article the lamp is ignited by means of a resonant circuit, which is switched off while the lamp is alight to avoid needless power consumption. Two thyristors connected in parallel opposition act as an inexpensive and reliable switch for this purpose. The switch is controlled via "diacs" by an R-C network connected across the lamp electrodes. No earthed metal fittings or ignition strip are required.

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, H. 't Lam & W. Heinle (Forschungsinstitut AEG-Telefunken, Ulm/Donau): The low-temperature velocity-field characteristic of *n*-type gallium arsenide.
Physics Letters **29A**, 596-597, 1969 (No. 10). *E*

H. J. Akkerman (Instituut voor Kernphysisch Onderzoek, Amsterdam): Precisiedraaibank met hydrodynamische hoofdaslagering.
Mikroniek **9**, 248-251, 1969 (No. 11).

H. J. Akkerman (Instituut voor Kernphysisch Onderzoek, Amsterdam): Een draaibank met numerieke beitelpositie-aflezing.
Mikroniek **9**, 258, 1969 (No. 11).

W. Albers & J. Verberkt: Isothermal substitutional growth of single crystals.
Philips Res. Repts. **25**, 17-20, 1970 (No. 1). *E*

E. Allamando, E. Constant, G. Salmer (all with Faculté des Sciences de Lille) **& A. Semichon:** Propriétés hyperfréquences des diodes à avalanche. Modes d'oscillation.
Acta electronica **12**, 211-253, 1969 (No. 3). *L*

J. A. Appels, E. Kooi, M. M. Paffen, J. J. H. Schatorjé & W. H. C. G. Verkuylen: Local oxidation of silicon and its application in semiconductor-device technology.
Philips Res. Repts. **25**, 118-132, 1970 (No. 2). *E*

A. J. F. de Beer & G. Diemer: Geïntegreerde schakelingen, 2. Opbouw en toepassingen.
Natuur en Techniek **37**, 318-327, 1969 (No. 9). *E*

K. Bethe: Über das Mikrowellenverhalten nichtlinearer Dielektrika.
Thesis, Aachen 1969. *H*

J. H. Bijleveld (Instituut voor Kernphysisch Onderzoek, Amsterdam): Het argonarc-lassen van aluminium.
Mikroniek **9**, 254-257, 1969 (No. 11).

G. Blasse: Lanthanide tellurates Ln_6TeO_{12}.
J. inorg. nucl. Chem. **31**, 3335-3336, 1969 (No. 10). *E*

G. Blasse: Thermal quenching of characteristic fluorescence.
J. chem. Phys. **51**, 3529-3530, 1969 (No. 8). *E*

G. Blasse & A. Bril: Energy transfer in Tb^{3+}-activated cerium(III) compounds.
J. chem. Phys. **51**, 3252-3254, 1969 (No. 8). *E*

R. Bleekrode & W. van Benthem: Resonance fluorescence of Cu atoms in the gas phase.
J. chem. Phys. **51**, 2757-2758, 1969 (No. 6). *E*

J. B. de Boer & D. A. Schreuder (Philips Lighting Division, Eindhoven): Betrachtungen über die Anwendung von Halogenlampen für die Kraftfahrzeugbeleuchtung.
Lichttechnik **21**, 88A-92A, 1969 (No. 8).

G. A. Bootsma & F. Meyer: Measurement of adsorption on semiconductors by ellipsometry and other methods.
Surface Sci. **18**, 123-129, 1969 (No. 1). *E*

C. J. Bouwkamp: Packing a rectangular box with the twelve solid pentominoes.
J. combin. Theory **7**, 278-280, 1969 (No. 3). *E*

C. J. Bouwkamp & N. G. de Bruijn (Technical University of Eindhoven): On some formal power series expansions.
Proc. Kon. Ned. Akad. Wetensch. **A72**, 301-308, 1969 (No. 4). *E*

J. C. Brice: Facet formation during crystal pulling.
J. Crystal Growth **6**, 205-206, 1970 (No. 2). *M*

J. C. Brice, G. W. Lelievre & P. A. C. Whiffin: Hydropneumatic apparatus for the Czochralski growth of crystals.
J. sci. Instr. (J. Physics E), ser. 2, **2**, 1063-1065, 1969 (No. 12). *M*

K. H. J. Buschow & H. J. van Daal: Evidence for the presence of the Kondo effect in the compound $CeAl_2$.
Phys. Rev. Letters **23,** 408-409, 1969 (No. 8). *E*

K. H. J. Buschow & A. S. van der Goot: The holmium-cobalt system.
J. less-common Met. **19,** 153-158, 1969 (No. 3). *E*

K. H. J. Buschow, A. S. van der Goot & J. Birkhan: Rare-earth copper compounds with $AuBe_5$ structure.
J. less-common Met. **19,** 433-436, 1969 (No. 4). *E*

K. H. J. Buschow, P. A. Naastepad (Philips Philite and Metalware Manufacturing Division, Eindhoven) **& F. F. Westendorp:** Preparation of $SmCo_5$ permanent magnets.
J. appl. Phys. **40,** 4029-4032, 1969 (No. 10). *E*

P. J. Buysman & G. A. L. Peersman: Stability of ceilings in a fluidized bed.
Proc. Int. Symp. on Fluidization, 1967, pp. 38-52. *E*

R. J. Carbone & W. J. Witteman: Vibrational energy transfer in CO_2 under laser conditions with and without water vapor.
IEEE J. Quantum Electronics **QE-5,** 442-447, 1969 (No. 9). *E*

H. B. G. Casimir: G. Holst: profile of a research director.
Science Journal **5A,** No. 1, 80-84, 1969. *E*

T. D. Clark & D. R. Tilley (Physics Dept., University of Essex, Colchester): Bulk mean field fluctuations in resistance of arrays of superconducting point contacts above T_c.
Physics Letters **29A,** 514-515, 1969 (No. 9). *E*

M. Davio & P. Piret: Les dérivées booléennes et leur application au diagnostic.
Rev. MBLE **12,** 63-76, 1969 (No. 3). *B*

Ph. Delsarte: A geometric approach to a class of cyclic codes.
J. combin. Theory **6,** 340-358, 1969 (No. 4). *B*

Ph. Delsarte & J. M. Goethals: Tri-weight codes and generalized Hadamard matrices.
Information and Control **15,** 196-206, 1969 (No. 2). *B*

R. Dessert: Applications aux hyperfréquences des diodes semiconductrices en régime d'avalanche.
Acta electronica **12,** 275-284, 1969 (No. 3). *L*

J. C. Diels & H. M. G. J. Trum: Gain measurements at 10.6 μ in pulse-generated CO_2 plasmas at high pressures.
Physics Letters **29A,** 697-698, 1969 (No. 11). *E*

W. F. Druyvesteyn & A. J. Smets: A technique for studying the anomalous penetration in the radio frequency size effect.
Physics Letters **30A,** 415-416, 1969 (No. 7). *E*

G. Engelsma: The influence of light of different spectral regions on the synthesis of phenolic compounds in gherkin seedlings, in relation to photomorphogenesis, VI. Phenol synthesis and photoperiodism.
Acta bot. neerl. **18,** 347-352, 1969 (No. 2). *E*

G. Engelsma: Low-temperature dependent development of phenylalanine ammonia-lyase in gherkin hypocotyls.
Naturwiss. **56,** 563, 1969 (No. 11). *E*

J.-F. Etaix: Etude du rayonnement lumineux émis lors d'une striction azimutale dans l'argon.
C. R. Acad. Sci. Paris **268B,** 230-233, 1969 (No. 3). *L*

R. Evrard & G.-A. Boutry (Conservatoire National des Arts et Métiers, Paris): An absolute micromanometer using diamagnetic levitation.
J. Vacuum Sci. Technol. **6,** 279-288, 1969 (No. 2). *L*

I. Flinn & P. J. Hulyer: Testing piezoelectric ceramics for power transducer applications.
Ultrasonics for Industry 1969 Conf. Papers, pp. 47-50. *M*

L. F. Gee, D. W. Parker & P. Swift: A M.O.S.T. store.
Information Processing 68, Proc. IFIP Congress, Edinburgh 1968, Vol. 2, pp. 778-782. *M*

R. J. Gelsing & K. van Steensel: A multilayer interconnexion system with gold beam leads.
Microelectronics and Reliability **8,** 325-329, 1969 (No. 4). *E*

Y. Genin: Théorie du contrôle optimal et calcul des variations.
Rev. MBLE **12,** 29-42, 1969 (No. 2). *B*

Y. Genin: Polynomial approximations in perturbational navigation and guidance schemes.
Advanced Problems of Mechanics for Space Flight Optimization, ed. B. Fraeys de Veubeke, Pergamon, Oxford 1969, pp. 13-24. *B*

A. H. Gomes de Mesquita & A. Bril: Preparation and cathodoluminescence of Ce^{3+}-activated yttrium silicates and some isostructural compounds.
Mat. Res. Bull. **4,** 643-650, 1969 (No. 9). *E*

H. C. de Graaff: Gate-controlled surface breakdown in silicon *p-n* junctions.
Philips Res. Repts. **25,** 21-32, 1970 (No. 1). *E*

C. A. A. J. Greebe & K. A. Ingebrigtsen (Norwegian Institute of Technology, Trondheim): On the electric coupling between mechanically transverse electroacoustic surface waves and an adjoining medium.
Physics Letters **30A,** 364-365, 1969 (No. 6). *E*

G. Groh & C. H. F. Velzel: Fehlerquellen bei der holographischen Bildvervielfachung.
Optik **30,** 257-272, 1969 (No. 3). *E, H*

W. van Haeringen & H.-G. Junginger: Empirical pseudopotential approach to the band structures of diamond and silicon carbide.
Solid State Comm. **7,** 1135-1137, 1969 (No. 16). *E, A*

W. van Haeringen & H.-G. Junginger: Pseudopotential approach to the energy band structure of graphite.
Solid State Comm. **7,** 1723-1725, 1969 (No. 23). *E, A*

S. H. Hagen & C. J. Kapteyns: The ionization energy of nitrogen donors in 6H and 15R SiC.
Philips Res. Repts. **25,** 1-7, 1970 (No. 1). *E*

C. M. Hargreaves: Anomalous radiative transfer between closely-spaced bodies.
Physics Letters **30A**, 491-492, 1969 (No. 9). *E*

J. C. M. Henning: Het meten van exchange met behulp van elektronenspin-resonantie.
Ned. T. Natuurk. **35**, 317-333, 1969 (No. 11). *E*

J. C. M. Henning, P. F. Bongers, H. van den Boom & A. B. Voermans: E.S.R. investigations on chromium doped $CdIn_2S_4$.
Physics Letters **30A**, 307-308, 1969 (No. 5). *E*

K. R. Hofmann: Stability criterion for Gunn oscillators with heavy surface loading.
Electronics Letters **5**, 469-470, 1969 (No. 20). *E*

F. N. Hooge, H. J. A. van Dijk & A. M. H. Hoppenbrouwers: $1/f$ noise in epitaxial silicon.
Philips Res. Repts. **25**, 81-86, 1970 (No. 2). *E*

F. N. Hooge & A. M. H. Hoppenbrouwers: Contact noise.
Physics Letters **29A**, 642-643, 1969 (No. 11). *E*

F. N. Hooge & A. M. H. Hoppenbrouwers: $1/f$ noise in continuous thin gold films.
Physica **45**, 386-392, 1969 (No. 3). *E*

A. M. H. Hoppenbrouwers & F. N. Hooge: $1/f$ noise of spreading resistances.
Philips Res. Repts. **25**, 69-80, 1970 (No. 2). *E*

K. Hoselitz & R. D. Nolan (University of Surrey, Guildford): Anisotropy-field distributions in barium ferrite micropowders.
Brit. J. appl. Phys. (J. Physics D), ser. 2, **2**, 1625-1633, 1969 (No. 12). *M*

B. B. van Iperen, H. Tjassens & J. J. Goedbloed: On the relation between microwave series resistance, capacitance, and output power of IMPATT diodes.
Proc. IEEE **57**, 1341-1342, 1969 (No. 7). *E*

A. Klopfer: Sauerstoffdesorption von Wolfram durch Elektronenstoß.
Vakuum-Technik **19**, 1-8, 1970 (No. 1/2). *A*

L. A. Ch. Koerts (Instituut voor Kernphysisch Onderzoek, Amsterdam): 20 jaar cyclotron.
Mikroniek **9**, 238-239, 1969 (No. 11).

W. G. Koster (Institute for Perception Research, Eindhoven): Translations of four articles by F. C. Donders (1818-1889) (preceded by a reprint of Sir William Bowman's "In Memoriam F. C. Donders"):
1) Official report of the ordinary meeting of the Royal Academy of Sciences, Department of Natural Sciences, on Saturday, 24 June, 1865;
2) On the speed of mental processes;
3) Two instruments for determining the time required for mental processes;
4) A short description of some instruments and apparatus belonging to the collection of the Physiological Laboratory and the Dutch Ophthalmic Hospital.
Acta psychol. **30**, 389-438, 1969.

E. Krätzig, K. Walther & W. Schilz: Investigation of superconducting phase transistions in Pb-films with acoustic surface waves.
Physics Letters **30A**, 411-412, 1969 (No. 7). *H*

W. Kwestroo, A. Huizing & J. de Jonge: The preparation of pure cadmium telluride and gallium phosphide using a generally applicable procedure.
Mat. Res. Bull. **4**, 817-824, 1969 (No. 11). *E*

J. Liebertz: Einkristallzüchtung von Paratellurit (TeO_2).
Kristall und Technik **4**, 221-225, 1969 (No. 2). *A*

J. Liebertz: Dimorphie von Lithiumjodat ($LiJO_3$).
Z. phys. Chemie Neue Folge **67**, 94-97, 1969 (No. 1-3). *A*

J. Liebertz: *P,T,f*-Werte von 4 n K_2CO_3-, Na_2CO_3- und NaOH-Lösungen unter hydrothermalen Bedingungen.
Chemie-Ing.-Technik **41**, 1231-1232, 1969 (No. 22). *A*

F. K. Lotgering: On the double-exchange interactions in $La_{1-x}Ba_xMnO_3$.
Philips Res. Repts. **25**, 8-16, 1970 (No. 1). *E*

F. K. Lotgering & G. H. A. M. van der Steen: Metal-deficient sulphospinels.
Solid State Comm. **7**, 1827-1829, 1969 (No. 24). *E*

M. H. van Maaren & H. B. Harland: An energy band model of Nb- and Ta-dichalcogenide superconductors.
Physics Letters **29A**, 571-573, 1969 (No. 9). *E*

M. H. van Maaren & H. B. Harland: Critical carrier concentration for superconductivity in a $CuRh_2Se_4$-based system.
Physics Letters **30A**, 204-205, 1969 (No. 3). *E*

F. A. M. M. van Meel, A. C. van Maaren & G. J. van Weezel (Philips Lighting Division, Eindhoven): The influence of annealing on the recrystallization behaviour and the mechanical properties of doped and undoped molybdenum wire.
High Temperature Materials, 6th Plansee Seminar, Reutte/Tyrol 1968, pp. 172-181.

R. J. Meijer: Rebirth of the Stirling engine.
Science Journal **5A**, No. 2, 31-37, 1969. *E*

F. Meyer: Plane specificity in the reaction of methanol and ethanol vapor with clean germanium.
J. phys. Chem. **73**, 3844-3848, 1969 (No. 11). *E*

F. Meyer & G. A. Bootsma: Ellipsometric investigation of chemisorption on clean silicon (111) and (100) surfaces.
Surface Sci. **16**, 221-233, 1969. *E*

F. Meyer & G. A. Bootsma: Ellipsometrie.
Ned. T. Natuurk. **35**, 154-160, 1969 (No. 5/6). *E*

J. Michel, R. Petit & A. Semichon: Elaboration des diodes à avalanche.
Acta electronica **12**, 255-273, 1969 (No. 3). *L*

R. F. Mitchell, W. Willis & M. Redwood (Queen Mary College, London): Electrode interactions in acoustic surface-wave transducers.
Electronics Letters **5**, 456-457, 1969 (No. 19). *M*

K. Mouthaan: Nonlinear characteristics and two-frequency operation of the avalanche transit-time oscillator.
Philips Res. Repts. **25**, 33-67, 1970 (No. 1). *E*

K. Mouthaan & H. P. M. Rupert: Second-harmonic tuning of the avalanche transit-time oscillator.
Proc. IEEE **57**, 1449-1450, 1969 (No. 8). *E*

K. Mulder, R. van Dantzig, J. E. J. Oberski & L. A. Ch. Koerts (Instituut voor Kernphysisch Onderzoek, Amsterdam): Constructie van de BOL strooiingskamer.
Mikroniek **9**, 240-245, 1969 (No. 11).

P. H. Müller & P. Knoop: L'analyse de verres par spectrométrie d'absorption atomique.
Silicates industr. **34**, 325-330, 1969 (No. 12). *E*

J. A. van Nielen: A simple and accurate approximation to the high-frequency characteristics of insulated-gate field-effect transistors.
Solid-State Electronics **12**, 826-829, 1969 (No. 10). *E*

W. C. Nieuwpoort & R. Bleekrode: On intensity alterations in C_2 "Swan" emission spectra.
J. chem. Phys. **51**, 2051-2055, 1969 (No. 5). *E*

S. G. Nooteboom (Institute for Perception Research, Eindhoven): The tongue slips into patterns.
Nomen, Leyden studies in linguistics and phonetics, 1969, pp. 114-132.

J. M. Noothoven van Goor: Impurity scattering in pure bismuth.
Physics Letters **29A**, 685-686, 1969 (No. 11). *E*

D. P. Oosthoek (Instituut voor Kernphysisch Onderzoek, Amsterdam): Precisie-opdamp-draadmasker.
Mikroniek **9**, 246-247, 1969 (No. 11).

A. van Oostrom: Application of a channel plate in field-ion microscopy.
Philips Res. Repts. **25**, 87-94, 1970 (No. 2). *E*

L. J. van der Pauw: Influence of diffraction on the four-pole impedance of electroacoustic delay lines with circular or rectangular plane-parallel transducers.
J. Acoust. Soc. Amer. **46**, 497-507, 1969 (No. 3, Part 1). *E*

L. J. van der Pauw: Four-pole properties of an electroacoustic delay line with plane-concave transducers.
J. Acoust. Soc. Amer. **46**, 508-516, 1969 (No. 3, Part 1). *E*

L. J. van de Polder: Beam-discharge lag in a television pick-up tube.
Adv. in Electronics and Electron Phys. **28A**, 237-245, 1969. *E*

R. G. Pratt & T. C. Lim (Imperial College of Science and Technology, London): Acoustic surface waves on silicon.
Appl. Phys. Letters **15**, 403-405, 1969 (No. 12). *M*

Eleanor D. Pyrah (Associated Semiconductor Manufacturers Ltd., Wembley, England): Negative resistance and the development of the Gunn effect oscillator.
Physics Education **4**, 333-341, 1969 (No. 6).

A. Rabenau & H. Rau: Über Sulfidhalogenide des Bleis und das Pb_4SeBr_6.
Z. anorg. allgem. Chemie **369**, 295-305, 1969 (No. 3-6). *A*

J. E. Ralph & M. G. Townsend (Electricity Council Research Centre, Capenhurst, Chester, Cheshire): Fluorescence and absorption spectra of Ni^{2+} in MgO.
J. Physics C **3**, 8-18, 1970 (No. 1). *M*

P. Reijnen: Non-stoichiometry and sintering of ionic solids.
Reactivity of Solids, Proc. 6th Int. Symp., Schenectady 1968, pp. 99-114. *E*

J. G. Rensen & J. S. van Wieringen: Anisotropic Mössbauer fraction and crystal structure of $BaFe_{12}O_{19}$.
Solid State Comm. **7**, 1139-1141, 1969 (No. 16). *E*

J. Roos: Electrets, semipermanently charged capacitors.
J. appl. Phys. **40**, 3135-3139, 1969 (No. 8). *E*

J. H. T. van Roosmalen: Adjustable saturation in a pick-up tube with linear light transfer characteristic.
Adv. in Electronics and Electron Phys. **28A**, 281-288, 1969. *E*

C. J. M. Rooymans: Chemical processes in high-pressure systems.
Reactivity of Solids, Proc. 6th Int. Symp., Schenectady 1968, pp. 743-761. *E*

F. L. J. Sangster & K. Teer: Bucket-brigade electronics — new possibilities for delay, time-axis conversion, and scanning.
IEEE J. Solid-State Circuits **SC-4**, 131-136, 1969 (No. 3). *E*

P. Schagen & A. A. Turnbull: New approaches to photoemission at long wavelengths.
Adv. in Electronics and Electron Phys. **28A**, 393-398, 1969. *M*

J. J. Scheer & J. van Laar: The influence of cesium adsorption on surface Fermi level position in gallium arsenide.
Surface Sci. **18**, 130-139, 1969 (No. 1). *E*

J. C. Schepers: De glasexpositie op de Glastechnische Dag 1969 te Petten, verzorgd door de Glastechnische Afdeling van het Philips Natuurkundig Laboratorium te Waalre.
Glastechn. Meded. **7**, 147-159, 1969 (No. 4). *E*

H. J. Schmitt & M. Lemke: Miniaturisierte Bauelemente in Streifenleitertechnik.
Int. elektron. Rdsch. **23,** 225-229, 270-272, 1969 (Nos. 9, 10). *H*

J. F. Schouten (Institute for Perception Research, Eindhoven): Perils of perception in radiant radiology.
12th Int. Congress of Radiology, Tokyo 1969; Tokyo Monitor 1969, p. 42.

J. Schröder & F. J. Sieben: Bildungsenthalpie von Wolframhexafluorid und Wolframpentafluorid.
Chem. Berichte **103,** 76-81, 1970 (No. 1). *A*

H. Schweppe: Electromechanical properties of bismuth germanate $Bi_4(GeO_4)_3$.
IEEE Trans. **SU-16,** 219, 1969 (No. 4). *A*

P. J. Severin: Superphase velocity effects caused by a single electron in a ferrite.
Physica **45,** 253-256, 1969 (No. 2). *E*

J. M. Shannon, J. Stephen (U.K. Atomic Energy Authority, Harwell, England) **& J. H. Freeman** (U.K. At. En. Auth., Harwell): MOS frequency soars with ion-implanted layers.
Electronics **42,** No. 3, 96, 98-100, 1969. *M*

J. G. Siekman: Nieuwe ontwikkelingen in de microbewerking, I. Microbewerkingen met gefocusseerde deeltjesbundels.
Ingenieur **81,** O 105-112, 1969 (No. 39). *E*

I. H. Slis & A. Cohen (Institute for Perception Research, Eindhoven): On the complex regulating the voiced-voiceless distinction, I, II.
Language and Speech **12,** 80-102, 137-155, 1969 (Nos. 2, 3).

F. A. Staas & A. P. Severijns: Vorticity in He II and its application in a cooling device.
Cryogenics **9,** 422-426, 1969 (No. 6). *E*

A. L. N. Stevels: Phase transitions in nickel and copper selenides and tellurides.
Thesis, Groningen 1969. *E*

T. G. W. Stijntjes, J. Klerk (both with Philips Electronic Components and Materials Division, Eindhoven) **& A. Broese van Groenou:** Permeability and conductivity of Ti-substituted MnZn ferrites.
Philips Res. Repts. **25,** 95-107, 1970 (No. 2). *E*

T. L. Tansley: Opto-electronic properties of heterojunctions — a review.
Opto-electronics **1,** 143-150, 1969 (No. 3). *M*

T. L. Tansley & J. E. Ralph: Analogue conductivity switch.
Electronics Letters **5,** 671, 1969 (No. 26). *M*

D. G. Taylor, C. H. Petley & K. G. Freeman: Television at low light-levels by coupling an image intensifier to a "Plumbicon".
Adv. in Electronics and Electron Phys. **28B,** 837-849, 1969. *M*

H. Tazieff & M. Jatteau: Mesure dans l'infrarouge des paramètres physiques des gaz éruptifs.
C. R. Acad. Sci. Paris **268D,** 767-770, 1969 (No. 5). *L*

F. D. Tisi: Schnelle Datenübertragung in Kanälen mit grosser Frequenzverwerfung.
Thesis, Zürich 1969. *E*

P. J. Turner, P. Cartwright, M. J. Southon (all with Department of Metallurgy, University of Cambridge), **A. van Oostrom & B. W. Manley:** Use of a channelled image intensifier in the field-ion microscope.
J. sci. Instr. (J. Physics E), ser. 2, **2,** 731-733, 1969 (No. 8). *E, M*

J. G. Verhagen & A. Liefkens: Gasbescherming bij het gasbooglassen.
Lastechniek **35,** 241-247, 1969 (No. 12). *E*

L. Verhoeven: Demodulation of PAL signals with two uncritical delay lines.
Electronics Letters **5,** 344-345, 1969 (No. 15). *E*

A. G. van Vijfeijken, A. Walraven & F. A. Staas: Energy and stability of vortex rings in liquid helium II; critical velocities.
Physica **44,** 415-436, 1969 (No. 3). *E*

A. T. Vink, A. J. Bosman, J. A. W. van der Does de Bye & R. C. Peters: Low temperature luminescence in GaP at very low excitation densities.
Solid State Comm. **7,** 1475-1481, 1969 (No. 20). *E*

M. T. Vlaardingerbroek: Output spectrum of IMPATT-diode oscillators.
Electronics Letters **5,** 521-522, 1969 (No. 21). *E*

A. G. C. Vogel (Instituut voor Kernphysisch Onderzoek, Amsterdam): Aandrijving voor de energieselectiespleet.
Mikroniek **9,** 252-253, 1969 (No. 11).

J. H. N. van Vucht, F. A. Kuijpers & H. C. A. M. Bruning: Reversible room-temperature absorption of large quantities of hydrogen by intermetallic compounds.
Philips Res. Repts. **25,** 133-140, 1970 (No. 2). *E*

W. L. Wanmaker, J. W. ter Vrugt & J. G. Verlijsdonk (Philips Lighting Division, Eindhoven): Luminescence of Mn^{2+}-activated spinels in the MgO-Li_2O-ZnO-Ga_2O_3-Al_2O_3 system.
Philips Res. Repts. **25,** 108-117, 1970 (No. 2).

C. H. Weijsenfeld: Empirical formula for the measured Hall angles of dirty type II superconductors.
Physica **45,** 241-252, 1969 (No. 2). *E*

K. Weiss: Die Löslichkeit von Silber in Silberbrômid.
Z. phys. Chemie Neue Folge **67,** 86-93, 1969 (No. 1-3). *E*

K. Weiss: Die Überführungswärme der Kationen in α-Ag_2S.
Berichte Bunsenges. phys. Chemie **73,** 683-690, 1969 (No. 7). *E*

K. Weiss: Die Überführungswärme von Frenkel-Defekten in AgBr.
Berichte Bunsenges. phys. Chemie **73,** 690-696, 1969 (No. 7). *E*

H. W. Werner: Investigation of solids by means of an ion-bombardment mass spectrometer.
Developm. appl. Spectroscopy **7A,** 239-266, 1969. *E*

J. S. C. Wessels: Isolation and properties of two digitonin-soluble pigment-protein complexes from spinach.
Progress in Photosynthesis Research, Vol. I, pp. 128-136, 1969. *E*

G. Winkler & P. Hansen: Calcium-vanadium-indium substituted yttrium-iron-garnets with very low line-widths of ferrimagnetic resonance.
Mat. Res. Bull. **4,** 825-837, 1969 (No. 11). *H*

S. Wittekoek & P. F. Bongers: Observation of Cr^{3+} absorption bands in $CdIn_2S_4$ (Cr) and the consequences for the interpretation of the absorption edge of $CdCr_2S_4$.
Solid State Comm. **7,** 1719-1722, 1969 (No. 23). *E*

J. P. Woerdman & B. Bölger: Diffraction of light by a laser induced grating in Si.
Physics Letters **30A,** 164-165, 1969 (No. 3). *E*

Contents of Philips Telecommunication Review **29,** No. 1, 1970:

C. M. de Zeeuw: HF communication systems (pp. 1-10).
E. Timmermans: New mechanical design for telephone transmission equipment (pp. 11-22).
C. Ziekman: The 8TR 601 pulse code modulation system (pp. 23-31).
G. Lind: Measurement of sea clutter correlation with frequency agility and fixed frequency radar (pp. 32-38).

Contents of Electronic Applications **29,** No. 2, 1969:

G. van Dijk & G. Wolf: Recent developments in circuits and transistors for television receivers: X. A mixer transistor for v.h.f., XI. Coupling a v.h.f. mixer transistor to a double-tuned band-pass filter (pp. 39-46, 47-55).
The 7A65X: a shadow-mask picture tube with 110° deflection (pp. 56-57).
F. T. Backers: An ultrasonic delay line for chrominance signal correction (pp. 59-70).
A counter tube for radiation dosimetry (p. 71).

Contents of Mullard Technical Communications **11,** No. 104, 1970:

J. Merrett: Practical transient suppression circuits for thyristor power-control systems (pp. 82-88).
P. Bissmire: Simple a.f.c. system for 625-line tv receivers (pp. 89-93).
K. W. Stanley: Reliability and stability of metal film resistors (pp. 94-100).

Contents of Mullard Technical Communications **11,** No. 105, 1970:

B. E. Attwood & B. J. Simpson: Transformerless field timebase for 90° 625-line colour (pp. 102-109).
Ratings of Mullard carbon film resistors (pp. 110-112).
J. M. Lavallee & J. Merrett: The basis of a design for a reversible drive with regenerative braking (pp. 113-120).

Contents of Valvo Berichte **15,** No. 4, 1969:

E. Pech: Ein integrierter Synchrondemodulator für Farbfernseh-Empfänger (pp. 109-121).
H. H. Feindt: Die Entwicklung eines integrierten RGB-Verstärkers für Farbfernseh-Empfänger (pp. 122-138).

Contents of Valvo Berichte **15,** No. 5, 1969:

J. Koch: Berechnung eines mit Ticonal 750 aufgebauten Lautsprecher-Kernmagnetsystems (pp. 139-155).
J. Brambring: Elektronenoptische Untersuchungen an Elektronenkanonen von Fernsehbildröhren (pp. 156-167).

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No. 3

In Memoriam Ir. S. Gradstein

On 21st September 1970, barely six months after his retirement as Editor-in-chief of our journal, Ir. Stephan Gradstein died, after a short illness, in Eindhoven.

With the death of Stephan Gradstein we lose the last man to be closely associated with the Review right from the start and through the whole 34 years of its existence. In the period of almost twenty years during which he was Editor-in-chief it was his achievement, building on the foundations laid by Holst and Oosterhuis, to develop the Review into what it had become upon his retirement.

The success that he achieved in this was due to his very special combination of talents and qualities of character. He had an exceptional feeling for language and was able to indicate the best method of treating even the most unyielding matter, but he also had pictorial ability of a high order and he set himself and others high standards.

We, who worked with him and learned so much from him, lose in Stephan an example difficult to match and above all a good friend.

J. W. Miltenburg

Ferrite-cored kicker magnets

H. O'Hanlon

In the CERN Proton Synchrotron twenty bunches of protons, after being accelerated, circulate with the speed of light. The development of kicker magnets has made it possible to select one or more of the bunches, and kick it out of the ring while leaving the other bunches unperturbed. This article on the subject, kindly written for us by Mr. H. O'Hanlon of CERN, seemed apt since it is an interesting example of the application of ferrites in proton synchrotrons, and is quite different from another application recently treated in our Review. When a material is used in a new way some previously unexceptional feature may take on a special importance. In the present case it turns out that ferrite is a clean material that is compatible with ultra-high-vacuum practice.

In high-energy particle accelerator techniques the injection of the particles into the accelerator before acceleration and the ejection from it after acceleration have always presented a delicate problem. Today kicker magnets are so widely used for these purposes that they can be considered as vital for the progress of high-energy nuclear physics. Ideally the kicker magnet provides a rectangular pulse of magnetic flux, deflecting the particle beam in an accurately defined way during a well defined short period of time.

H. O'Hanlon, M. Sc. is with the Intersecting Storage Rings Division, CERN, Geneva.

The wide use of kicker magnets — now and in the near future — is illustrated in *fig. 1*, which shows the accelerator complex at CERN. This complex consists of the proton synchrotron *CPS*, in which protons can be accelerated to nearly 30 GeV, and the intersecting storage rings *ISR*, in which protons circulating in opposite directions will be stored. (The purpose is to cause head-on collisions between protons at the intersections of the two rings.) The proton synchrotron has been in use for high-energy experiments since 1959 [1]. In the near future protons will be injected into it from a 800 MeV booster synchrotron (*PSB* in fig. 1), which is expected to come into operation in 1972.

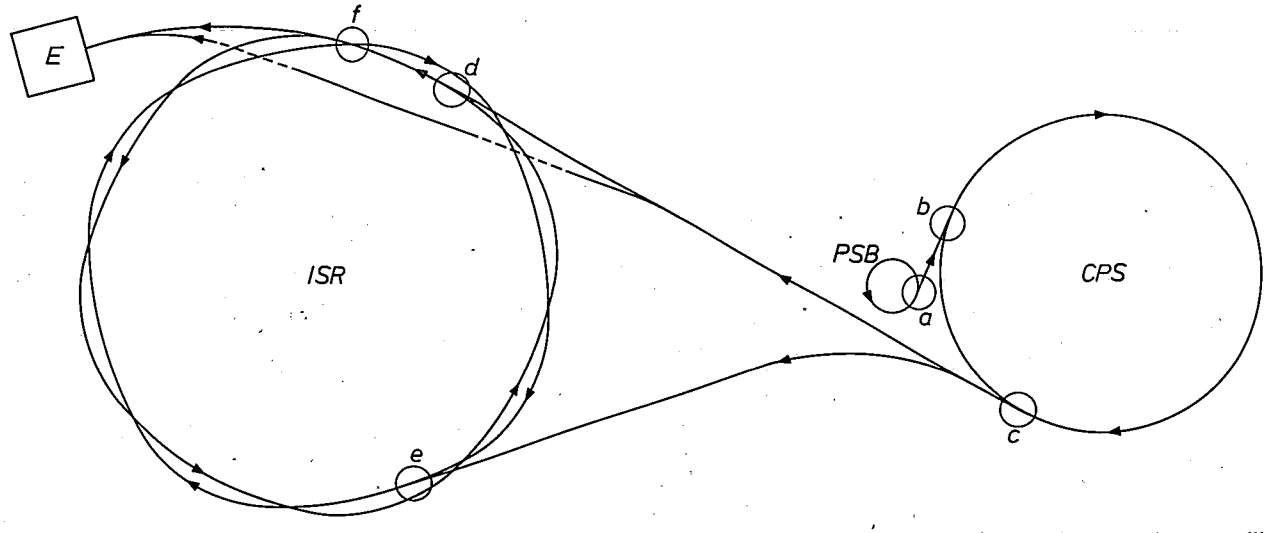


Fig. 1. The accelerator complex at CERN, now nearing completion. Protons, preaccelerated in the proton synchrotron booster *PSB*, will be accelerated to 28 GeV in the CERN proton synchrotron *CPS*. Protons from many synchrotron cycles will be stacked in the two intersecting storage rings *ISR*; at the intersections of the storage rings beam-collision experiments will be performed. *E* experimental hall. The diameter of the CPS ring is 200 m, the diameter of the ISR is 300 m. Fast kicker magnets are or will be used in this complex for (*a*) ejection from the booster, (*b*) injection into and (*c*) ejection from the main synchrotron, and (*d*, *e*) injection into and (*f*) ejection from the intersecting storage rings. The proton synchrotron *CPS* has been in use since 1959; the first circulation of a beam in the storage rings is expected at the beginning of 1971[*]; the 800 MeV booster is expected to come into operation in 1972. At present injection into the main synchrotron is direct from a 50 MeV linear accelerator [1].

The storage rings are nearing completion; the first injection of a beam is expected to take place at the beginning of 1971[*]. In fig. 1 *a-f* indicate the zones where kicker magnets are, or will be, used for switching the protons from one orbit into another.

Fig. 2 shows the way in which kicker magnets are employed in a synchrotron, where the particles in the beam are bunched. This bunching is a result of the acceleration by r.f. electric fields. In fig. 2 it is

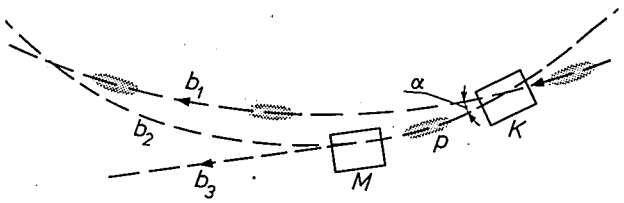


Fig. 2. The use of a fast kicker (*K*) for ejection from a synchrotron. b_1 proton equilibrium orbit in the synchrotron. *K* is energized during the passage of one bunch of protons *p*, which is deflected by a small angle α into the orbit b_2 oscillating about b_1. A d.c. bending magnet *M* deflects it further away from the synchrotron guide fields. b_3 ejected beam.

assumed that one bunch has to be ejected. During the passage of this bunch the kicker magnet *K* is excited, giving a constant field that deflects the bunch from the equilibrium orbit; before and after the passage the field is zero. The deflection, though small, is sufficient to cause the bunch to enter the aperture of a strong d.c. bending magnet *M* that deflects the bunch further away from the synchrotron guide fields. More bunches can be deflected by applying a longer pulse to the kicker magnet. For the *injection* of a bunched beam into an accelerator or storage ring the reverse order is followed.

The required rise and decay time of the magnetic field pulse depend upon the function of the kicker magnet. Ejection from a synchrotron, for instance, requires a discrete number of complete bunches to be deflected, and rise and decay time must be shorter than the separating period between consecutive bunches, which is typically 50-200 ns. In the case of a homogeneous (non-bunched) beam (for example, a beam stacked in a storage ring) rise and decay time determine the efficiency of the ejection. Particles passing the magnet during the rise or decay of the pulse are partially deflected; they neither remain in the equilibrium orbit nor enter the bending magnet aperture and are lost.

Kicker magnets in existing proton synchrotrons usually have a low repetition rate: one pulse every one to three seconds is typical. In electron synchrotrons they operate much faster, for instance at 60 pulses per second.

A rise time of about 100 ns implies frequency components of about 10 MHz in the magnetic-field pulse. The magnet, and in particular the magnet core, must function properly at these frequencies. Ferrites, well known as magnetic materials for other high-frequency applications [2], have also been found to be suitable as core material for kicker magnets, as we shall explain below.

Generation of a rectangular magnetic field pulse

The ideal way to create a rectangular pulse of magnetic field was first indicated by G. K. O'Neill [3]. The magnet is built in the form of a *delay line* or transmission line by connecting suitable capacitors across its single magnetizing winding which consists of a central conductor and a return lead (*fig. 3*). Such a delay line acts as a low-pass filter for electrical signals. Ideally it has a flat response for frequencies well below a certain cut-off frequency. If the cut-off

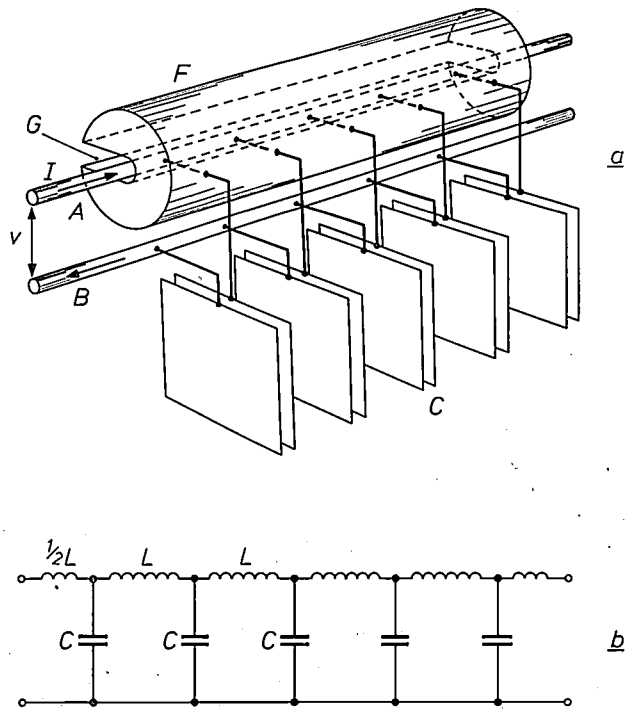


Fig. 3. *a*) Delay-line kicker magnet, schematic. *F* ferrite core. *G* gap. The core is magnetized by the current *I* in a central conductor *A* and a return lead *B*. The delay-line characteristics are adapted to the requirements by the capacitors *C* between central and return conductor. *b*) Equivalent circuit of the delay line.

[*] Tests have now been run successfully in which beams have circulated in both storage rings; the magnets of figs. 8-11 have performed according to their design specifications. (*Ed.*)

[1] For more information about the CERN proton synchrotron (and four other proton synchrotrons), see R. Gouiran, Five major proton synchrotrons, Philips tech. Rev. 30, 330-365, 1969 (No. 11/12).

[2] Another application of ferrites in synchrotrons has recently been described in this Review by F. G. Brockman, H. van der Heide and M. W. Louwerse, Ferroxcube for proton synchrotrons, Philips tech. Rev. 30, 312-329, 1969 (No. 11/12).

[3] G. K. O'Neill, Proc. Int. Conf. on High-Energy Accelerators and Instrumentation, CERN 1959, p. 125.

frequency is sufficiently high it can propagate electrical signals with no measurable distortion. For such distortionless propagation to take place in practice, input and output must be matched to the characteristic impedance of the delay line.

If a rectangular electrical pulse is propagated through the delay-line magnet, a constant magnetic field is present throughout the magnet gap during the time that the pulse completely fills the delay line (*fig. 4*). During the passage of the leading edge of the electrical pulse the flux is building up, and it is removed when the trailing edge passes. It follows that a substantially rectangular magnetic-flux pulse is created if the velocity of propagation of the electrical pulse is high.

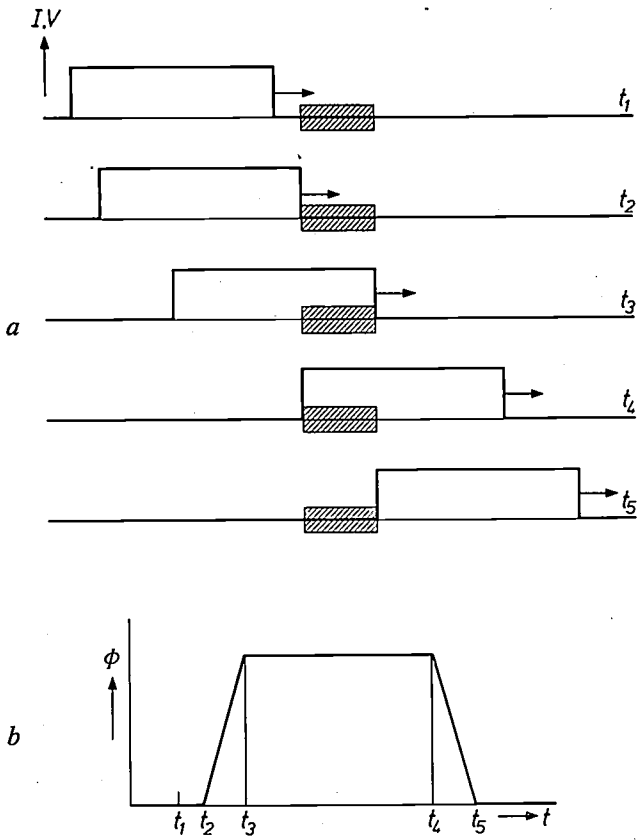


Fig. 4. *a*) A rectangular electrical pulse propagating in a long transmission line, part of which consists of the delay-line magnet (hatched). At each position along the gap the magnetic field is proportional to the current *I*. *b*) Total flux Φ in the magnet as a function of time. t_1, t_2, ... times corresponding to (*a*). The flux pulse is substantially rectangular if the propagation time of electrical signals through the magnet (t_3—t_2 and t_5—t_4) is small.

The complete circuit used to create the magnetic flux pulse is shown schematically in *fig. 5*. The input of the delay-line magnet *K* is connected to a network that forms the rectangular electrical pulse. This network includes a delay line *D* which is usually a coaxial cable. While the switch *S* is open, it is charged to a voltage V_0 by a high-voltage source *G* through a high resistance R_c. The output of *K* is connected to a terminating load R_t. *D* and R_t are matched to the characteristic impedance *Z* of *K*.

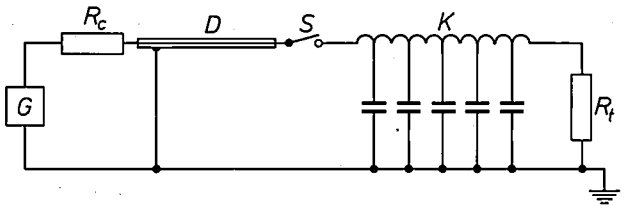


Fig. 5. Circuit for generating a magnetic flux pulse. *K* delay-line magnet. *D* coaxial delay line. R_t terminating resistor. *D* and R_t are matched to *K*. While the switch *S* is open, *D* is charged through a high resistance R_c by the source *G*. On closing *S*, *D* discharges through *K*, giving the required pulse.

When the switch *S* is closed *D* and *K* together form a transmission line that is effectively open-circuited at one end (because $R_c \gg Z$), and matched at the other (because $R_t = Z$). The fundamental solutions for a transmission line show that waves propagate forwards or backwards with a propagation constant characteristic of the line. Consequently when *S* is closed the stationary rectangular voltage on *D* will separate into two rectangular components of half the amplitude, one travelling forwards and the other backwards. This backward component is then reflected at the open-circuited end and travels forward, following the first component. The process is sketched in *fig. 6*. It follows that the electrical pulse propagating through *K* is of voltage $V = \frac{1}{2}V_0$, and its electrical length is twice the electrical length of *D*.

The length of the flat top of the magnetic flux pulse can be made variable, if so desired, by intro-

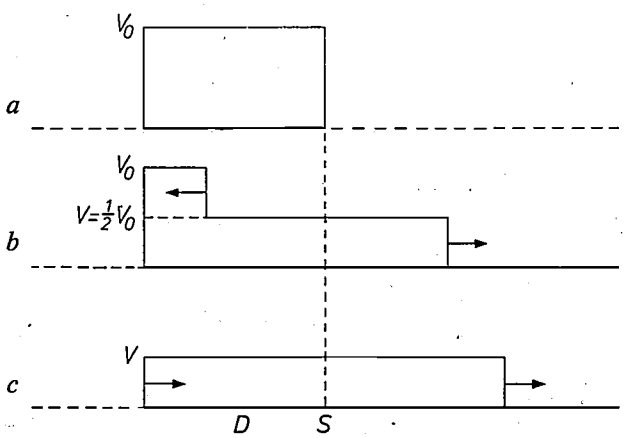


Fig. 6. Discharge of a coaxial cable (*D* in fig. 5). When the switch (*S* in fig. 5) is closed, the rectangular voltage on *D* (*a*) separates into two rectangular pulses of half the original amplitude, one travelling forwards, the other backwards. The backward-travelling pulse is reflected at the open-circuited end and links up with the forward-travelling pulse (*b*). In (*c*) the trailing edge of the backward-travelling pulse has just been reflected.

ducing switches to direct part of the electrical pulse into another matched load. This can be used to vary the number of proton bunches that are deflected.

Design of a kicker magnet

We shall outline some quantitative considerations that enter into the design of a kicker magnet of the type shown in fig. 3. First we estimate the voltages and currents that are required to build up a given "kick strength" (a concept defined below) in a magnet of given dimensions within a given time.

Let the length of the magnet gap be l, the width w and the height h. The core represents a magnetic short-circuit if its permeability is sufficiently high, and all of the field induced by the current appears across the gap. Thus, if there is a single winding only, as in fig. 3:

$$I = hH . \quad \ldots\ldots\ldots \quad (1)$$

The field H is required to deflect particles of charge e and momentum p over a certain angle α. If ϱ is the radius of curvature of the particle orbit in the magnet, we have $\alpha = l/\varrho$. Combining this with the relation [4] $p = \varrho eB$, where B is the flux density ($B = \mu_0 H$), we find that the product Bl must have the value:

$$Bl = \alpha p/e . \quad \ldots\ldots\ldots \quad (2)$$

Bl is by definition the "kick strength" of the magnet [5].

The dimensions w and h are determined by the cross-section of the beam if the magnet is contained within the vacuum chamber or by the cross-section of the vacuum chamber if the magnet is external. This is discussed later. For a given kick strength Bl, the length l of the magnet can in principle be traded for magnetic flux density B. Usually l is made as large as the available space permits in order to reduce the current required (see eq. 1), which in practice turns out to be high. Consider a simple but not unrealistic example. A 30 GeV beam is to be deflected over 2 mrad in a gap of dimensions $l = 1$ m, $h = 2$ cm, $w = 4$ cm. We then have $pc \approx 30$ GeV (where $c = 3\times10^8$ ms^{-1} is the velocity of electromagnetic radiation), $p/e = 30$ GV/3×10^8 ms^{-1} $= 100$ Vs/m. $\alpha = 2\times10^{-3}$, so that $Bl = 0.2$ Vs/m and $B = 0.2$ Vs/m^2. Hence $H = B/\mu_0 = 16\times10^4$ A/m. The required current in this case is $I = hH = 3200$ A.

In this simplified treatment the total inductance L_s of the magnet is determined solely by the dimensions of the gap, since $L_s = \Phi/I$, where Φ, the total flux in the gap, is equal to lwB. Using eq. (1) we obtain:

$$L_s = \mu_0 lw/h .$$

The required voltage is directly related to the required rate of rise of the field, as can be seen by considering the delay-line characteristics of the magnet. A magnet like the one of fig. 3 is closely related to a coaxial delay line. If the extra capacitors were omitted and the outer conductor distributed over the circumference it would be very similar to a coaxial line. The extra capacitors needed to optimize the design result in a network that is more suitably approximated by a "lumped" delay line, with n lumped inductances L ($L_s = nL$), each connected to earth by a lumped capacitor C. Such a line is characterized by a *cut-off frequency* $f_c = 1/\pi\sqrt{LC}$, and, for electromagnetic waves with Fourier components of frequency well below f_c, by a *time of propagation* $t_a = n\sqrt{LC}$, and a *characteristic impedance* $Z = \sqrt{L/C}$. Multiplying t_a and Z we obtain $Zt_a = nL = L_s$. The voltage V is therefore $V = ZI = L_sI/t_a$, or:

$$V = \Phi/t_a . \quad \ldots\ldots \quad (3)$$

The required voltage on the line is therefore proportional to the required flux and to the required rate of rise of the flux. In the example given above we have $\Phi = lwB = 8\times10^{-3}$ Vs. If this flux is to be built up within 100 ns, the required voltage is 80 kV. The impedance of the line would be $Z = V/I = 25\ \Omega$.

The total capacitance $C_s = nC$ that must be connected to the magnet is found by dividing t_a by Z:

$$C_s = t_a/Z = t_a^2/L_s .$$

It has been assumed above that a perfectly rectangular pulse propagates through the delay line undistorted (see fig. 4). If this were true the rise and decay time of the magnetic flux would be equal to the time of propagation t_a of the leading and trailing edge of the electrical pulse through the magnet. In fact the rise and decay time of the flux are longer because in the first place the input pulse itself has a non-zero rise (and decay) time t_b, and secondly the leading and trailing edge of the electrical pulse are distorted while passing the delay-line magnet. The latter effect is characterized by a "pulse degeneration time" t_c. These spreading effects add a term $\sqrt{t_b^2 + t_c^2}$ to the flux rise time, giving an approximate total rise time of:

$$\tau = t_a + \sqrt{t_b^2 + t_c^2} . \quad \ldots\ldots \quad (4)$$

The rise time t_b of the electrical pulse is in practice determined by the switch S in fig. 5; we shall return

[4] A derivation of this well-known relation can be found in reference [1], fig. 8 (p. 336).

[5] More precisely, the kick strength is defined as the integral of the flux density B of the magnet along the path of the particles, $\int B\,dl$. This is equal to the product Bl if B is homogeneous along the length of the magnet gap and if fringe fields are zero.

to this later. It can easily become comparable to t_a.

The pulse degeneration time t_c has been shown emprically [6] to be

$$t_c \approx (1.1/n)^{2/3}\, t_a .$$

In practice t_c is of little significance for the flux rise time if the number of sections n is greater than 10. In the ideal case of a continuous lossless line ($n \to \infty$, L and $C \to 0$ such that Z and t_a remain constant) there would be no distortion of a pulse passing the delay line ($f_c \to \infty$, $t_c \to 0$).

Kicker-magnet types; two examples

The useful lifetime of a beam of charged particles is limited by the vacuum pressure in the vacuum chamber. In proton synchrotrons the lifetime must be of the order of seconds, and the vacuum is in the range 10^{-5} to 10^{-7} torr. In the storage rings the beam can have a lifetime as long as a day and the vacuum must be in the range 10^{-9} to 10^{-10} torr. When designing a kicker magnet consistent with these vacuum requirements one is faced with the choice of building the magnet either *in* or *around* the vacuum chamber.

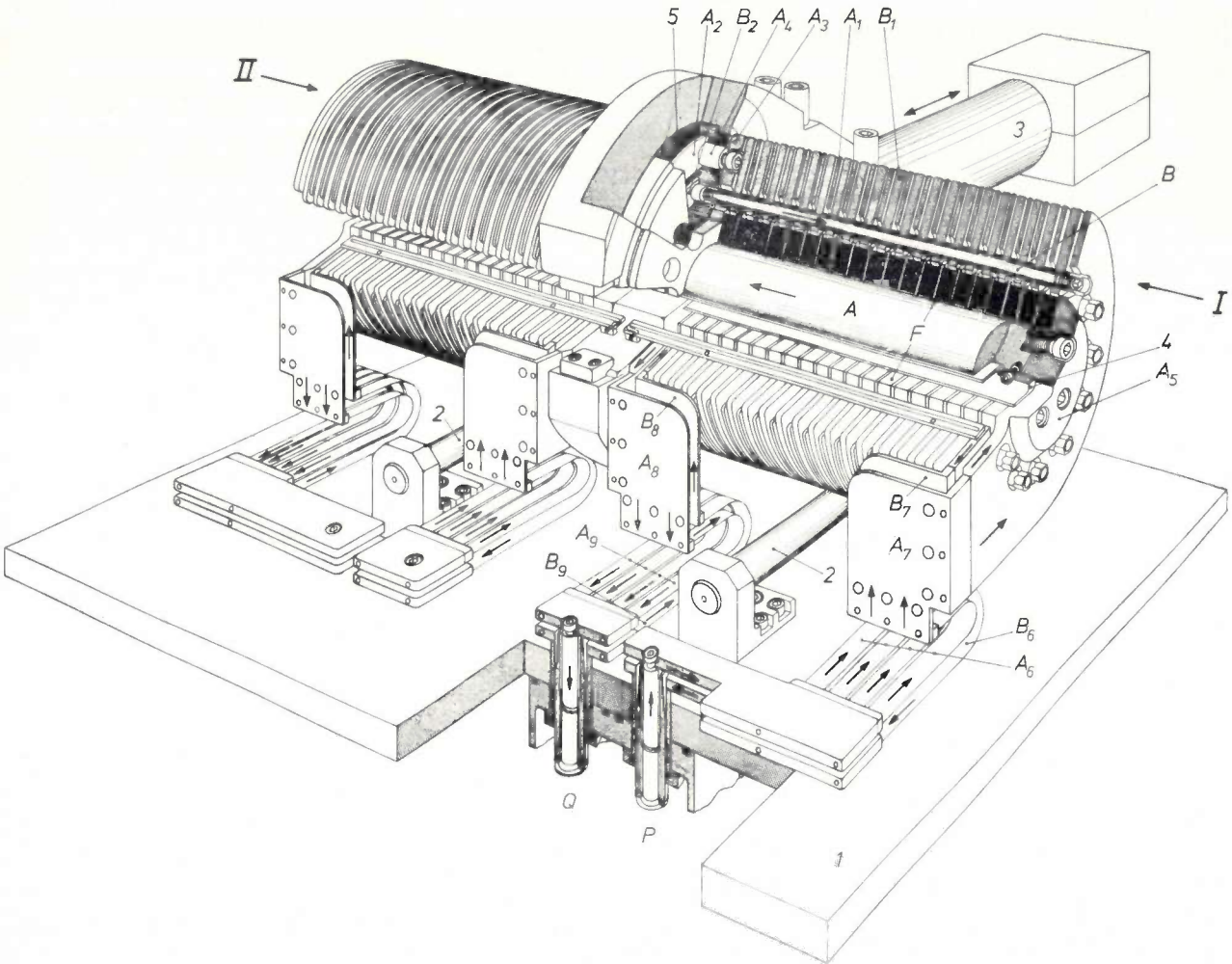


Fig. 7. Cut-away perspective view of the CERN proton synchrotron ejection magnet. The magnet, being of the small-aperture type, operates completely in vacuum and is mechanically shifted into and out of its working position each cycle. The vacuum chamber is locally enlarged to a tank (*1* bottom of the tank). The magnet rests on a carriage on rails *2*, and is connected by a shaft *3* to a hydraulic activator outside the tank.
The magnet consists of two symmetrical parts *I* and *II*. Each of these consists of a central conductor *A* with large thin capacitor plates A_1 connected to it, and a set of eleven return leads *B* with small thick capacitor plates B_1. The core of the magnet is made of ferrite rings *F* carried by the central conductor *A*; they are separated by the capacitor plates A_1 and aligned at the gap by the shims *4*. The central conductor and the return lead, each with the capacitor plates connected to it, form two interlaced but mechanically independent and electrically isolated structures; they are both supported only from the central plate *5*. This plate, made of epoxy resin reinforced with glass fibre, is mounted to the shaft *3*. A steel ring A_2 is cast into the plate *5*, and the central conductor *A* is mounted to it by a flange A_3 and studs A_4. Another steel ring B_2 in the central plate carries the return lead structure *B*. The capacitor plates and the ferrite rings form a stack held together by the flange A_5 bolted to the central conductor *A*. The capacitor plates are of polished aluminium.

Entry and exit of the delay-line system *A-B* is by the coaxial traversals *P* and *Q* in the bottom of the vacuum tank. A current pulse entering at *P* successively traverses the flexible parallel strip conductors A_6 and B_6, the endplates A_7 and B_7 supporting the strips, the actual delay-line system *A-B*, the plates A_8 and B_8 and the strips A_9 and B_9; the pulse leaves the system at *Q*. At any position along the delay line a current in system *A* is accompanied by a countercurrent in system *B*. *P* and *Q* are connected to the pulse generator and to the terminating load. By a "field converter" these connections can be interchanged, leading to pulse propagation in the opposite direction and to field inversion.

Also of importance in the design is the considerable reduction in cross-section of the beam during acceleration. It is estimated [7] that in the CERN proton synchrotron the width and height of the beam shrink by a factor of 10.

Consequently there are two types of kicker magnets. *Large-* or *full-aperture* kickers enclose the wide unaccelerated beam, and possibly the vacuum chamber as well. *Small-aperture* kickers only enclose the shrunken accelerated beam, and these are always inside the vacuum chamber.

If a kicker is to be used for injection into a synchrotron it must be of the full-aperture type. The vacuum chamber, if embraced by the magnet, must be made of ceramic or epoxy resin in the magnet gap, since pulsed magnetic fields cannot penetrate conducting surfaces.

When kickers are used outside the vacuum chamber special care must be paid to the insulation between the capacitor elements; for example they may be immersed in oil. For kickers used inside the chamber the vacuum provides an ideal insulating medium.

Small-aperture kickers have the advantage that for a given kick strength, length and rise time the required currents and voltages are relatively small (see eqs. (1) and (3)). If they are to be used for ejection from a synchrotron there is a complication however. They must not be in the way of the beam at injection and must be mechanically *shifted into position* only after the beam has shrunk sufficiently. In practice displacements of 5-10 cm must be performed with actuation times of the order of 0.1 s.

An example of a kicker magnet that is in current use in the CERN proton synchrotron is shown in *fig. 7*. It is a small-aperture, displaceable magnet used for ejection (area *c* in fig. 1). Details of its construction and operation are given in the caption of fig. 7 and in *Table I*. Its aperture is 20×22 mm and this may be contrasted with the 50×100 mm dimensions of a typical full-aperture kicker, that of the Alternating Gradient Synchrotron in Brookhaven [8].

Fig. 8 and *9* show one of the kicker magnets that will be used for inflection in the intersecting storage rings at CERN (areas *d* and *e* in fig. 1). A cross-section is shown in *fig. 10* and numerical details are given in Table I. At each inflection area two of these magnets will be used, giving a total maximum deflection of 2.5 mrad to 28 GeV protons. The magnets to be used inside the vacuum chamber can have a small aperture (20×44.5 mm) because it is the thin accelerated beam that is to be deflected. These magnets will be displaceable; they are to be removed after completion of the stacking process during which the protons of many synchrotron cycles are stacked

Table I. Data for the CERN synchrotron ejection magnet (fig. 7) and the inflection magnets for the storage rings (fig. 8-11). The ejection magnet consists of two units (*I* and *II* in fig. 7); the data in this table apply to a single unit. Similarly, two magnets will be used in each storage-ring inflection area; again the data apply to a single magnet.

	Synchrotron ejection magnet (per unit)	Inflection magnets for storage rings (per magnet)
Maximum flux density B	0.186 T	0.085 T
Effective magnetic length l	0.40 m	1.37 m
Maximum kick strength Bl	0.075 Tm	0.117 Tm
Maximum deflection angle α (for 28 GeV protons)	0.80 mrad	1.25 mrad
Flux pulse duration	0.1-2.1 μs	2.4 μs
Flux pulse rise time τ	85 ns	0.31 μs
Useful aperture: height h	22 mm	19.2 mm
width w	20 mm	44.5 mm
Inductance L_s	0.9 μH	4.3 μH
Capacitance C_s	7.7 nF	22.0 nF
Characteristic impedance Z	10.8 Ω	14 Ω
Maximum voltage on magnet V	35 kV	19.3 kV
Maximum current through magnet I	3.5 kA	1.37 kA
Moving mass	200 kg	400 kg (magn.) + 2100 kg (tank)
Stroke of movement	250 mm	−60 to +60 mm
Minimum actuation time	0.1 s	ca. 30 s

[6] See for example J. Millman and H. Taub, Pulse and digital circuits, McGraw-Hill, New York 1956, pp. 293-294.
[7] See reference [1], p. 364.
[8] E. B. Forsyth and C. Lasky, IEEE Trans. NS-12, No. 3, 882, 1965.

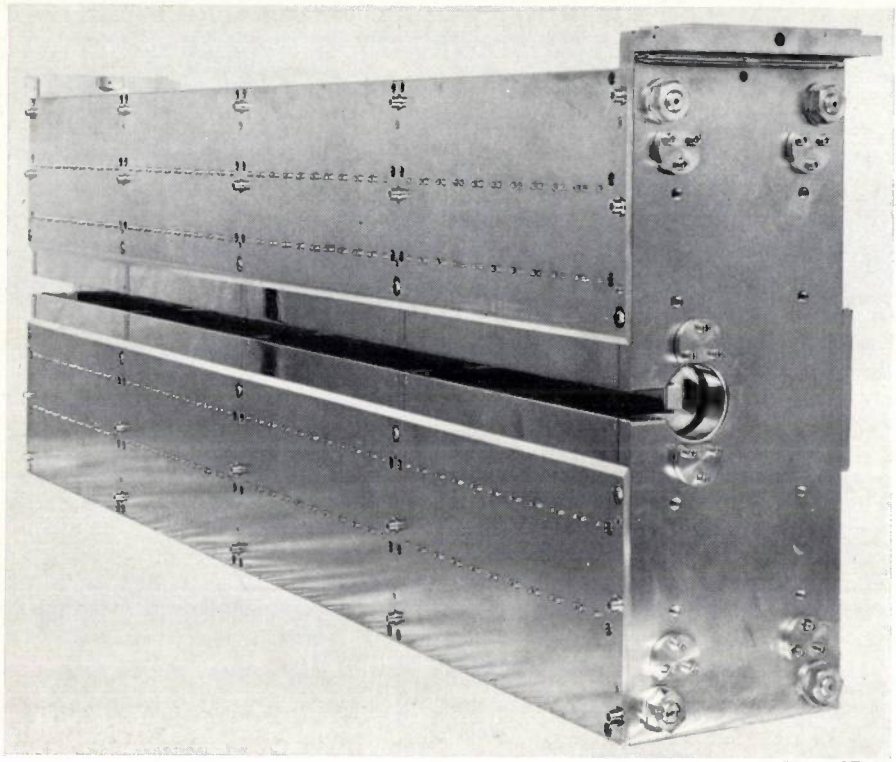


Photo CERN

Fig. 8. A kicker magnet for inflection into the storage rings. Two such magnets will be used in each of the areas *d* and *e* of fig. 1, giving a total deflection of 2.5 mrad to 28 GeV protons. Further details are given in Table I.

into a single beam. This is necessary as the stacked beam must be relocated in the horizontal plane for optimum use of the vacuum chamber since the beam "blows up" during its stored lifetime due to scattering by residual gas. Magnet and vacuum tank form one displaceable unit; the tank will be flexibly connected to the main vacuum chamber structure.

The decay time of the flux pulse (equal to the rise time, see Table I) does not need to be extremely short by synchrotron standards, as the beam of one synchrotron cycle is only occupying two-thirds of a storage ring, so that the magnets can be switched off in one-third of the proton orbital period in the storage rings. The greatest problem with these magnets is to

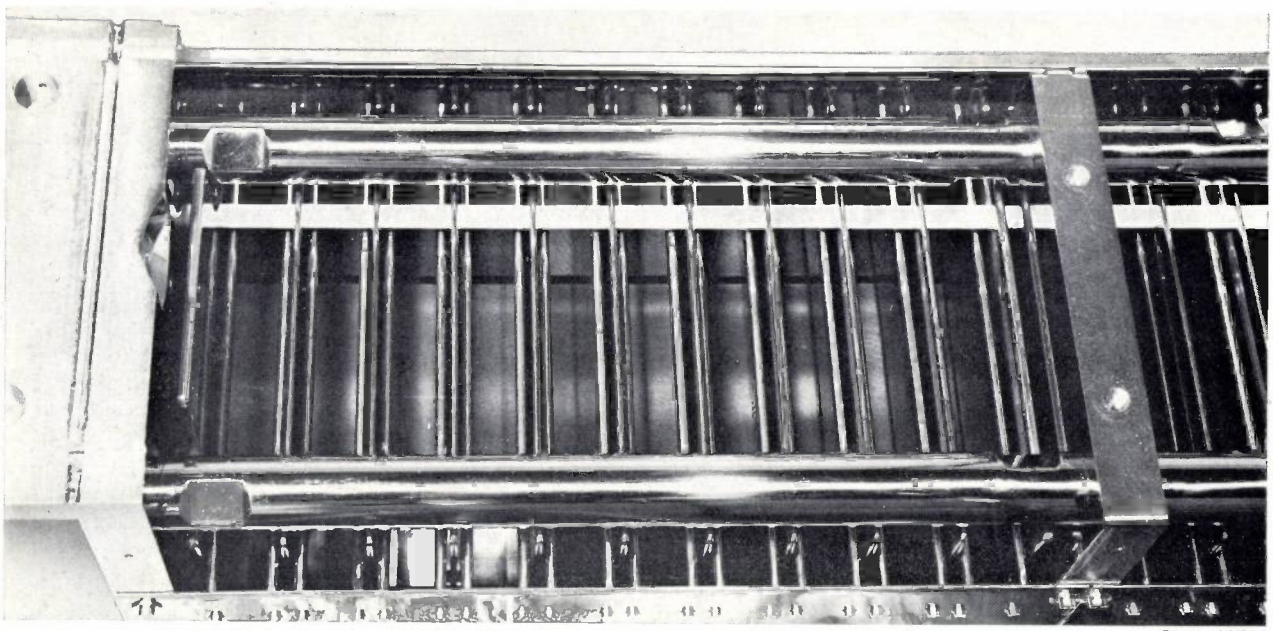


Photo CERN

Fig. 9. View from top into the inflection magnet of fig. 8, showing the capacitor plates. Between the plates the top of the ferrite blocks can be seen.

keep the mismatch reflections which occur in the first microsecond after the end of the pulse to less than 1 %, since the beam on the injection orbit continues to circulate through the kicker-magnet aperture until it is stacked.

These magnets are to be used in ultra-high vacuum (below 10^{-9} torr) and will be baked in their tank up to 300 °C, for complete outgassing. The capacitor plates are made of titanium, a material known to have good outgassing properties. The outgassing properties of the ferrite core (to which we shall return below) are also satisfactory. Because of the ultra-high vacuum and the high-voltage requirements, an elaborate cleaning and polishing procedure is demanded during the construction of such a magnet. In a test assembly this magnet has operated with pulsed voltages up to 25 kV, and a vacuum with a residual pressure of 2×10^{-10} torr has been achieved. A magnet of this type mounted in its vacuum tank is shown in *fig. 11*.

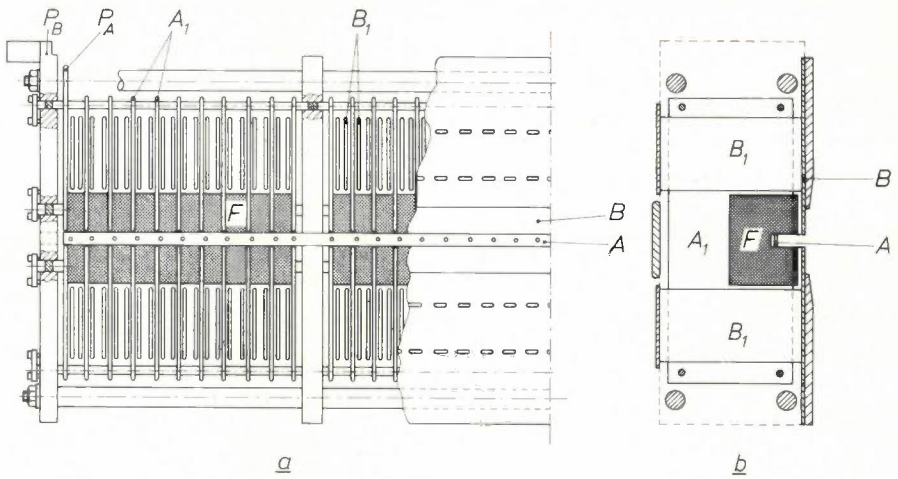


Fig. 10. *a*) Part of longitudinal section of the inflection magnet of fig. 8. *b*) Cross-section. *A* central conductor. *B* return conductor. A_1 capacitor plates connected to *A*. B_1 capacitor plates connected to *B*. *F* ferrite blocks. P_a is connected to the inner conductor and P_b to the outer conductor of the coaxial input to the system. The coaxial output is at the other end of the magnet. The magnet consists of four sections (each containing 10 ferrite blocks); the first section and part of the second are shown in (*a*).

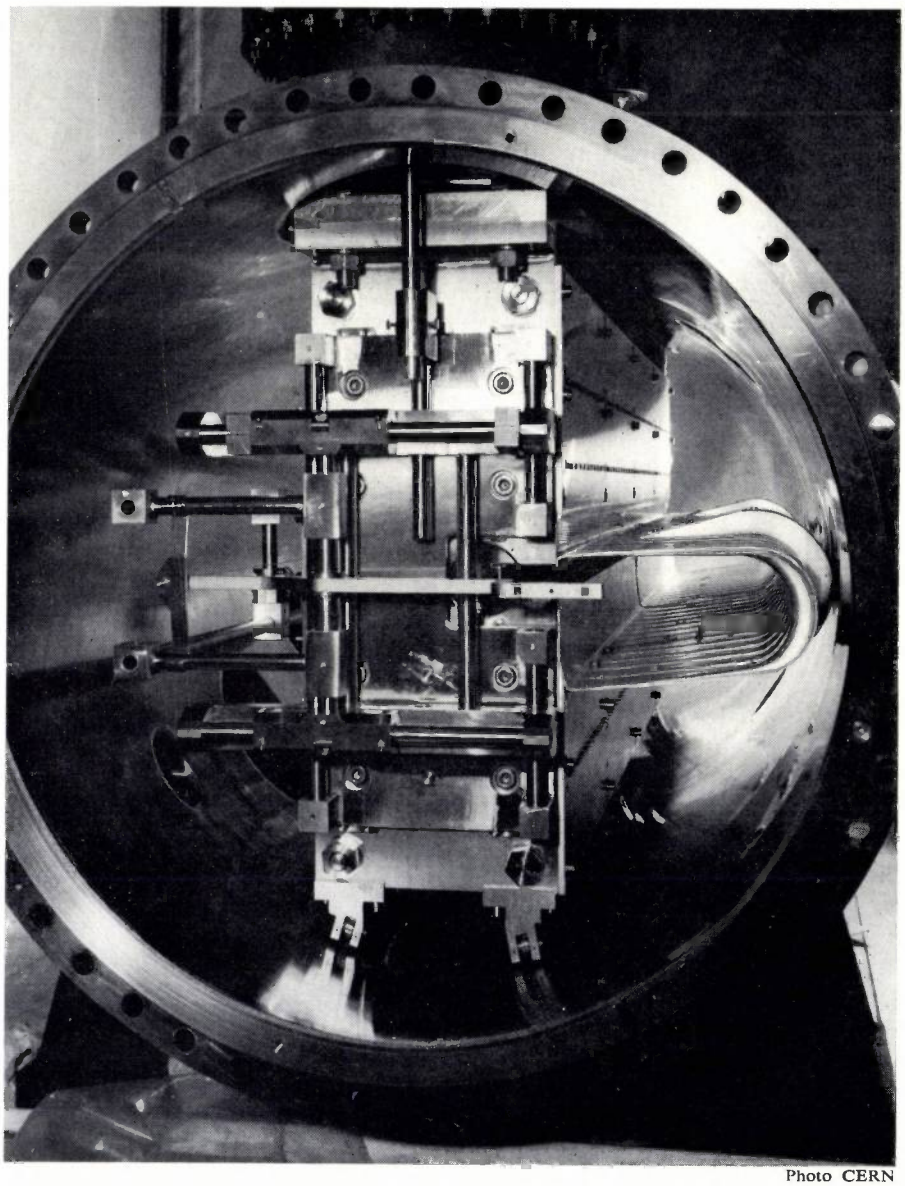


Photo CERN

Fig. 11. A storage-ring inflection magnet (see fig. 8), mounted in its ultra-high vacuum tank. A scanning machine for measurements of the magnetic field (see fig. 12) is temporarily mounted on the magnet. The system of horizontal and vertical bars guides the scanning loop. The movable screen on the right is for screening the stacked beams in the storage rings from the magnet. Magnet and tank together will be mounted in the storage rings; they may be displaced through ± 6 cm (12 cm total) as a single unit.

Core material for kicker magnets

As noted before, ferrites are the obvious magnetic materials for high-frequency applications at the present time. This is because they are effectively insulators and their high resistivity precludes eddy current losses, which are the main limitation with the traditional magnetic materials at high frequencies. For eddy-current losses to be negligible, the skin-depth must be large compared to the size of the material. At 1 MHz, the skin depth for iron turns out to be in the micron range, and laminating the material — the traditional remedy — is no longer feasible. In ferrites the skin depth lies in the centimetre to metre range at 1 MHz. This advantage of ferrites completely offsets the apparent disadvantages of a typical permeability of 100-1000 which is smaller than the value of 3000 for iron, and a smaller saturation polarization up to 0.4 tesla, compared with up to 2 teslas for iron. In practice the saturation polarization is not often a limiting factor, as the required flux density is reduced as much as possible by making the magnet as long as space permits (see eq. 2) so as to reduce the current requirements.

In this particular application which involves high voltages, the high resistivity of the ferrite has the additional advantage that high-voltage components can be inserted in the core without further insulation.

The two most commonly used ferrite varieties are ferroxcube 3 (manganese-zinc ferrites) and ferroxcube 4 (nickel-zinc ferrites). The latter is to be preferred, since again the advantage of a higher resistivity (10^4 Ωm in ferroxcube 4 compared to 1 Ωm in ferroxcube 3) offsets the disadvantage of a smaller permeability and saturation polarization. The losses in ferroxcube 4 (including those not caused by eddy currents) remain reasonably low, and the permeability reasonably high, over the full frequency range required for the magnetic-field pulse, that is, up to 20 MHz [9].

For the storage-ring inflector magnets at CERN (*figs. 8-11*), the sub-variety 4A of the ferroxcube 4 series has been chosen, because it has the lowest coercivity (see Table III of reference [2]). This is of importance for the suppression of any remanent flux that would perturb the beam before and after the magnet is excited.

A final point to note is that ferrite, despite its porosity, is a clean material that can be employed in ultra-high vacuum provided care is exercised during machining and subsequent cleaning of the material. During its manufacture it is sintered at temperatures above 1000 °C, and is devoid of organic contaminants. Preliminary comparative measurements on ferrite and stainless steel under the same conditions [10] have shown that, after careful cleaning, the outgassing rate of ferrite is only about 20 times as great as that of stainless steel (which is around 10^{-12} torr l/s cm^2). This is sufficiently low to permit its use in ultra-high vacuum.

The pulse-forming network

There are some practical problems in making the pulse-forming network described above. In the first place, attenuation due to losses in the delay line (*D* in fig. 5) will cause the flat top of the electrical pulse to droop, since the trailing end of the pulse travels a longer distance along the line than the leading end. Such a droop gives a decrease in the kick strength during the magnetic pulse (see *fig. 12*), and this must be restricted to within the tolerances set for the kick strength, which is ± 1 % for the storage ring inflector.

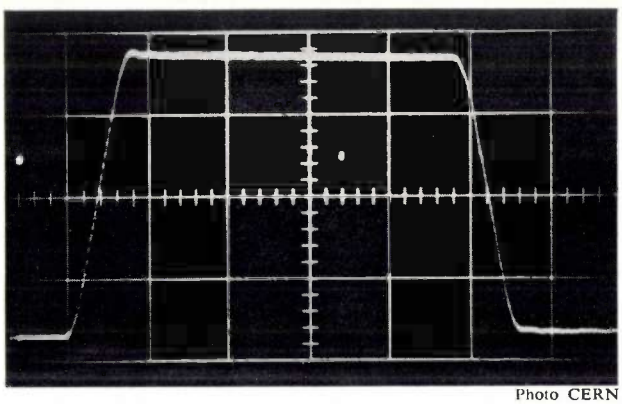


Photo CERN

Fig. 12. Magnetic-flux pulse, as measured with the scanning apparatus shown in fig. 11. The result is obtained with a single-turn loop 3.7 mm wide and 1.6 m long. The flux Φ is obtained from the induced voltage $d\Phi/dt$ by means of an integrator with $RC = 500$ μs at the input to the oscilloscope. The horizontal time scale is 0.5 μs/division. About 0.5 % of the 1.5 % droop of the flat top is due to the finite RC time of the integrator, and therefore the remaining 1 % is due to the attenuation in the delay line of the pulse-forming network.

Losses, which are more significant in low-impedance cable, may indeed form a problem for a delay line to be matched to a low-impedance magnet. A solution to this problem is to use several cables of higher impedance in parallel, but there is an economic and practical limit to this technique. In the network used to energize the storage-ring inflection magnet, the delay line *D* consists of two parallel 28 Ω coaxial cables with polyethylene insulation, to match the 14 Ω magnet.

Secondly, the switch in the pulse-forming network (*S* in fig. 5) is a critical element. It determines the rise time t_b of the electrical pulse, and, if not sufficiently fast, may limit the rate of rise of the magnetic flux pulse (see eq. 4). Moreover, it must be able to conduct the very high currents (thousands of amperes)

common in this work. Deuterium tetrode thyratrons have been successfully employed in cases where rise times of 30-50 ns were sufficiently short. The maximum rate of rise of current that these switches can handle is about 100 kA/μs. A faster switch is the spark gap; this can have a rise time as short as 5-20 ns. It should be noted that the effective value of τ may be increased by jitter or other disturbing effects.

To maintain the rectangular pulse form as closely as possible, all connections between the storage delay line, the switch, the magnet and the terminating resistor must be coaxial and matched. Finally, it should be noted that the peak power in the high-frequency components of the pulse is often many megawatts (greater than 30 MW in the storage ring inflection-magnet circuits); therefore, all components must be adequately screened to prevent interference with control and measuring equipment.

As noted in the introduction, the kicker magnet is rapidly becoming an essential piece of apparatus in high-energy nuclear physics. Particle beams can very efficiently be injected or extracted with kicker magnets; beam extraction has been performed with an efficiency greater than 90% [11]. Extraction is not only of importance for the transfer of beams from one ring to the next (fig. 1), but for any experiment to be performed with the accelerated beam [12]. The alternative method of setting up an experiment, that of inserting internal targets, is less flexible and gives rise to more induced radioactivity.

The author is indebted to Dr. B. de Raad and Dr. W. C. Middelkoop of the Beam Transfer Group, Intersecting Storage Rings Division, CERN for their helpful assistance in the preparation of this article, and to Dr. B. Kuiper of the Proton Synchrotron Machine Division, CERN for permission to publish details of the synchrotron ejection magnet.

Summary. Kicker magnets are designed to provide short rectangular pulses of magnetic flux. They are used for deflecting electrons or protons into and out of synchrotrons and storage rings. The CERN accelerator complex, comprising the main synchrotron (operating since 1959), an injector synchrotron (under construction) and the intersecting storage rings (nearing completion), contains six deflection areas where kicker magnets are or will be used.

Electrically the kicker magnet is equivalent to a delay line. A charged coaxial cable is discharged through the magnet to excite it. Currents in the kiloamp range and voltages in the kilovolt range are required in order to build up a useful kick strength within tenths of a microsecond in a gap of practical size.

Since a particle beam in a synchrotron is wide at injection but thin after acceleration, there are two types of kicker magnets: "full-aperture" kickers for injection, and "small-aperture" kickers for other applications. A small-aperture kicker operates in vacuum and must usually be rapidly displaceable.

As examples short descriptions are given of an ejection kicker in current use in the CERN proton synchrotron and of an inflection magnet for the storage rings which is still under construction.

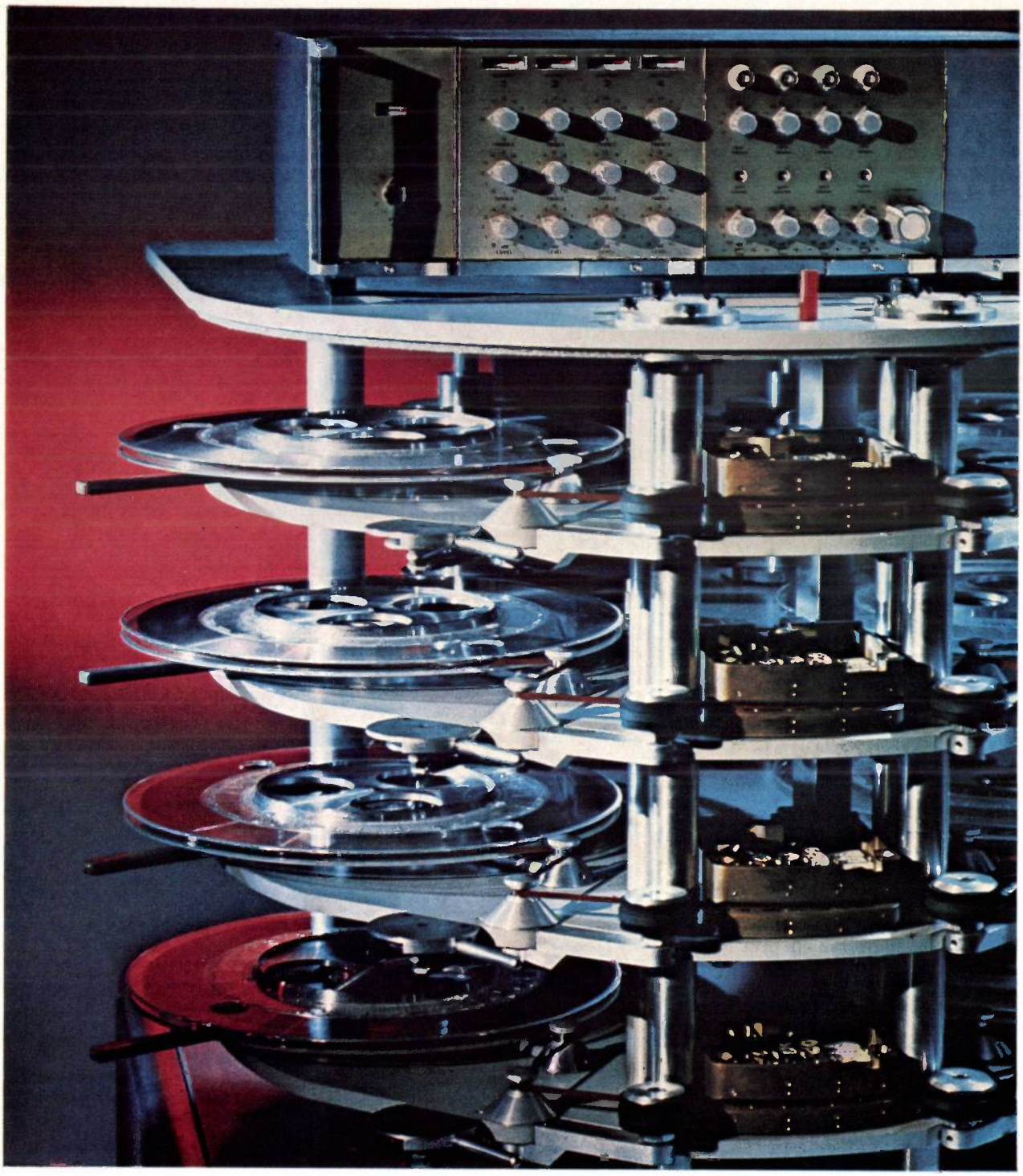
Because the pulse rise time must be short, the core has to be made of ferrite. Measurements on outgassing rates have shown the ferrite to be compatible with ultra-high-vacuum practice.

[9] For a more extensive discussion of the relevant properties of ferroxcube, see reference [2].

[10] J. Sherwood, CERN Report No. AR-SG/64-13. Industrial Report 1964.

[11] R. Bertolotto *et al.*, Proc. 1963 Dubna Int. Accelerator Conf., p. 669; B. Kuiper and G. Plass, Proc. 1965 Frascati Int. Accelerator Conf., p. 579.

[12] See for example K. H. Reich, Beam extraction techniques for synchrotrons, Progress in nuclear techniques and instrumentation 2, 161-215, 1967.



This machine records the programmes of music on the tapes for Philips musicassettes. Four tapes are prerecorded simultaneously, each with four sound tracks; this is done at a speed which is 32 times the nominal tape speed. Cassettes and cassette players are the subject of the article on the adjoining page.

Audio tape cassettes

P. van der Lely and G. Missriegler

"Speech (or singing) recorded on the cylinder can be reproduced as often as desired, with no weakening of the recording, and the timbre of the voice comes out well . . . The reproduced speech is of great purity and clarity, without annoying background noise. The later instruments reproduce with extraordinary fidelity not only what is spoken and sung, but also what is whispered into the microphone; even the faint sound of breathing can be reproduced."

Since V. Poulsen described his recording of sound on steel wire in these enthusiastic terms in the Annalen der Physik for 1900 (the "cylinder" served for winding on the wire), a great deal has changed in the technique of magnetic recording. In the thirties steel tape was still being used as the recording medium: this had to be played at a speed of two metres per second. However, the development of iron-oxide coated tape and high-frequency biasing led to a rapid increase in the use of magnetic sound recording after the Second World War, and to its wider popularity among the general public. Growing requirements for convenient operation and for effective protection of ever-thinner tapes resulted in the introduction of tape cartridges and cassettes. Among these the Compact Cassette developed by Philips, and now internationally standardized, is outstanding for its convenient shape and small dimensions, and has won a considerable and steadily growing popularity.

Why cassettes?

In the last ten years the use of tape recorders has increased to an extent that at one time would have seemed almost impossible. The popularity of the do-it-yourself sound recording is perhaps only to be compared with the popularity achieved through the years by home movies with 8 mm film.

The technical improvements that have accompanied this development, and are reflected in the products, have taken various directions. They have led not only to a higher quality of reproduction but also, and no less significantly, to simpler operation of the recorders and to a reduction of their volume and weight. This development has also brought the tape recorder into a special field, that of the dictation machine.

Like loading a camera with 8 mm cine film, threading the tape into a tape recorder or dictating machine is a relatively awkward operation for the inexperienced user. And now that recording equipment is so much more portable, tape often has to be loaded in difficult conditions, e.g. in a moving vehicle. The cartridge or cassette provides the answer to this problem. It relieves the user of the need to manipulate the tape himself and it offers effective protection.

The protection of the tape is the second important function of the cassette. With the development of thinner tapes, and with the lower tape speeds and narrower sound tracks made possible by improvement of the magnetic properties, it has become imperative to protect the tape from dust and fingerprints. In fact, the use of a cassette is almost essential if the full potentialities of present-day tapes are to be realized [1].

What type of cassette?

In recent years many and various types of audio tape cassettes (often referred to as "cartridges") have appeared on the market. They may be divided into two groups: one-reel and two-reel cassettes.

The one-reel cassette is shown in *fig. 1a*; at the end of the tape there is a catch-piece which is fed into a second, empty cassette after the cassette has been

P. van der Lely is with the Hasselt (Belgium) branch of the Philips Radio, Television and Record-playing Equipment Division; Dipl.-Ing. G. Missriegler is with Oesterreichische Philips Industrie, Werk Wirag GmbH, Vienna.

[1] In view of the two advantages noted here it is not surprising that tape cassettes are also being used for video recording. A video cassette and associated colour video recorder for use in the home are at present under development.

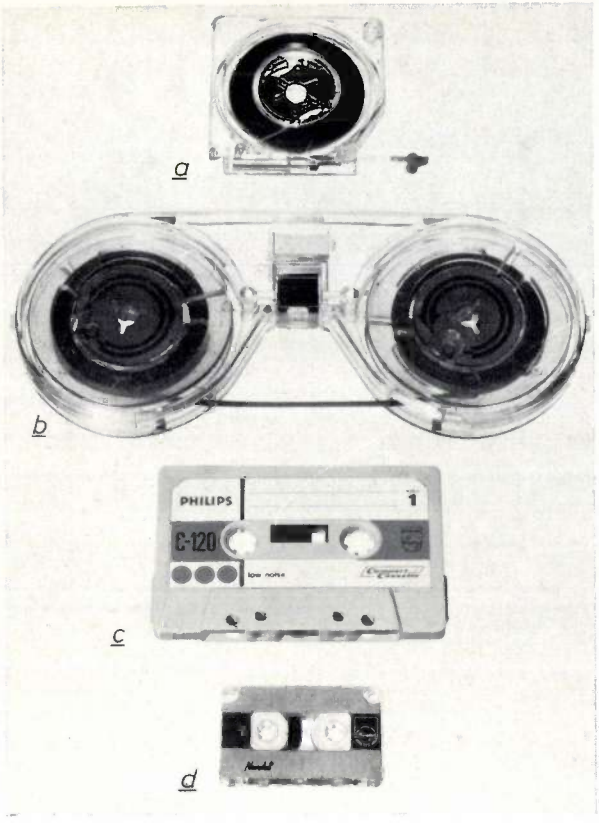


Fig. 1. Various types of audio tape cassette. *a*) One-reel cassette. *b*) Reel-to-reel cassette. *c*) Compact Cassette. *d*) Cassette for the Pocket Memo dictating machine.

loaded in the machine: the empty reel engages with the catch-piece and begins to wind on the tape. The one-reel cassettes are used in dictating machines. In another form of one-reel cassette the tape has an endless loop; since the fast forward winding and rewinding needed for finding a recording quickly is not possible with this kind of cassette, it is only suitable for some applications (announcements, back-ground music).

Unlike the one-reel cassettes the two-reel cassettes contain both a supply reel and a take-up reel. In their original form they were no more than an encapsulation of the two reels of a tape recorder together with the length of tape between them. At first, therefore, they were rather bulky (fig. 1*b*). The trend of development, however, was towards smaller dimensions. At Philips this development led to the Compact Cassette (fig. 1*c*), which is now very widely used not only for making recordings but also as the "musicassette" with pre-recorded tape [2]. The compactness of the Philips cassette is partly due to the use of flangeless reels with tape which is narrower than usual — 3.81 mm (0.15 inch) instead of 6.25 mm (0.25 inch). But an even more important factor is that the quality of the magnetic tapes has advanced to such a stage that the design of the cassette could be based on a tape speed of only 4.76 cm/s ($1\frac{7}{8}$ inch/s). For general use the Compact Cassette seems to have won the day from the one-reel cassettes, not just because it is compact but also because it can simply be taken from the machine without first having to wind it back, and because the equipment that the cassette fits into does not have to be as complicated.

Even more advanced miniaturization than with the Compact Cassette was possible in the design of a cassette for a pocket dictation machine (fig. 1*d*). Here, since the quality did not have to be so high, there was no need to ensure a constant tape speed by using a drive capstan and pressure roller: the tape is transported directly by the take-up reel, and runs faster as the diameter of the winding on the reel increases. The result was not only a very small cassette but also an extremely compact transport mechanism.

This machine — the Philips Pocket Memo — is shown in *fig. 2* beside the smallest of the Philips range of cassette players. Later on we shall look more closely at some of the special features of both these machines, but first we shall look more generally at the magnetic recording process and the relationship between tape speed and quality of reproduction.

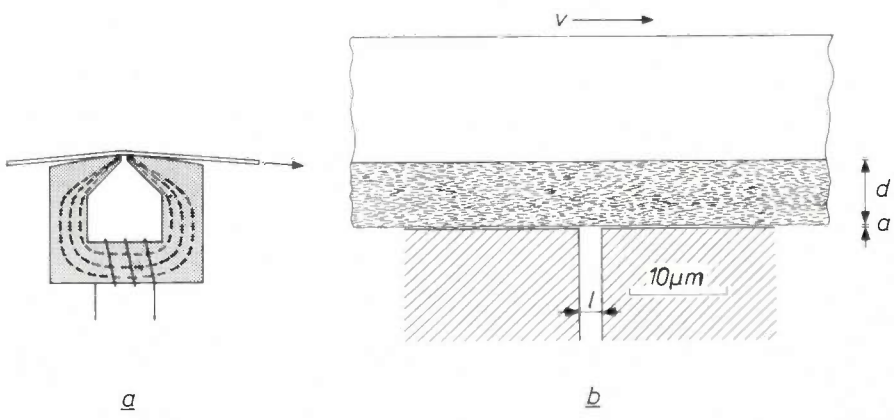


Fig. 3. *a*) The record/playback head has a soft-iron core or yoke with a very short air gap The tape is in contact with the head at the location of the gap. The magnetization of the tape sends its lines of force partly through the soft-iron core of the head and thus through the winding round the head.
b) Scale drawing, magnified about 1300 times, of a tape 18 μm thick moving at a speed v past the gap, 2 μm long, of the record/playback head. The tape consists of a plastic base coated with a magnetic layer. d thickness of the magnetic layer. a average distance between tape and head. l length of the gap. The dimensions correspond to those in actual cassette players.

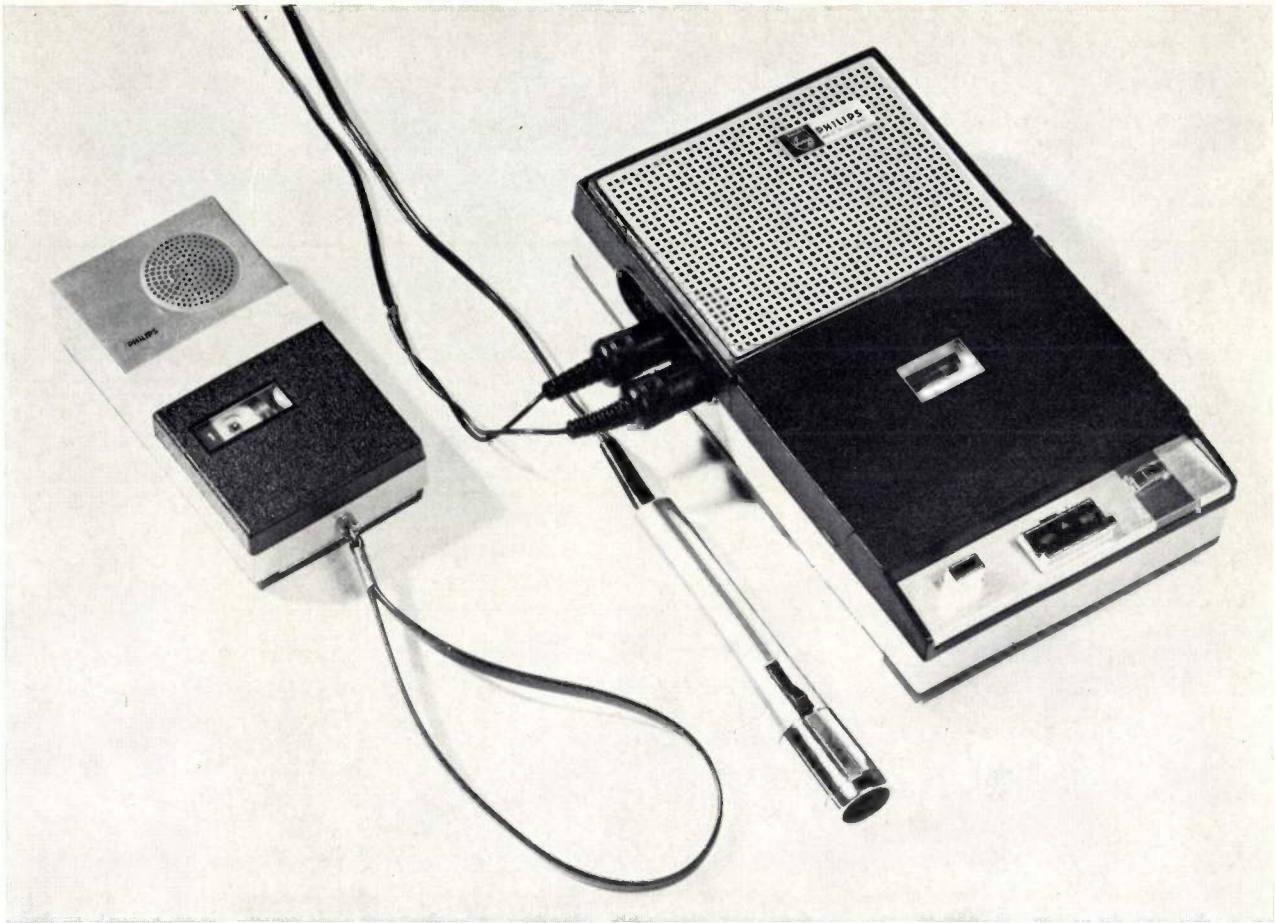


Fig. 2. Left: Pocket Memo dictating machine; right: cassette player with microphone.

Recording and playback at low tape speed

The design of the Compact Cassette is based on a tape speed of 4.76 cm/s. This is low compared with the tape speeds of 9.53 cm/s and 19.05 cm/s commonly used for sound recording, and this means that when a higher frequency f is being recorded the wavelength λ of the periodic magnetization on the tape is small, since $\lambda = v/f$, where v is the tape speed. The recording and reproduction of such small wavelengths sets certain requirements on the recording and playback process. The success achieved in meeting these requirements has been such that in normal use frequencies up to 10 kHz can readily be recorded and reproduced at a tape speed of 4.76 cm/s; a performance that ensures satisfactory reproduction of music.

Within the limited scope of this article the treatment of the recording and playback process which now follows is necessarily highly simplified. There are some avantages in describing the playback process first.

Playback

The requirements to be met by the playback process relate mainly to the geometry of tape and head. *Fig. 3a* illustrates the way in which a tape moves past a recording or playback head whose soft-iron core has a gap at the location where the tape is in contact with it. Fig. 3*b* shows a scale drawing, at about 1300 times full size, of a head with a gap length of 2 μm [3] and a tape 18 μm thick, "triple-play" tape. These dimensions apply to the cassette player.

Let us assume that the magnetic layer in fig. 3*b* is sinusoidally magnetized. The magnetization in and around the gap sends its lines of force partly through the soft-iron core of the head, and hence through the winding around it. The movement of the tape causes the magnetic flux Φ enclosed by the coil to vary, thus

[2] The Compact Cassette has been internationally standardized: IEC Publication 94, Addition 1. In addition to its use for sound recording, the Compact Cassette is also beginning to be used for recording digital signals; for program input in smaller computers and for use in cash registers it can be an improvement on the punched tape. These applications require a somewhat modified construction, and sometimes cassettes are used that have been comprehensively inspected for "drop-outs" (momentary interruptions of the signal), which cause more of a nuisance in digital recording than in sound recording. Preparations for the standardization of digital cassettes have reached an advanced stage.

[3] It is customary to call the dimension of the gap in the direction of tape travel the "length", although it is much smaller than the dimension perpendicular to the plane of the drawing in fig. 3, which is called the "width".

inducing in the coil a voltage E which is equal to $-N\mathrm{d}\Phi/\mathrm{d}t$ (N being the number of turns). If we represent the flux Φ, which varies sinusoidally with time t, by $\Phi = \Phi_0 e^{j\omega t}$, where $\omega = 2\pi f$, then we may write

$$E = -N\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -Nj\omega\Phi. \quad . \; . \; . \quad (1)$$

We see that the amplitude of E increases linearly with the frequency f. Thus, each time the frequency doubles the level of E increases by 6 dB. This 6 dB increase per octave is represented by the straight line A in *fig. 4*.

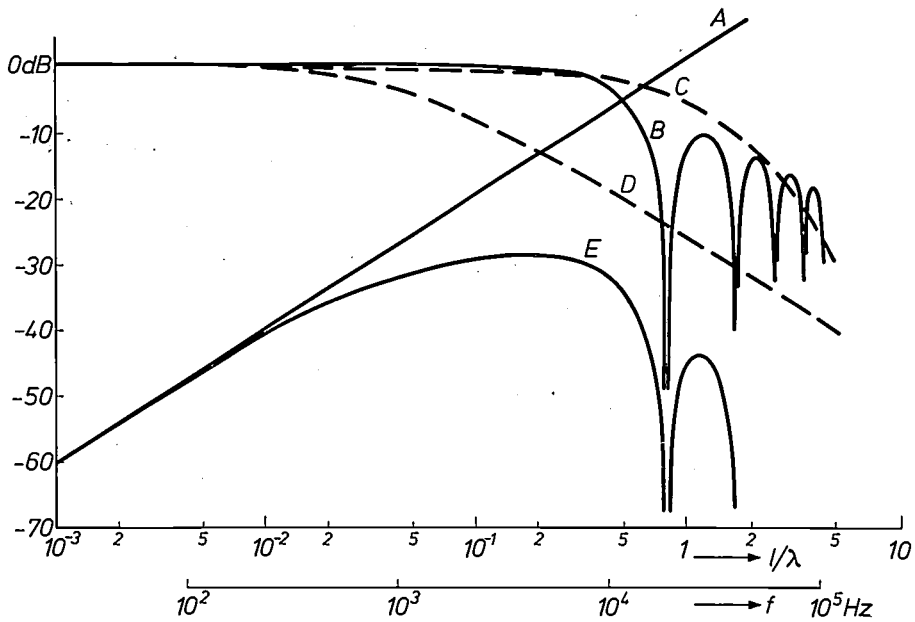


Fig. 4. The frequency characteristic of the playback process is determined by a number of effects. With a constant magnetic flux through the playback head the induced voltage at the terminals of the head increases linearly with the frequency f (curve A). The flux through the playback head drops to zero, however, when the frequency is so high that the wavelength λ of the magnetization on the tape or an integral multiple of it is approximately equal to the gap length l (curve B). Furthermore, with rising frequency the loss increases, partly because there is a certain spacing a between head and tape (curve C) and partly because at higher frequencies a decreasing part of the thickness d of the magnetic layer contributes to the output (curve D). The addition of all these losses results in the frequency characteristic E. With certain simplifying assumptions the curves were calculated for a tape speed of 4.76 cm/s, $l = 2$ μm, $a = 0.2$ μm, and $d = 6$ μm, i.e. values that are normally encountered with cassette recorders.

We shall use this straight line as the basis for calculating the frequency characteristic of the playback process. In practice it is only at low frequencies that any approximation to this characteristic is found [4]: at high frequencies there are losses because of certain geometrical factors.

To begin with, there are the gap losses. Clearly, the detailed pattern of the tape magnetization between the beginning and end of the gap cannot be observed by the playback head. The wavelength of the magnetization on the tape must therefore be large compared with the gap length l. The nearer the wavelength approaches to the gap length the smaller the recorded signal; this is illustrated by curve B of fig. 4. If the wavelength or an integral multiple of it is approximately equal to the gap length, then there is no transfer at all. In practice the minimum wavelength is taken to be $2l$, at which the gap loss is 5 dB. In the cassette player of today the gap length is 2 μm. The limiting wavelength is thus 4 μm, which corresponds to an upper cut-off frequency of 12 kHz.

In addition to the gap losses, attenuation occurs at short wavelengths because the tape is not in complete contact with the head but is displaced from it by an average distance a. To calculate this attenuation we start with the flux which is induced in the head by a thin layer of the tape of thickness $\mathrm{d}y$ and located at a distance y from the head; this flux is proportional to $e^{-2\pi y/\lambda}$ [5]. If the whole magnetic layer of the tape is uniformly magnetized — this is not in fact true but we shall permit ourselves the assumption to make things easier — then the total flux Φ through the head is proportional to the average value of this function over the thickness d of the tape, i.e. proportional to

$$\frac{1}{d}\int_a^{d+a} e^{-2\pi y/\lambda}\,\mathrm{d}y = \frac{\lambda}{2\pi d}\, e^{-2\pi a/\lambda}(1 - e^{-2\pi d/\lambda}). \quad . \quad (2)$$

The factor $e^{-2\pi a/\lambda}$ represents the attenuation due to the distance a; this attenuation is 54.5 dB per wavelength distance between tape and head. In practice the distance is 0.1 μm to 0.3 μm, which, at a wavelength of 4 μm, corresponds to an attenuation of 1.4 dB to 4.1 dB. This "spacing loss", calculated for a spacing a of 0.2 μm, is given by curve C of fig. 4.

The factor $(\lambda/2\pi d)(1 - e^{-2\pi d/\lambda})$ indicates that as the wavelength becomes shorter the contribution to the output signal comes more and more from the magnetization close to the tape surface. There is a "thickness loss", and its magnitude appears in fig. 4 from curve D for the tape-head configuration of fig. 3*b*, i.e. for a layer thickness of 6 μm.

If at each frequency we add up the three different sorts of losses mentioned above and then subtract the sum from the theoretical 6 dB/octave characteristic A, we obtain the curve E of fig. 4. This curve gives the frequency response of the playback process, and is valid for the assumption of homogeneous magnetization over the whole thickness of the tape. It is evident that the cut-off frequency is determined primarily by the gap losses (curve B). One of the related problems with a mass-produced product like the cassette player is that of making the very narrow gaps in the heads accurately to size and with properly finished edges; an increase of 1 micron in the effective gap length could bring the theoretical cut-off frequency mentioned above from 12 kHz to 8 kHz.

There is another possible cause of losses at high frequencies, and this is also significant because the cassette recorder is a mass-produced product. We have tacitly assumed in our analysis that the tape is magnetized by a recording head in which the width of the gap, like that of the playback head, is perpendicular to the direction of travel of the tape, i.e. perpendicular to the plane of fig. 3. The correct setting is made as accurately as possible by individual adjustment of what is usually called the "azimuth" of the head. A difference in angle α between the recording and reproducing gaps gives an effective lengthening of the gap and additional losses at short wavelengths. These losses, again expressed in dB, are given by

$$20 \log_{10} \frac{\sin(\pi w \tan \alpha/\lambda)}{\pi w \tan \alpha/\lambda} \text{ dB}, \quad \ldots \quad (3)$$

where w is the width of the sound track. A plot of (3) as a function of frequency gives a curve similar to curve B of fig. 4. In monaural recording on a cassette tape the track width is 1.5 mm; a 3 dB attenuation of the cut-off wavelength of 4 μm occurs here at an azimuth error α of only 4′. Regular adjustment of the azimuth of the head is obviously not possible in cassette machines, and therefore azimuth errors may cause significant deterioration of the frequency characteristic. They are mainly significant when a tape is played back on a different machine from the one on which the recording was made.

Recording

To obtain a linear relationship between magnetization and signal current during the recording process, a high-frequency a.c. current is passed through the recording head together with the signal current [6]. The amplitude of this bias current is so high as to cause the magnetic field at the gap to periodically exceed the coercivity of the magnetic material in the tape; its frequency is several times higher than the audio frequencies to be recorded. The tape acquires its remanent magnetization when the tape has passed the gap and traverses a zone in which the peak value of the magnetic field has decreased to a value exactly equal to the coercivity (*fig. 5*). This recording zone has a certain extent, partly because all the magnetic particles in the tape do not have exactly the same coercivity. It is desirable that the instantaneous value of the signal should not change much during the time that a point of the tape passes through the recording zone, in other words that the

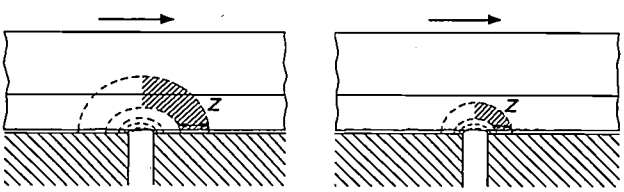


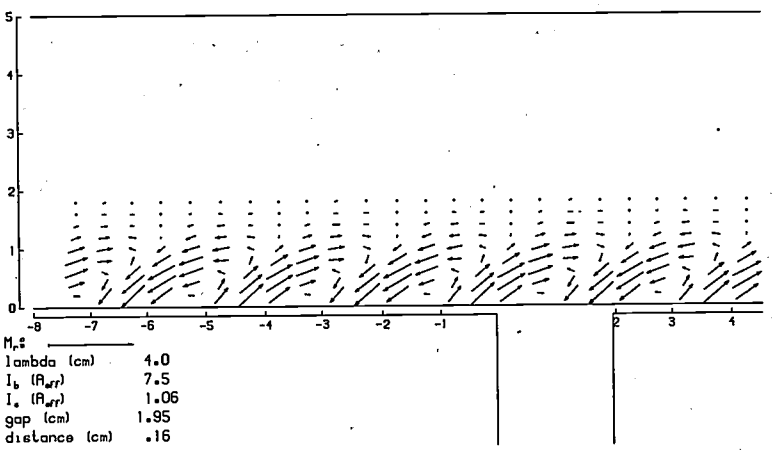
Fig. 5. The magnetization of the tape takes place in a zone Z, in which the magnetic field strength has decreased to the coercivity of the tape. The path through this zone is longer at a high bias current (left) than at a small bias current (right).

extent of the recording zone should be small with respect to the wavelength to be recorded. The width of the recording zone increases with the magnitude of the bias current (fig. 5). For short wavelengths it is therefore best to use a smaller bias current. The recording zone then contracts around the gap and no longer covers the full thickness of the magnetic layer. This is no disadvantage, however, since as we saw above at short wavelengths the contribution to the output signal

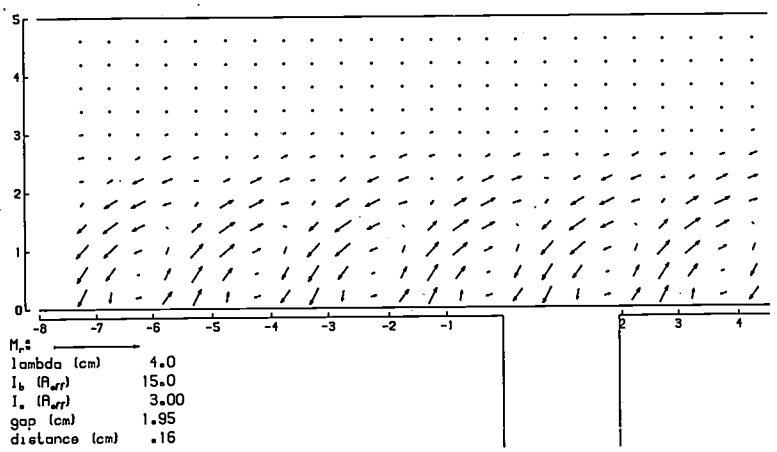
[4] Unless the wavelength on the tape is much larger than the length of tape in contact with the head; this is not the case for the cassette player, where the wavelength at 50 Hz is only about 1 mm.

[5] R. L. Wallace, Jr., The reproduction of magnetically recorded signals, Bell Syst. tech. J. **30**, 1145-1173, 1951.

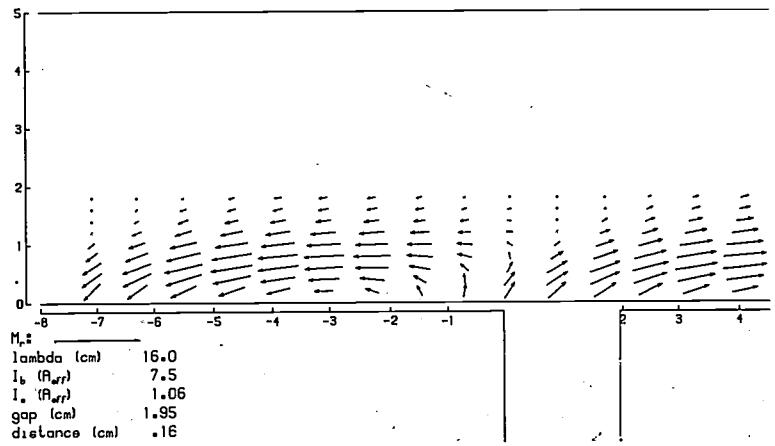
[6] The idea of the high-frequency bias current seems to have had three different originators; see W. K. Westmijze, Studies on magnetic recording, Thesis, Leyden 1953. Its linearizing action is dealt with by the same author in: The principle of the magnetic recording and reproduction of sound, Philips tech. Rev. **15**, 84-96, 1953/54.



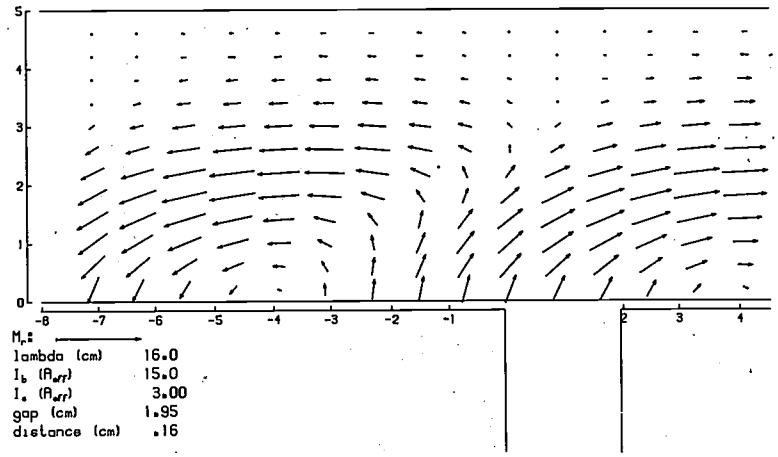
a



b



c



d

Fig. 6. The magnetization inside a magnetic tape, measured on a 5000 : 1 scale model. The magnitude of the magnetization is expressed by the length of the arrows; for comparison the remanence M_r of the magnetic material is also given. The gap in the head is shown beneath the longitudinal section of the magnetic layer in the position which it took up at one of the instants when the signal current passed through zero. *a*) Short wavelength and low bias current I_b. *b*) Short wavelength and high bias current; although the signal current I_s is three times greater than in (*a*), the remanent magnetization is lower. *c*) Long wavelength and low bias current; only part of the layer thickness is magnetized. *d*) Long wavelength and high bias current; almost the entire layer is magnetized.

comes only from the surface of the magnetic layer. On the other hand at low frequencies, i.e. at long wavelengths it is useful to magnetize the layer over its whole thickness, and for this purpose a higher bias current is desirable.

The effect of the bias current on the recorded remanent magnetization at long and short wavelengths is illustrated in *fig. 6*, obtained from measurements on a scale model of the magnetic recording process magnified 5000 times [7]. For a short wavelength a comparison of fig. 6*b* and fig. 6*a* shows that doubling the bias current I_b gives a smaller remanent magnetization even though the signal current I_s is made three times as large. In this case the magnitude of the magnetization finally impressed on the tape is determined not only by the signal current but also by the bias current. At long wavelengths the situation is different; in fig. 6*c,d* it can be seen that when the wavelength is increased by a factor of four the remanent magnetism is the same both for high and low bias current, though with the high bias current a greater depth of the layer is magnetized. At this wavelength the result is an increased output voltage on playback.

The relative levels at which high and low frequencies are recorded on the tape therefore depend partly on the magnitude of the bias current. This has to be taken into account when devising the corrections to the frequency response that will give the correct relative levels on playback; more will be said about this when we describe the cassette player on page 88. In any case the choice of the bias current has in practice to be a compromise between what is desirable for high frequencies and what is desirable for low frequencies [8]. The compromise adopted for the cassette player rather favours the high frequencies; the bias current is lower than in normal sound recording.

When the magnetic coating of the tape is saturated, the maximum output level of the tape has been reached. Exceeding this level will cause non-linear distortion. When making a recording the aim is to magnetize the tape to a level at which the loudest passages do not quite reach the maximum output level; this is often defined as the signal amplitude at which the third harmonic caused by non-linear distortion of the fundamental reaches 5% of the amplitude of the fundamental. This definition applies for low frequencies. At high frequencies it is not usually possible to drive the tape into saturation. This is because an increase of the signal current, like an increase of the bias current, makes the recording zone broader, so that the recorded signals are smaller for the short wavelengths, which are of the same order of magnitude as the width of the recording zone or less. At short wavelengths, therefore, the impressed magnetization does not increase linearly with the signal current but reaches a maximum. This deviation from linearity at high frequencies is the cause of non-linear distortion. At high frequencies, also, the permissible distortion determines the maximum signal level that can be recorded on the tape, but in this case it is not connected with magnetic saturation. At a higher bias current the deviation from linearity will start at lower signal currents, which means that the maximum undistorted output level of the tape is lower.

The tape

Audio tapes are made by coating a plastic base with a thin layer consisting of a suspension of magnetic material in a volatile solvent, mixed with an organic binder. After drying, the magnetic material — particles of iron oxide or nowadays chromium dioxide — remain on the base embedded in the binder. The following four factors have contributed significantly to the improvement in tape quality.

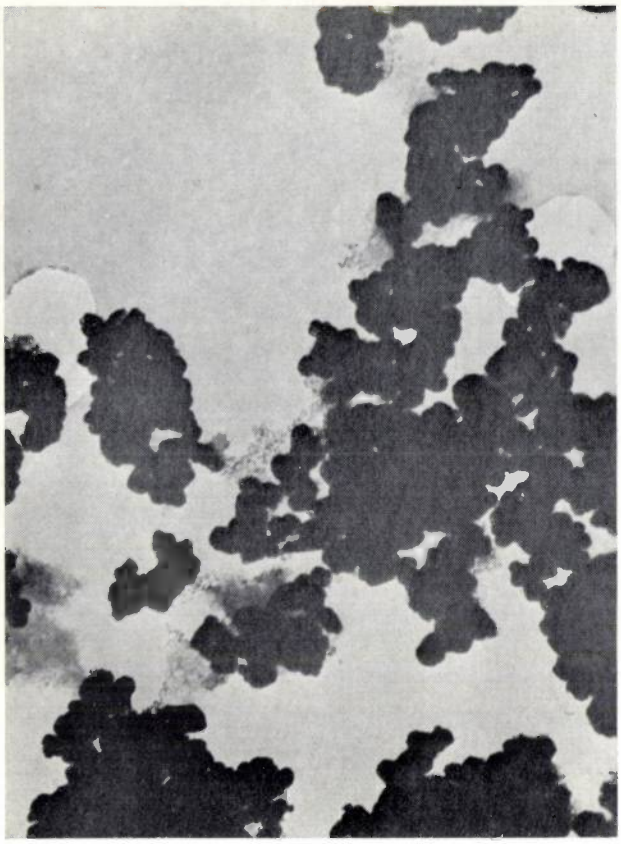
1. Improvement in the shape of the particles, giving the material better magnetic properties;
2. A more uniform distribution of the particles in the binder;
3. The use of binders with better resistance to wear;
4. Finishing treatment of the surface to make it smoother.

Originally cubic iron-oxide particles were used (*fig. 7a*). An advance came with the use of smaller, needle-shaped particles, about 1 μm long and 0.2 μm thick (fig. 7*b*). Smaller particles are important because the average size of a particle should preferably be less than half the shortest wavelength to be recorded. The shape anisotropy of the needles gives a greater coercivity; moreover, needle-shaped particles can be oriented parallel to the direction of travel of the tape while the coating is still wet. This treatment increases the remanent magnetization of the tape and thus gives a higher output voltage on playback.

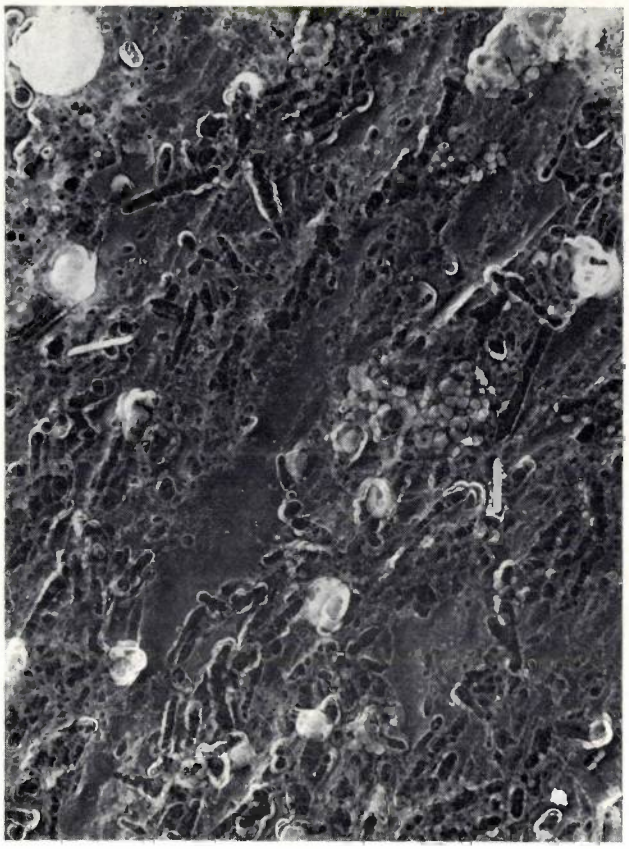
Since the particles are magnetic, they tend to cluster together in the binder while it is still wet. This effect ("clumping") makes it difficult to obtain a homogeneous distribution. During the preparation, which takes place at high temperature (about 350 °C), the iron-oxide particles are sometimes found to be sintered together. This leads to a high noise level and a rough surface, which increases the average spacing between tape and head and consequently increases the spacing

[7] D. L. A. Tjaden and J. Leyten, A 5000: 1 scale model of the magnetic recording process, Philips tech. Rev. 25, 319-329, 1963/64.

[8] There is no need to be content with a compromise if low and high frequencies are recorded one after the other, each under its own optimum conditions, on the same sound track: E. de Niet, K. Teer and D. L. A. Tjaden, Magnetic recording of audio signals at low tape speeds, 4th Int. Congress on Acoustics, Copenhagen 1962, paper No. N 11.



a



b

loss. Clumping can be countered by adding dispersion stabilizers to the suspension. These are organic substances which adhere to the iron-oxide particles. Nowadays more effective dispersion stabilizers are available, and in addition there are other effective methods of dispersion. Modern tapes are therefore more homogeneous and have a smoother surface.

A smoother surface gives a lower spacing loss. Another improvement in this respect comes from the use of binders which have better resistance to wear, so that there is less contamination of the head from abrasion of the magnetic layer. Moreover most tapes are now given a finishing treatment to make the surface smoother. This is a calendering process, in which the tape is passed between heated rollers under pressure.

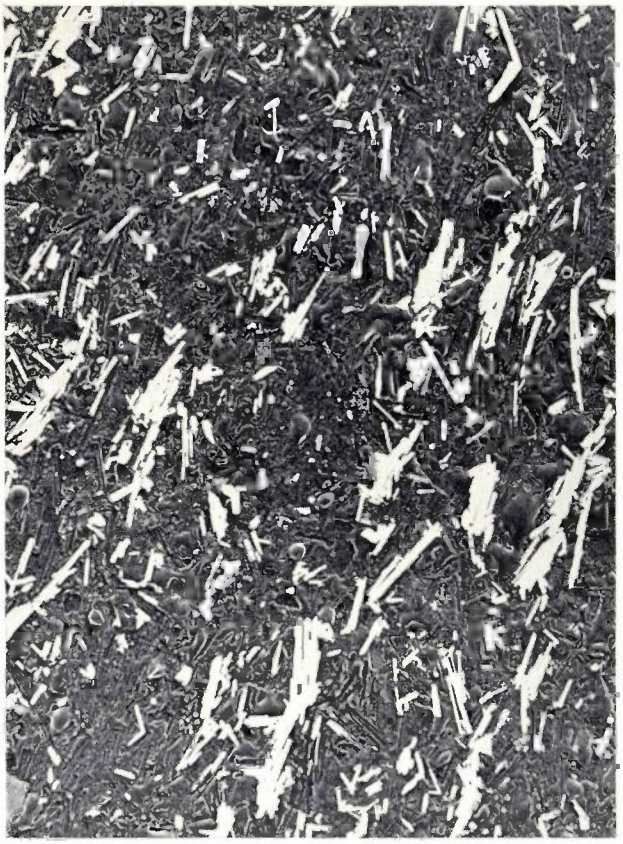
Tapes with chromium dioxide have particularly good properties. This is mainly because the chromium-dioxide needles are more uniform in size than the iron-oxide needles and also because they do not sinter together during preparation (fig. 7*c*). There is no sintering since, unlike the iron oxide, the chromium-dioxide particles are prepared in a wet chemical process at ordinary temperatures. As they do not form clusters to the same extent, the chromium-dioxide needles can also be better oriented, giving a higher remanence.

Apart from the improvements in the magnetic coating, there have also been important improvements in the base. The oldest tapes, intended for a tape speed of 76.2 cm/s, were subjected to considerable mechanical stress and the paper base had therefore to be very thick; the thickness of these "standard" tapes was 54 μm. The availability of stronger base materials (in the order of their appearance: cellulose acetate, polyvinyl chloride and polyester) and the reduction of tape speeds made it possible to introduce thinner tapes (see *Table I*). Tapes with thicknesses of 12.5 μm and 9 μm, which are the latest development in this field, are exclusively for use in cassettes; since the bias current in cassette recorders is lower, the magnetic coating of these tapes is thinner. The greater flexibility of thin tapes gives better contact between tape and head.

An important aspect of the improvement in tapes is

Table I. Stages in the development of thinner audio tapes.

Total thickness (μm)	Thickness of magnetic layer (μm)	Name	Type designation of cassette
54	15	standard tape	—
35	10-11	longplay tape	—
25	10-11	double-play tape	—
18	6	triple-play tape	C 60
12.5	3-4	—	C 90
9	3-4	—	C 120



c

Fig. 7. Electron photomicrographs of the magnetic layer of *a*) a tape with cubic iron-oxide particles, in common use about 20 years ago; *b*) a tape with needle-shaped iron-oxide particles as commonly used today, and *c*) a recently marketed tape with chromium-dioxide particles. The magnification of all three photomicrographs is 14 500×. Those in (*b*) and (*c*) were made by the replica technique; in (*c*) the white bars are chromium-dioxide needles that have stuck to the replica, while other chromium-dioxide needles have only left impressions in the replica.

that it has led to an increase in the density of the information that can be recorded on magnetic tape. Contributory factors here have been the introduction of thinner tapes, improved characteristics at short wavelengths, higher remanence and improved noise characteristics; the last two factors have made narrower sound tracks possible. A reel with a diameter of 18 cm, for example, can hold 360 metres of standard tape. In the old days, recording over the full width of the tape and using a tape speed of 38.1 cm/s, which was then necessary for high quality, the playing time was 15 minutes. The same sound quality can now be obtained with a four-track recording and a tape speed of 9.5 cm/s. The same 18 cm reel can now take 1080 metres of triple-play tape and thus at this speed gives a playing time of 4×180 minutes. This means a 48-fold increase of the information per unit volume.

With such a long playing time on one reel of tape it becomes difficult to find a particular passage or programme; the reel with four-track tape has become a relatively inaccessible carrier of information. Here again the cassette provides a welcome solution, since it can be taken out of the machine without first having to wind the tape back, thus offering increased information density in a smaller package.

[9] L. F. Ottens, The Compact Cassette for audio tape recorders, J. Audio Engng. Soc. **15**, 26-28, 1967.

The Compact Cassette

We shall now deal in somewhat more detail with the design of the Compact Cassette [9]. We have already mentioned that it was designed for a tape width of 3.81 mm and for a tape speed of 4.76 cm/s. Two sound tracks are recorded on the tape, one in the forward direction and one in the reverse direction. Depending on the thickness of the tape used, the cassette gives a playing time of 2×30 minutes, 2×45 minutes or 2×60 minutes; the three types are designated by the numbers C 60, C 90 or C 120 (see Table I). In stereophonic equipment the 1.5 mm wide sound tracks are each subdivided into two tracks of 0.6 mm, with a separation of 0.3 mm, for the left-hand and right-hand channels (*fig. 8*). This division of the tracks makes monaural and stereo recordings fully compatible. Crosstalk from one channel to the other is minimal; the channel separation

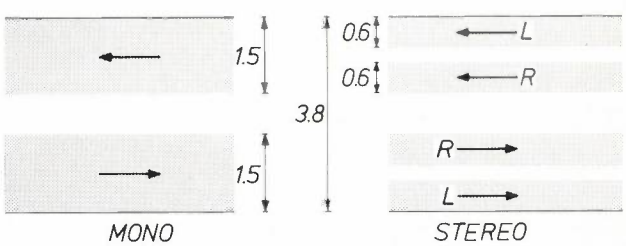


Fig. 8. Arrangement of the sound tracks on a cassette tape. *L* left-hand channel, *R* right-hand channel.

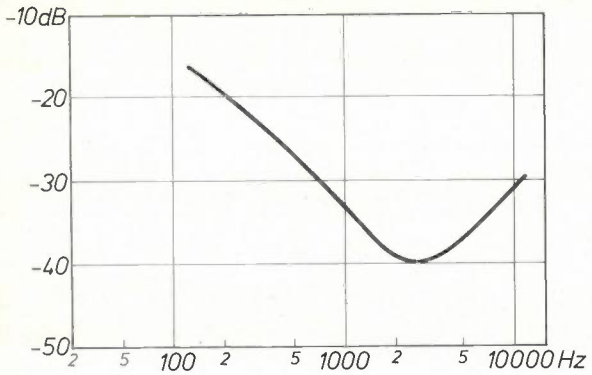


Fig. 9. Separation between the two channels for playback of a stereophonic programme on a Compact Cassette.

wound on to hubs *1* and fastened to them with a clamping-piece *2*, shaped in such a way as to preserve as far as possible the circular shape of the hub. A lining *3* inside the cassette holds the tape in place at both sides. To keep the power required for fast winding as low as possible, the tape is not threaded over fixed guide spindles but over rollers *4*. When the cassette is loaded into the device and tape transport is switched on, the erase head *5* and the record/playback head *6* are brought into contact with the tape through openings in the cassette; at the same time the pressure roller *7* is pressed against the driving spindle or capstan *8*. Each cassette contains high-permeability screening *9*, which

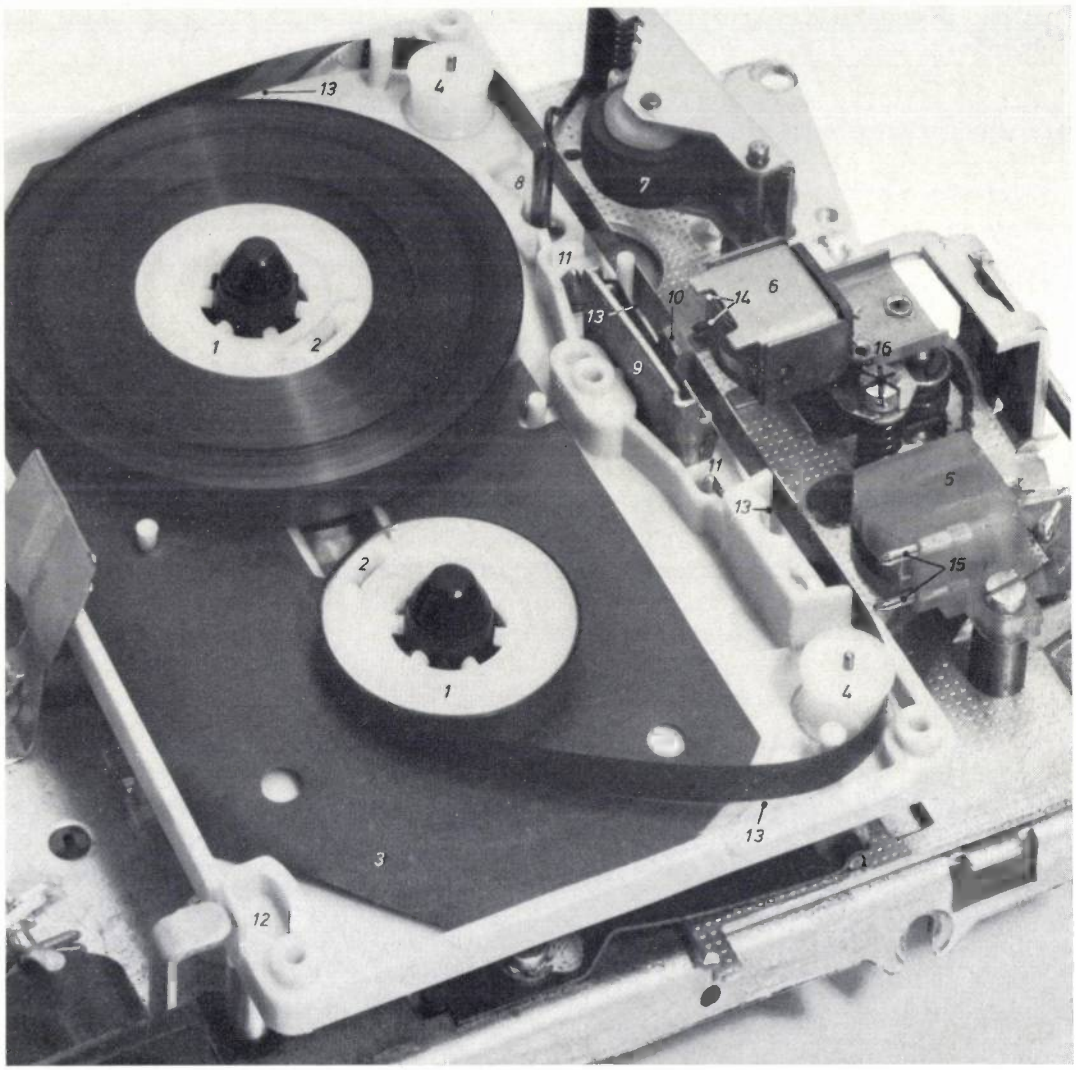


Fig. 10. View inside a cassette in a cassette player. *1* hubs. *2* clamping piece. *3* lining. *4* rollers. *5* erase head. *6* record/playback head. *7* pressure roller. *8* capstan. *9* high-permeability screening. *10* felt pressure pad. *11* reference holes and pins. *12* recording lock. *13* support points for tape transport. *14* tape guides. *15* tape guides. *16* screw for adjusting azimuth.

is presented in *fig. 9* as a function of frequency. The channel separation is more than sufficient to give an unimpaired stereo effect and is in fact greater than the separation on stereo gramophone records.

Fig. 10 shows the interior of a cassette. The tape is connects up with the screening in the cassette player to protect the head against stray fields, and a felt pressure pad *10*, which presses the tape against the head with a force of 0.1 N to 0.2 N. Two reference holes *11* correspond to the two locating pins in the cassette player; a

spring pushes the cassette towards the heads. A recording can be protected against accidental erasure by breaking the lip of the recording lock *12*; there is one lock for each track. A pawl in the cassette equipment then falls into the hole vacated by the lip and prevents recording. At both ends of the tape there is usually a thicker polyterephthalate strip which takes the strain when the tape suddenly stops at the end of fast forward winding or rewinding.

A few support points *13* help to guide the tape in the cassette. To give more accurate guidance of the tape past the head, the capstan draws the tape between two accurately aligned guides *14* fitted to the housing of the record/playback head. There are two similar guides *15* on the erase head. We have already mentioned the importance of accurate guidance of the tape in avoiding azimuth errors. The azimuth of the record/playback head is individually adjusted in every cassette player by means of an adjusting screw *16*.

Philips cassette equipment

Transport mechanism

The transport mechanism of a cassette recorder or player has to feed the tape over the record/playback head at an accurately constant speed. Fluctuations in this speed, due for example to non-circularity of rotating parts, result in fluctuations of pitch in the reproduced music, which can be very annoying and are referred to as "wow" or "flutter". Apart from feeding the tape at a constant speed, the transport mechanism for battery equipment also has to take very little current. In addition the mechanism must be accommodated in a small volume of prescribed shape.

The tape is transported by the capstan, with a rubber roller pressing the tape against the capstan. For uniform tape feed it is of prime importance that the motor driving the capstan should run at a constant speed, which does not decrease when the battery voltage drops or when the mechanical load increases. During playback the load does increase since the radius of the roll of tape becomes smaller as it unwinds while the braking torque acting on it remains constant; this causes an increase in the tension on the tape. To keep the motor speed constant, Philips cassette machines contain an electronic control circuit (*fig. 11*). This circuit keeps the tape speed constant to within 0.5% during varying load, and continues to operate as long as the battery voltage, which is 7.5 V nominal, has not dropped below 5 V.

In the control circuit the motor M forms one of the branches of a bridge circuit whose out-of-balance voltage is applied across the emitter-base junction of a transistor *Tr 1*. This transistor drives a second one, *Tr 2*, which determines the current through the whole bridge. If the motor speed changes, the opposing voltage induced in the motor coil changes in proportion to the speed, and this change reacts on the balance of the bridge in such a way as to oppose the speed variation. This is accompanied by a change in the current through the motor and hence by a change in the voltage drop across the ohmic resistance in the motor. The resistor R_1 is given a value such that the voltage drop across it balances the drop across the ohmic resistance of the motor. The copper-wound resistor R_{Cu} has the correct temperature dependence to compensate various temperature effects in the motor and semiconductors.

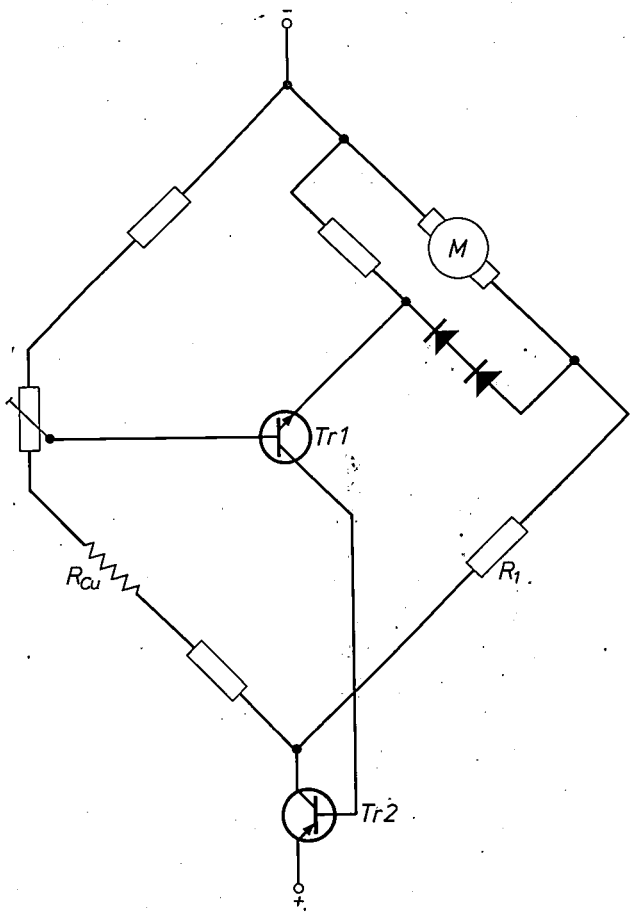


Fig. 11. Electronic control circuit which keeps the speed of motor M constant with decreasing battery voltage and varying load. A decrease in motor speed causes a decrease in the induced voltage generated in the motor; this puts the bridge circuit out of balance, causing transistor *Tr 1* to drive transistor *Tr 2* in a direction such that the current through the bridge circuit increases, and with it the motor speed. The temperature dependence of the copper-wound resistor R_{Cu} compensates the sum of a number of temperature effects in the motor and semiconductors.

A motor speed control system prevents slow variations in tape speed, but not the faster variations that cause wow and flutter.

To suppress these faster variations a flywheel is fitted to the capstan of a tape device. Because of the limited space available in the cassette transport mechanism, the flywheel can only be fairly small in diameter (see *fig. 12*), but this is compensated by using a thin, fast-running capstan. This has the disadvantages that the capstan has to be machined even more accurately and

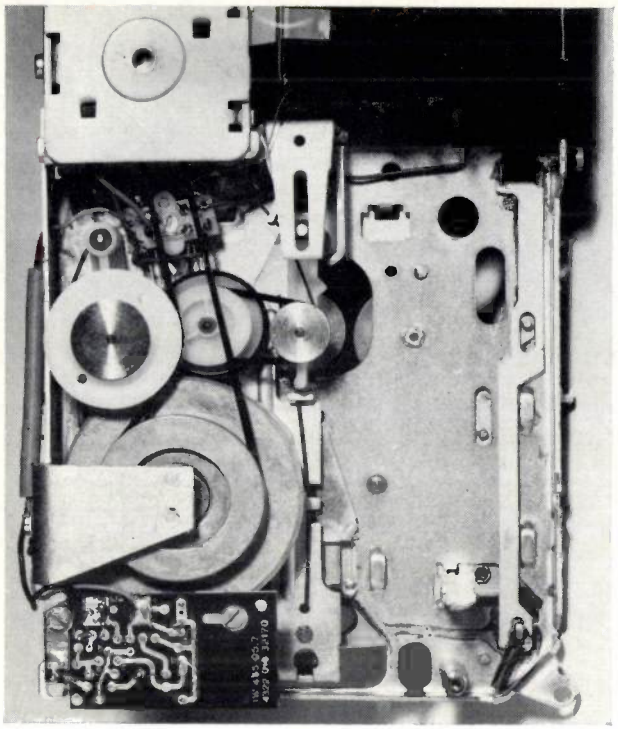


Fig. 12. View of the cassette transport mechanism from below.

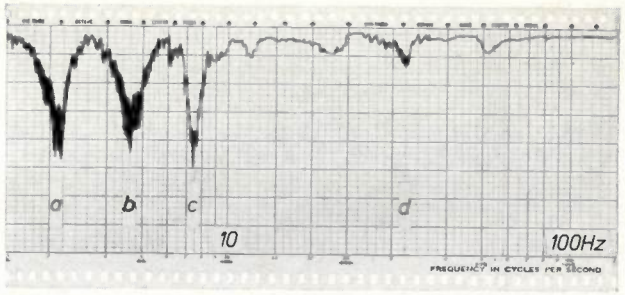


Fig. 13. Frequency analysis of the flutter in a cassette player. The amplitude is marked on the scale from top to bottom. Four frequency components may be observed, originating from the rubber belt (*a*), the slipping clutch (*b*), the capstan (*c*) and the motor (*d*).

that it may bend under the lateral force exerted by the pressure roller. Misalignment of the capstan affects the tape feed, causing effects such as azimuth error. In Philips cassette equipment the compromise taken is to make the capstan diameter 2 mm and the speed about 7.5 revolutions per second.

The capstan with flywheel is driven by a rubber belt and pulley system. This gives better damping of any fast ripple in the speed of the motor than a system of intermediate pulley wheels. Ripple of this kind can occur because the rotating rotor tends to "cling" at certain angular positions. Another advantage of such a drive system is that it gives greater freedom in the layout of a tape-transport mechanism. For these advantages we are prepared to accept the disadvantages of possible flutter due to variations of belt thickness or slipping of a stretched belt. Precision-ground belts are used that have thickness variations of less than 2%. The same belt drives both the flywheel and the take-up reel, but the take-up reel is driven via a slipping clutch which is needed because the speed of revolution of the take-up reel decreases as it fills up with tape.

To measure the magnitude of the flutter, which consists of frequency modulation of the tones recorded on the tape, these tones are demodulated. The modulating signal can then be analysed into its constituent frequencies. A frequency analysis carried out in this way for the flutter of a cassette player is shown in *fig. 13*. The four principal frequency components are seen to lie at the rotational frequency of the belt (about 3 Hz), the slipping clutch (4.6 Hz), the capstan (7.5 Hz) and the motor (32 Hz).

The magnitude of the flutter is expressed as the percentage frequency variation of a tone. The annoyance caused by flutter depends on the rapidity of the variations; the most annoying are variations with a frequency in the region of 4 Hz. All frequency components in the flutter are thus not equally troublesome. It is customary to "weight" them against a certain specification, such as the widely used weighting curve (*fig. 14*) laid down in the German standard DIN 45507. For the simple Philips cassette players the flutter, when weighted from this curve, must be no more than 0.4%; more expensive equipment has to meet a stiffer specification.

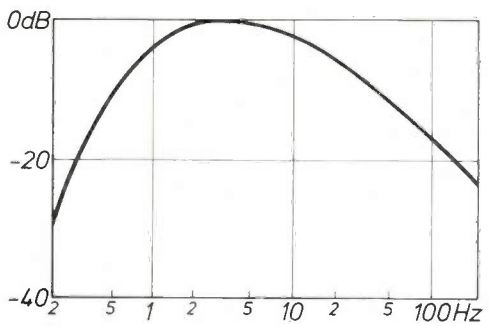


Fig. 14. Weighting curve after DIN 45507; which shows the weighting that the various frequency components of the flutter in a sound recording have to be given in order to express the flutter in a single numerical value corresponding to the degree of annoyance experienced.

Current consumption of the transport mechanism

The tape-transport mechanism in the Philips cassette player takes a current of 75 mA. When the battery voltage has dropped to 5 V, the electrical power supplied to the player is distributed among the components of the mechanism as shown in *fig. 15*. The battery voltage is usually higher than 5 V; the speed control then has an appreciably larger share, because it dissipates the surplus power. The friction in the bearings of

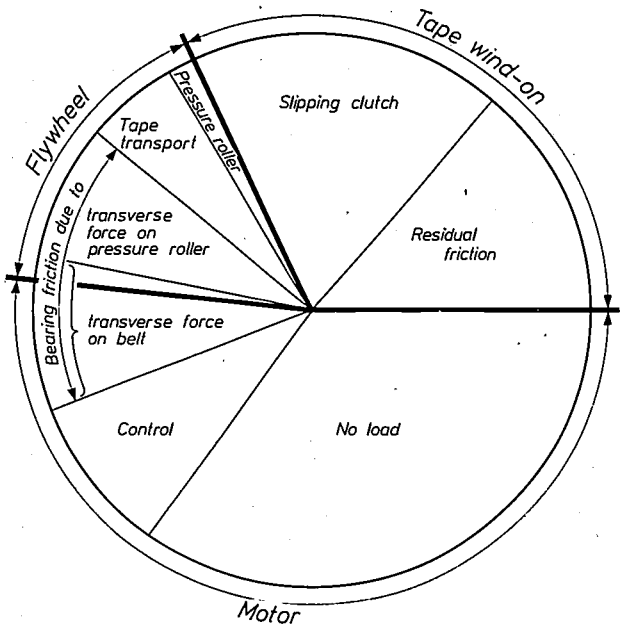


Fig. 15. Distribution of electrical power among the components of the transport mechanism when a tape cassette is played on a Philips cassette player. The fractions of the bearing friction due to the lateral forces originating from the rubber belt and pressure roller are presented separately. The distribution shown here relates to the situation when the battery voltage has decreased to 5 V. At a higher battery voltage the share of the control system is much larger.

motor and flywheel is considerably increased by the lateral forces exerted by the drive belt and the pressure roller; the flywheel bearing friction is in fact almost entirely due to these lateral forces.

Electroacoustic characteristics

The electroacoustic characteristics of a tape player, such as frequency response and signal-to-noise ratio, are determined not only by the characteristics of the tape and the tape geometry (see fig. 4) but also by the processing which the signals presented for recording and reproduction receive in the electronic circuits of the device. The frequency characteristic resulting from the recording and playback process is not usable until it has been corrected (or "equalized") in these circuits. The purpose of the equalization may be given a somewhat wider formulation: to obtain the best signal-to-noise ratio, maximum use must be made at all frequencies of the dynamic range of the tape, and the replay output characteristic must be flat within the widest possible range of frequencies.

The dynamic range of the tape is not identical for all tapes and at higher frequencies it also varies with the bias current (see page 83). To avoid a situation in which every manufacturer introduced his own corrections, and tapes would not necessarily be interchangeable, an international standard has been laid down for the variation of the magnetization on the tape with frequency, for a constant signal at the input of the recording amplifier [10]. The magnetization is defined in this standard as the magnitude of the "surface induction", which is the flux density at right angles to the surface of the tape when the tape is moved along against an ideal reproducing head. Different surface-induction frequency curves have been laid down for the various standard tape speeds, and curve *M* in *fig. 16* is an example for the tape speed of 4.76 cm/s. In practice the playback amplifier is equalized with the aid of a reference tape, taken to be magnetized in accordance with this standard; the amplifier is given a frequency characteristic such that when the reference tape is played, a virtually frequency-independent signal amplitude appears at the output. The replay characteristic of the Philips cassette players is given by curve *P* of fig. 16.

When the surface induction is the same at different frequencies, then — neglecting for a moment the losses indicated in fig. 4 — the induced voltage at the terminals of the playback head is also the same at these different frequencies. This explains why the playback characteristic *P* in fig. 16 is very like the mirror image of the magnetization curve *M*. The standard curve *M* thus determines to a large extent the playback characteristic of a cassette player; at high frequencies, however, *P* rises steeply to offset the gap losses occurring on playback (fig. 4, curve *B*), and this high-frequency equalization may differ from one type of cassette player to another.

What frequency characteristic should we now give to the recording amplifier to ensure that the overall frequency characteristic is flat when a tape is played back through the playback amplifier corrected in the manner described? The answer to this question depends on how large we decide to make the bias current. We have explained on page 83 that the relative levels at which low and high frequencies are recorded depend on the magnitude of the bias current, and that the maximum output level of the tape at high frequencies also depends on the bias current. With a higher bias current the high fre-

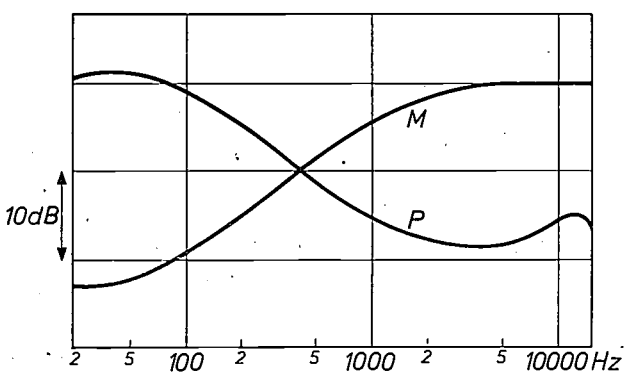


Fig. 16. *M* internationally standardized curve [10] showing the variation of tape magnetization with frequency for constant voltage at the input of the recording amplifier; the curve relates to a tape speed of 4.76 cm/s. *P* frequency characteristic of the playback amplifier in the Philips cassette player.

[10] IEC Publication 94.

quencies give a smaller recorded signal, and at the same time the maximum undistorted output level of the tape decreases at high frequencies. On the other hand, we also saw that good recording of the low frequencies requires a relatively high bias current. If we are to meet this requirement as well we shall have to compensate the attenuation of the high frequencies by an appropriate high-frequency "lift" upon recording. This can in fact be done without exceeding the maximum output level of the tape, since the energy content of the higher frequencies (higher than 2 kHz) in an average music signal is smaller than the energy content of the lower frequencies. The aim, however, is not only to obtain an optimum frequency response but also to achieve an optimum distortion-free modulation of the tape. We try to achieve this by setting the bias current to a value at which the low-frequency and high-frequency signals in an average music signal both approach the limit for distortion-free modulation on the tape to about the same extent after the high-frequency compensation in the recording amplifier. If a signal exceeds the permissible amplitude, then there will be simultaneous distortion of both low and high frequencies. If the bias current is higher than this optimum value, more high-frequency compensation will be needed; this conflicts, however, with the reduced modulation limit of the tape and the high frequencies are the first to show distortion. At too low a bias current the low frequencies become distorted first.

The term "average music signal" used here suggests that practical experience must also have a say in the choice of bias current and high-frequency compensation. In the Philips cassette player the recording amplifier has been given the frequency characteristic shown by curve *R* of *fig. 17*. This lifts not only the high frequencies but also the frequencies below 200 Hz. This corresponds with the shape of curve *M* in this region, and the object is to improve the ratio of the signal to any hum in equipment supplied from the mains. The frequency characteristic resulting from all the corrections applied in recording and playback is represented by curve *T* of fig. 17. The corresponding signal-to-noise ratio for present-day players and tapes is about 45 dB.

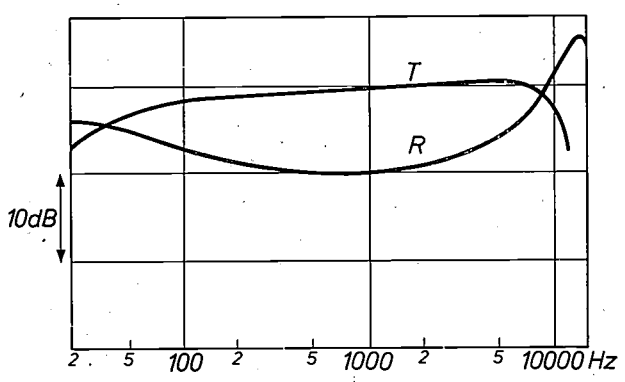


Fig. 17. *R* frequency characteristic of the recording amplifier in the Philips cassette player. *T* overall frequency characteristic, measured via recording and playback (not including microphone and loudspeaker).

"Musicassettes"

All musicassettes are prerecorded with stereo programmes. To speed up production the music programmes on the master tape are not transferred to the cassette tapes at the nominal tape speed but at a considerably higher speed, 32 times higher in the latest production machines. This means that the master tape, which is modulated at a speed of 19.05 cm/s, is played at the rather astonishing speed of slightly more than 6 metres per second. The cassette tapes, four of which are modulated simultaneously with the signal from the master tape, are run at a speed of more than 1.5 m/s. All four tracks are prerecorded simultaneously, one pair from back to front.

During the copying process all frequencies are multiplied by a factor of 32; the amplifiers cover a frequency range of 200 Hz to 500 kHz and the playback and recording characteristics are correspondingly transformed. The high-frequency bias current has a frequency of 2.4 MHz. Ferrite heads are used to avoid eddy current losses and rapid wear; the gap length of the recording heads is 4 μm.

Large numbers of programmes are successively played on to a single cassette tape 1500 metres long, the programmes being punctuated by signal tones which act as "cues" in the semi-automated assembly process for the cassettes. It would exceed the scope of this article to go into the details of this process [11]. During recording great care is taken to minimize flutter, which is less than 0.06% when properly weighted by the curve of fig. 14. The perpendicular alignment of the recording gap is also carried out very accurately, the angular errors being less than 2′. This is very important, because azimuth errors in playback equipment can seriously impair the quality of reproduction when musicassettes are played; everything possible is done to ensure that the recording process contributes as little as possible to azimuth errors that may ultimately exist.

Pocket Memo

The development of the Philips Pocket Memo (fig. 2) was prompted by the desire to market an extremely small dictation machine that could be carried in the pocket and used for making quick verbal notes. Small size was the principal requirement which the design had to meet; a playing time of 2 × 10 minutes (or 2 × 15 minutes with

[11] For further information see J. L. Ooms, Multiple speed tape duplicating, J. Audio Engng. Soc. **14**, 343-355, 1966.

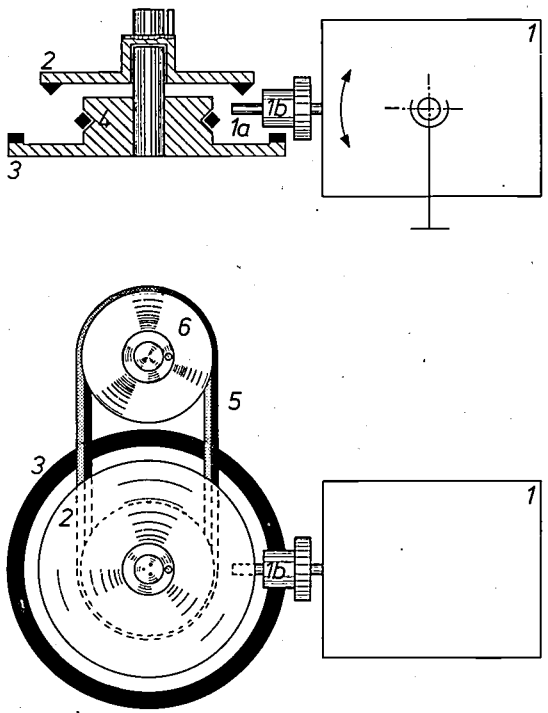


Fig. 18. Transport mechanism of Pocket Memo. *1* motor, *1a, b* pinions on motor shaft. *2* turntable of take-up reel. *3* driving wheel for supply reel with *4* pulley wheel. *5* rubber belt. *6* turntable of supply reel.

9 μm tape) was desired, and also a rewind facility. A cassette was here the obvious means of making the audio tape manageable. We have already mentioned above that the Pocket Memo does not use a capstan for tape transport but uses instead the mechanically much simpler system of transporting the tape directly by means of the take-up reel. Apart from saving space and simplifying the transport mechanism, this has the advantage that less current is required. This is because there is no need of the slipping clutch used in capstan drive of the take-up reel, and this slipping clutch usually takes a fairly large amount of power (see fig. 15 which shows the distribution of the power taken by the transport mechanism of the cassette).

Transport mechanism

The simplicity of the transport mechanism in the Pocket Memo can clearly be seen in *fig. 18*. Attached to the shaft of the motor *1* is a double pinion *1a, b*. In the "record/playback" position the motor shaft is tilted upwards and presses pinion *1a*, which has a diameter of 1 mm, against a rubber rim on the turntable *2* of the take-up reel. In the "rewind" position the motor shaft is tilted downwards, and pinion *1b* runs on the rubber rim of wheel *3*. Mounted on the same spindle as wheel *3* is a pulley wheel *4* which drives the supply reel via belt *5* in the direction to wind the tape back on to the reel.

The rotating parts are given rotational frequencies that are as far removed as possible from the value of 4 rev/s (*fig. 19*), since the human ear is most sensitive to flutter at a frequency of 4 Hz (page 88). A frequency analysis of the flutter is shown in *fig. 20*, in which the motor frequency f_m and some of its harmonics can be seen.

The Pocket Memo is provided with the same electronic motor speed control as the cassette player. The circuit is shown in fig. 11. In the Pocket Memo a signal is derived from the control circuit to give an indication of the battery voltage. Superimposed on the collector d.c. voltage of *Tr 2* is an a.c. voltage due to the commutation of the d.c. motor. By means of a battery-check switch this voltage can be applied to the playback amplifier. If the battery voltage falls too low, there is no longer any voltage across *Tr 2*, which then becomes effectively a short-circuit to earth, and consequently the a.c. voltage will no longer give an audible signal.

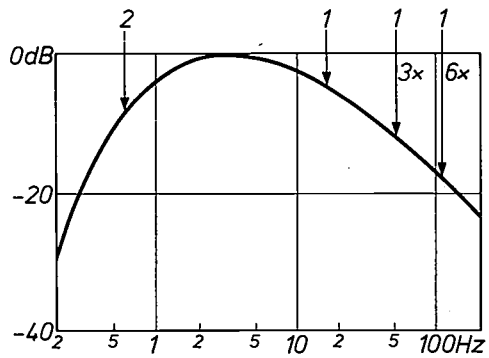


Fig. 19. The motor *1* and the take-up reel *2* (see fig. 18) are given rotational frequencies which are as far removed as possible from the frequency 4 Hz, which is the most annoying flutter frequency. An indication is given of where they lie in relation to the weighting curve of fig. 14. Unbalance of the rotor windings, magnetic "cling" at certain angular positions and commutator effects can lead to fluctuations with frequencies of 3 and 6 times the motor speed.

Tape modulation, frequency characteristic

The Pocket Memo incorporates a moving-coil microphone, which also acts as loudspeaker when the recording is played back. To ensure a good signal-to-noise ratio it is desirable that the tape should be modulated to a reasonable output level, irrespective of the distance from which the microphone is spoken into. In most tape recorders, and in the cassette players as well, the recording level is adjusted by hand; in the Pocket Memo, however, this would make it unacceptably complicated to use. The Pocket Memo is therefore given a fixed recording level, which is set fairly high; signal limitation is provided by the natural saturation of the tape. Steps have been taken to ensure that the recording amplifier remains well within the limits of its range of linear operation even when the tape is modulated into saturation, so that the tape is in fact the only limiting element. The manner in which the tape limits the signal gives less annoying non-linear distortion than overdriving an amplifier, and the intelligibility of speech is not so greatly affected.

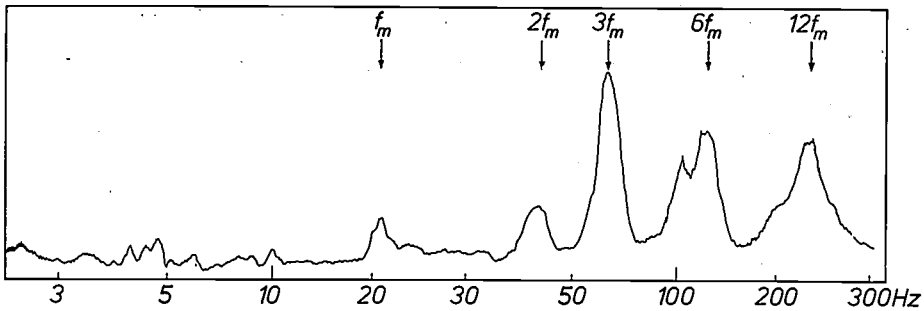


Fig. 20. Frequency analysis of the flutter in the transport mechanism of a Pocket Memo. In the frequency band from 2.5 Hz to 300 Hz shown here the flutter originates entirely from the motor. This turns at a speed of f_m revolutions per second; the component at the frequency 3 f_m is due to magnetic unbalance of the rotor windings, the 6 f_m frequency component is due to magnetic "cling" and commutator effects, and the 12 f_m component is due to the commutator alone.

Intelligibility of speech is again the significant factor that should be borne in mind when examining *fig. 21*, which shows the frequency characteristic of the Pocket Memo, measured at the output terminals with an ohmic load. The frequency characteristic does not therefore comprise the microphone/loudspeaker; it can be seen that it favours the frequencies which are of main importance for the intelligibility of speech.

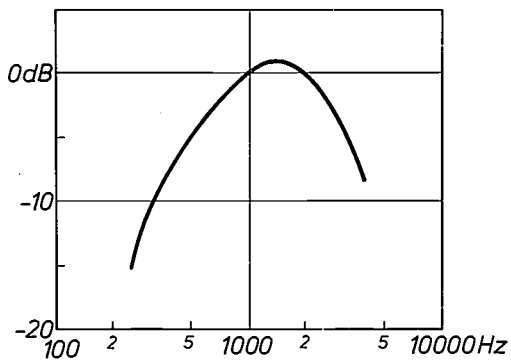


Fig. 21. Frequency characteristic of the Pocket Memo for recording and playback, excluding the frequency response of the microphone/loudspeaker. The characteristic favours the frequencies important for intelligibility of speech.

Summary. Because they are easy to manipulate and provide effective protection of the tape, audio tape cassettes are coming into ever-increasing use. Largely because of the continued improvement of tapes, it is now possible to use simple mass-produced cassette equipment to record frequencies up to 10 kHz at a tape speed of 4.76 cm/s, and to obtain a signal-to-noise ratio of 45 dB on a track width of 1.5 mm. This has led to the development of the Philips Compact Cassette with a playing time of 2 × 60 minutes (on a tape 9 μm thick); it is also marketed with a stereophonically prerecorded tape (Musicassette, maximum playing time 1 × 45 min). Cassette recorders and players are mostly portable and can work from batteries; an electronic control circuit keeps the motor speed constant as the battery voltage decreases.

The Philips miniature dictating machine, the Pocket Memo, is equipped with an even smaller cassette, specially developed for this machine, which gives a maximum playing time of 2 × 15 min. The tape is transported directly by the take-up reel; with this system the cassette and machine can be kept small and simple, and the current required is small.

A non-destructive measurement of pressures in incandescent lamps

In the development of incandescent lamps with an internal frosting layer of "Aerosil" powder [1] an initial difficulty was the short life of the experimental lamps. It was soon suspected that the difficulty was due to water vapour being released from the free surface of the frosting layer (amounting to about 20 m² in each lamp). It is known that tungsten filaments react with water vapour. A small amount of water vapour is sufficient to transport a considerable amount of tungsten to the bulb wall (in the "water cycle"). The release of an excessive amount of gas in an incandescent lamp can also lead to electrical breakdown.

Tests confirmed this suspicion: at the usual bulb-wall temperature of 150 °C it was found that the pressure in the frosted experimental lamps was 10 to 12 times higher than in the lamps without frosted bulb. Measurements of this type, using the method described below, also made it possible to establish which pumping process is needed to produce lamps that do have a sufficiently long life.

In performing the measurements we were in the almost ideal situation of being able to use the lamp filaments as a Pirani pressure gauge [2]. We were thus able to measure pressures of about 0.1 to 10 pascals[3] in the experimental lamps without having to interfere with them structurally in any way. *Fig. 1* shows the test circuit; the filament forms one arm of the bridge. The resistance of the filament is measured by balancing the bridge, with an electronic voltmeter as the null indicator. The pressure can be found from this resistance by using a calibration chart. The calibration curve depends on the temperature of the bulb wall.

The calibration charts were derived from measurements with a calibrated pressure gauge in lamps that had been opened and filled with an adjustable quantity of air. To obtain useful calibration charts the filament had to be 100 to 200 °C hotter than the bulb wall.

Everything depended, of course, on the assumption that a calibration chart established for one particular lamp would also be valid for the other lamps. In practice this assumption was sufficiently accurate, provided that the filaments of the lamps were coiled from the same wire on the same mandrel, and on the same machine (*fig. 2*). Measurements and calibration were of course carried out at the same bulb temperature.

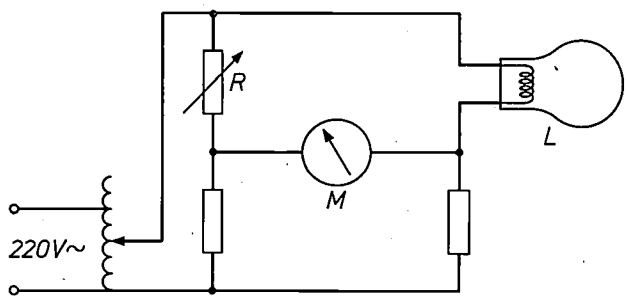


Fig. 1. Non-destructive pressure measurement in an incandescent lamp, using the filament as a Pirani gauge. The bridge circuit is used for measuring the resistance of the filament of the lamp *L*. This resistance is a measure of the gas pressure in the bulb, since the gas pressure affects the temperature of the filament and hence its resistance. *M* electronic voltmeter, acting as null indicator for the bridge. *R* balancing resistor.

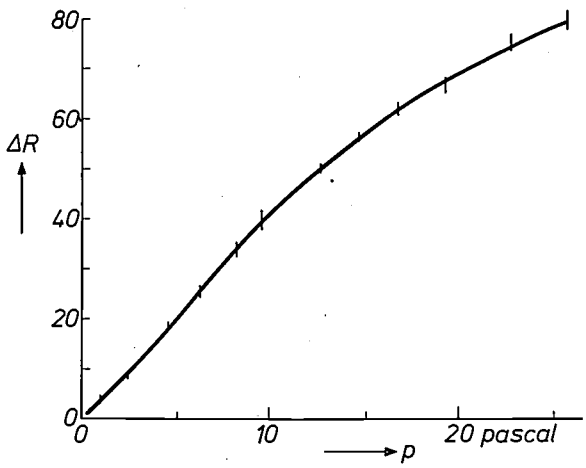


Fig. 2. The result of calibration measurements on seven test lamps whose filaments were wound in the same way from the same wire. The horizontal axis shows the selected pressure *p*, and the vertical axis the decrease ΔR in the balancing resistance (in arbitrary units). The scatter in the measured points, indicated by vertical dashes, is extremely slight.

The method of measurement described has also been used for monitoring in a lamp-production unit. During a period of a year we made a routine measurement of the pressure in the lamps immediately after evacuation. This check provided valuable information about the reliability of the pumping process.

H. J. H. Beuvens
J. H. Dettingmeyer

[1] "Aerosil" (SiO_2) is a Registered Trade Mark of Degussa of Frankfurt-am-Main. The specific surface of "Aerosil" powder is about 200 m² per gramme.

[2] See for example M. Pirani and J. Yarwood, Principles of vacuum engineering, Chapman and Hall, London 1961.
A recent article on the operation of a Pirani gauge is the one by L. Heijne and A. T. Vink: A Pirani gauge for pressures up to 1000 torr and higher, Philips tech. Rev. **30**, 166-169, 1969 (No. 6/7).

[3] The SI unit of pressure, 1 N/m², is the pascal (Pa); 1 torr = 1/760 atm = 133.322 pascals.

H. J. H. Beuvens and J. H. Dettingmeyer are 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.

H. J. Akkerman (Philips Research Laboratories Amsterdam): Draaibank met numerieke maatindicatie voor enkelfabricage in de fijnmechanische techniek.
Metaalbewerking **35**, 299-305, 1969 (No. 13).

V. Belevitch: Minimum-gyrator cascade synthesis of lossless *n*-ports.
Philips Res. Repts. **25**, 189-197, 1970 (No. 3). *B*

K. Bethe & J. Verweel: On dielectric losses at microwaves: sensitive measurements and results.
IEEE Trans. **MAG-5**, 474-477, 1969 (No. 3). *H*

R. Bleekrode & W. van Benthem: Spectroscopic investigations of high-current hollow-cathode discharges in flowing nitrogen at low pressures.
J. appl. Phys. **40**, 5274-5280, 1969 (No. 13). *E*

P. F. Bongers: Faraday rotation of infrared and visible light by magnetic materials.
IEEE Trans. **MAG-5**, 472, 1969 (No. 3). *E*

J. Bootsma: The effect of viscosity variations with temperature on the performance of spiral groove bearings.
ASLE Trans. **12**, 287-296, 1969 (No. 4). *E*

J. C. Brice, O. F. Hill & P. A. C. Whiffin: A modification to the Czochralski method of crystal pulling.
J. Crystal Growth **6**, 297-298, 1970 (No. 3). *M*

J.-J. Brissot: Des cristaux pour les lasers.
Atomes **265**, 303-306, 1969. *L*

J. van den Broek & A. Netten: Photoconductivity in lead-tin monoxide.
Philips Res. Repts. **25**, 145-154, 1970 (No. 3). *E*

H. B. G. Casimir: Remarks on the formulation of magnetic theory.
IEEE Trans. **MAG-5**, 159-161, 1969 (No. 3). *E*

J. A. Clarke: Light-gathering properties of curved fibre optic faceplates.
Optica Acta **17**, 123-130, 1970 (No. 2). *M*

W. F. Druyvesteyn & A. J. Smets: Experimental demonstration that the electron velocity is normal to the Fermi surface.
Physics Letters **31A**, 1-2, 1970 (No. 1). *E*

G. Engelsma: Photoinduction of phenylalanine deaminase in gherkin seedlings, IV. The role of the temperature.
Planta **90**, 133-141, 1970 (No. 2). *E*

U. Enz, W. Lems, R. Metselaar, P. J. Rijnierse & R. W. Teale (University of Sheffield, England): Photomagnetic effects.
IEEE Trans. **MAG-5**, 467-472, 1969 (No. 3). *E*

L. Fraiture & J. Neirynck: Theory of unit-elements filters.
Rev. HF **7**, 325-340, 1969 (No. 12). *B*

J. G. C. de Gast: Dynamic behavior of double film hydrostatic bearing with variable flow restrictor.
ASME Winter Annual Meeting, Nov. 1969, Los Angeles, paper 69-WA/Lub-12; 12 pp. *E*

A. A. van der Giessen & C. J. Klomp: The preparation of iron powders consisting of submicroscopic elongated particles by pseudomorphic reduction of iron oxides.
IEEE Trans. **MAG-5**, 317-320, 1969 (No. 3). *E*

A. van de Grijp, T. Holtwijk & R. M. G. Wijnhoven: SATAN: a versatile memory array tester.
IEEE Trans. **MAG-5**, 656-661, 1969 (No. 3). *E*

G. Groh & M. Kock: 3-D display of X-ray images by means of holography.
Appl. Optics **9**, 775-777, 1970 (No. 3). *H*

H. J. M. de Haan: THEMIS, three-hole element memory with integrated selection.
IEEE Trans. **MAG-5**, 403-407, 1969 (No. 3). *E*

N. Hazewindus: Calculation of particle trajectories in a cyclotron axial injection system with an electrostatic deflector.
Nucl. Instr. Meth. **76**, 273-277, 1969 (No. 2). *E*

J. C. M. Henning & H. van den Boom: An orthorhombic Fe^{2+} centre in Cs_3ZnCl_5.
Philips Res. Repts. **25**, 155-170, 1970 (No. 3). *E*

P. Holst & M. Lemke: Ferrite substrates for microwave integrated systems.
IEEE Trans. **MAG-5**, 478-480, 1969 (No. 3). *H*

E. P. Honig & J. H. Th. Hengst: The point of zero charge and solid-state properties of silver bromide.
J. Colloid Interface Sci. **31**, 545-556, 1969 (No. 4). *E*

H.-G. Junginger & W. van Haeringen: Energy band structures of four polytypes of silicon carbide calculated with the empirical pseudopotential method.
Phys. Stat. sol. **37**, 709-719, 1970 (No. 2). *A, E*

E. M. H. Kamerbeek: On the theoretical and experimental determination of the electromagnetic torque in electrical machines.
Thesis, Eindhoven 1970. *E*

F. M. Klaassen: On the geometrical dependence of $1/f$ noise in MOS transistors.
Philips Res. Repts. **25**, 171-174, 1970 (No. 3). *E*

F. M. Klaassen & J. Prins: Noise of field-effect transistors at very high frequencies.
IEEE Trans. **ED-16**, 952-957, 1969 (No. 11). *E*

A. Klopfer: Desorption durch Elektronenbeschuß von Wolfram mit adsorbiertem Wasser.
Vakuum-Technik **19**, 37-42, 1970 (No. 3). *A*

A. Klopfer: Desorption durch Elektronenbeschuß von Molybdän mit absorbierten Gasen: H_2, O_2, H_2O.
Surface Sci. **20**, 129-142, 1970 (No. 1). *A*

J. A. Kok: The corona ignition voltage of an electrical discharge in a gas.
Appl. sci. Res. **21**, 303-305, 1969 (No. 3/4). *E*

P. van der Laan (Philips Information Systems and Automation Division, Eindhoven): Simple distribution-free confidence intervals for a difference in location.
Thesis, Eindhoven 1970.

J. Loeckx: Analyse syntaxique et transducteurs à pile.
Math. Comm. Twente Univ. **5**, 1-20, 1969 (No. 1). *B*

J. Loeckx: Grammaires formelles et automates.
Bull. Soc. Math. Belg. **21**, 145-165, 1969 (No. 2). *B*

F. A. Lootsma: Boundary properties of penalty functions for constrained minimization.
Thesis, Eindhoven 1970. *E*

J.-M. Martinache (Faculté des Sciences de Lille), **A. Semichon, E. Constant** (Fac. Sci. Lille) **& A. Vanoverschelde** (Fac. Sci. Lille): Sur le bruit de fond d'une diode à avalanche à barrière métal - semi-conducteur.
C.R. Acad. Sci. Paris **269B**, 644-647, 1969 (No. 14). *L*

A. Mijnheer: Compact counterflow heat exchanger has high thermal efficiency.
Process Engng. June 1969, 85-88. *E*

J. Monin (Conservatoire National des Arts et Métiers, Paris), **J. Houdard & G.-A. Boutry** (Cons. Nat. A. et M.): Nouvelle méthode pour la détermination des paramètres définissant une vibration elliptique.
C.R. Acad. Sci. Paris **270B**, 200-203, 1970 (No. 3). *L*

K. Mouthaan: Nonlinear analysis of the avalanche transit-time oscillator.
IEEE Trans. **ED-16**, 935-945, 1969 (No. 11). *E*

D. de Nobel & H. G. Kock: A silicon Schottky barrier avalanche transit time diode.
Proc. IEEE **57**, 2088-2089, 1969 (No. 11). *E*

L. G. Pittaway: Computer aided design of a new mass spectrometer ion source.
Proc. Int. Conf. on Ion Sources, Saclay 1969, pp. 449-458. *M*

J. E. Ralph: Cathodoluminescence of Er^{3+} in YGaG and YAG.
J. Phys. Chem. Solids **31**, 507-516, 1970 (No. 3). *M*

G. W. Rathenau: Reflections on the INTERMAG conferences.
IEEE Trans. **MAG-5**, 154-156, 1969 (No. 3). *E*

A. G. Roederer (SODERN, Paris): Calculation of the electromagnetic field radiated by a log-periodic dipole antenna.
Philips Res. Repts. **25**, 175-188, 1970 (No. 3).

H. Schweppe: Elastic and piezoelectric properties of paratellurite (TeO_2).
Ultrasonics **8**, 84-87, 1970 (No. 2). *A*

J. W. Slotboom: Iterative scheme for 1- and 2-dimensional d.c.-transistor simulation.
Electronics Letters **5**, 677-678, 1969 (No. 26). *E*

M. J. Sparnaay: Analogies between semiconductors and electrolyte solutions (in particular of surface effects due to space charges and Gouy layers).
J. Chimie phys., No. spécial 19e Réunion, 87-99, 1969. *E*

J. Verweel: On the HF permeability of dense ferrites in polarizing fields.
IEEE Trans. **MAG-5**, 622-625, 1969 (No. 3). *H*

J. F. Verwey: The introduction of charge in SiO_2 and the increase of interface states during breakdown of emitter-base junction of gated transistors.
Appl. Phys. Letters **15**, 270-272, 1969 (No. 8). *E*

Q. H. F. Vrehen, H. G. Beljers & J. G. M. de Lau: Microwave properties of fine-grain Ni and Mg ferrites.
IEEE Trans. MAG-5, 617-621, 1969 (No. 3). *E*

M. V. Whelan: Leakage currents of n^+p silicon diodes with different amounts of dislocations.
Solid-State Electronics **12**, 963-968, 1969 (No. 12). *E*

R. M. G. Wijnhoven: Magnetoelectronic mass stores.
IEEE Trans. MAG-5, 640-641, 1969 (No. 3). *E*

E. E. Windsor: A sensitive high vacuum gauge using an electron multiplier.
Vacuum **20**, 7-9, 1970 (No. 1). *M*

P. L. Wodon: Data structure and storage allocation.
BIT **9**, 270-282, 1969 (No. 3). *B*

P. Wurtz: Thermographe médical.
Acta electronica **12**, 339-351, 1969 (No. 4). *L*

Contents of Philips Telecommunication Review **29**, No. 2, 1970:

P. W. L. van Iterson: Automatically tuned HF transmitter systems for 10 and 30 kW (pp. 41-54).
H. P. J. Grubben & P. Veldkamp: The small EBX 15 private branch telephone exchange using mini-reed contacts (pp. 55-63).
W. Beijnink: The 8TR 352 channel and group modulation equipment for carrier telephone systems (pp. 64-76).
T. Poulsen, H. L. Bakker & W. van Vlijmen: 12 MHz line equipment for coaxial cables type 8TR 317 in operation on the Copenhagen-Aarhus route (pp. 77-83).
H. Bouwman: Modems for data transmission over the public telephone network (pp. 84-86).

Contents of Electronic Applications **29**, No. 3, 1969:

The D13-500, a high-frequency instrument cathode-ray tube (pp. 73-76).
M. J. Köppen: Vertical deflection amplifier for 300 MHz oscilloscope (pp. 77-91).
Semiconductor dE/dx detector (pp. 92-93).
J. Nicot, M. Basque, J. C. Lavenir & A. Rousset: 500 nanosecond 3D memory (pp. 94-102).
Reliability of the ZM1200 indicator tube (pp. 103-106).
W. M. C. Swanenburg: A graphical method for finding transistor parameters (pp. 107-110).
Improved delay line for colour television receivers (pp. 115-116).

Contents of Electronic Applications **29**, No. 4, 1969:

D. J. G. Janssen: Inexpensive digital voltmeter (pp. 117-124).
F. May: Wide-range multivibrator using integrated circuits (pp. 125-128).
J. M. Rosa Bunge & B. W. Banega: Instrument for measuring incremental inductance (pp. 129-133).
Magnetic catches and seals (pp. 134-137).

Contents of Mullard Technical Communications **11**, No. 106, 1970:

D. Mason: High-grade long-life electrolytic capacitors: the 106/107 series in power supply design (pp. 122-129).
D. R. Armstrong: TTL interfacing with GRL111 and GRL101 (pp. 130-138).

Contents of Mullard Technical Communications **11**, No. 107, 1970:

C. V. Newcomb: A new adjustor for Mullard ferrite inductor cores (pp. 142-146).
J. Thörig & W. A. M. Peters: High-quality audio amplifiers (pp. 147-152).
J. Thörig: High-quality pre-amplifier (pp. 153-155).
A. Ciuciura: A simple stabilised power supply for transistorised monochrome receivers (pp. 156-167).

Volume 31, 1970, No. 3 pages 65-96 Published 26th March 1971

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No.4

The "bucket-brigade delay line", a shift register for analogue signals

F. L. J. Sangster

Information theory states that a signal of bandwidth B is completely characterized by 2B samples per second. If these values are stored in the stages of a kind of shift register (for example in the form of charges on capacitors) the attractive features of the shift register — especially as a delay line — can also be used for analogue-signal handling. The transfer of analogue information in a shift register has for years been a great problem, to which the author has now found a simple and elegant solution. By analogy with the old fire-fighting method, in which buckets of water are passed along the line, circuits of this type are called "bucket-brigade delay lines".

In the processing of analogue signals, such as audio or video signals, there have always been problems in achieving a time delay, since hitherto there has been no universal electronic method that could be used for this purpose. For operations such as amplification, modulation, detection and filtering, simple devices or circuits exist, but for delaying the signals it is generally necessary to resort to non-electronic systems. It is true that a signal can be delayed by passing it through an electrical transmission line, e.g. a coaxial cable or an *LC* network, where the distributed inductance and capacitance of a cable are lumped in a number of coils and capacitors. Such transmission lines, however, can only be used for delays of a few microseconds at the most for video signals and a few milliseconds for audio signals. As a rule the requirements are much higher for both delay and bandwidth. For some colour-television systems, for instance, a delay of 64 microseconds is required at a bandwidth of 1 MHz; for this application delay lines have been made that use the propagation of ultrasonic waves in glass [1]. In audio applications, delays of a few tens of milliseconds are sometimes required to simulate reverberation: the effect is achieved either by using a magnetic recording disc, which is rather bulky, or an arrangement in which mechanical vibrations propagate in a metal strip, spring or plate [2].

The idea of making a shift register for analogue signals and using it as a delay line dates back to the beginning of the fifties [3]. The principle of such a register is quite simple. Sampled values of the analogue signal are stored in the form of charges on a series of capacitors. Between each of these storage capacitors is a type of "switch" that transfers the charges from one capacitor to the next on a command from a clock pulse. Since each storage capacitor cannot take up its new charge until it has passed on the old one, only half the capacitors carry information and the ones in between are empty. A "bucket-brigade delay line" of this kind

F. L. J. Sangster is with Philips Research Laboratories, Eindhoven.

[1] See C. F. Brockelsby and J. S. Palfreeman, Ultrasonic delay lines and their applications to television, Philips tech. Rev. **25**, 234-252, 1963/64, and F. Th. Backers, A delay line for PAL colour television receivers, Philips tech. Rev. **29**, 243-251, 1968.

[2] See R. Vermeulen, Stereo reverberation, Philips tech. Rev. **17**, 258-266, 1955/56.

[3] J. M. L. Janssen, Discontinuous low-frequency delay line with continuously variable delay, Nature **169**, 148-149, 1952.
G. A. Philbrick, Bucket-brigade time delay — a palimpsest on the electronic analog art, Philbrick Researches, Boston, 1955.

is illustrated schematically in *fig. 1*; in (*a*) the even-numbered capacitors carry information and the ones in between are empty, and in (*b*) the information has been passed on via switches S_2 and S_4 to the odd-numbered capacitors, while a new sample has been supplied to C_1, and a sample has been delivered to the output. In the next step the information is passed on to the even-numbered capacitors. The switches are thus driven at the sampling frequency, but the even and odd

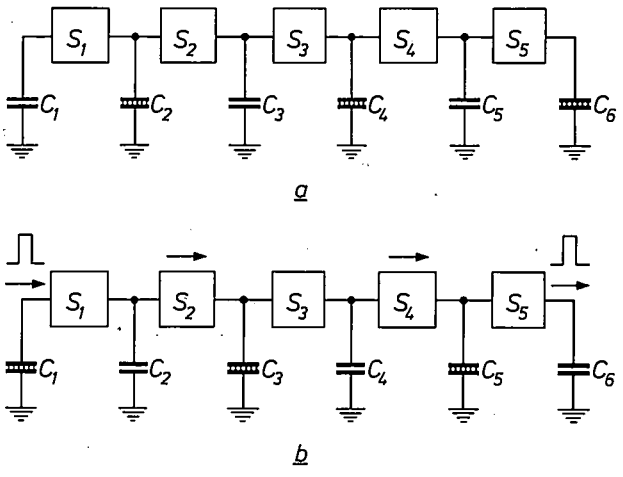


Fig. 1. The operation of an analogue shift register (bucket-brigade circuit). Samples of the signal are stored in the form of charges on capacitors, and are shifted by "switches" S from left to right. At any given time only half of the storage capacitors contain a signal sample, e.g. only those of even number (*a*), and the capacitors in between are discharged. While the information from the even-numbered capacitors is passed on by switches S_2 and S_4 to the odd-numbered capacitors, a new signal sample arrives at the input and a sample is delivered at the output (*b*). Switches S_1, S_3 and S_5 then pass on the information to the even-numbered capacitors, and so on.

switches are driven with a relative phase difference of half a period. The shifting and sampling are in practice effected by the same clock pulse.

The delay τ obtainable with such a shift register depends on the bandwidth B of the signal, for we know from the sampling theorem that, to characterize an analogue signal completely, at least $2B$ samples per second are needed [4]. To give a delay of τ_0 the register must therefore be capable of storing $2B\tau_0$ samples — in the case illustrated in fig. 1 this calls for $4B\tau_0$ capacitors — and the frequency of the clock pulses that effect the sampling and signal-shift must be $2B$. For a greater τ a longer register is necessary, but without changing the length of the register a smaller delay can be obtained by increasing the clock frequency. The delay can thus be continuously adjusted by varying the clock frequency, but this must not of course be lower than $2B$.

These delay lines have not however come into general use, because of the inevitable complexity and bulk of the "switches" S, which have to ensure a correct and complete transfer of the signal sample to the next capacitor, without losses and without being affected by non-uniformity of the different capacitors. Even with integrated-circuit techniques it is not possible to design these switching devices in a compact and economic form.

It will be shown in this article that we can get out of this impasse by transferring the signal in quite a different way, in which the charge is not shifted in the direction of signal travel but in the *opposite* direction. We start again from the situation in fig. 1*a* in which the even-numbered capacitors carry a sample (we shall confine ourselves here to positive values). The odd-numbered capacitors, which carry no signal, are now however no longer "empty", but are "full", that is to say they are charged up to a particular reference voltage, which is higher than the signal voltages. The transfer of information now takes place by transferring the charge from these "full" capacitors to the *left*, until the even capacitors in their turn are "full". It is clear that the charges that remain in the odd capacitors after this are equal to the charges that were originally stored in the preceding even capacitors, so that the signal has shifted to the right over a distance of one capacitor.

If we design a bucket-brigade delay line on this new principle, then all we need for the "switches" is one transistor per stage [5]. The whole circuit is now very suitable for integration; for example, a complete shift register of say 72 stages can be produced in the form of a monolithic circuit. Such a device can be used as an inexpensive and compact delay line for analogue signals, with the advantages of low distortion, a variable time delay and a high $B\tau$ product. This bucket-brigade circuit provides a considerable simplification not only as a delay line but also for various other forms of analogue signal processing.

Circuit of the bucket-brigade shift register

The operation of this new type of bucket-brigade shift register will first be discussed with reference to the basic diagram of *fig. 2*. It can be seen there that each stage of the register consists of an *NPN* transistor and a capacitor. These storage capacitors are numbered C_1, C_2 etc.; C_1 is the input capacitor. All the capacitors are equal and have the value C. The incoming signal samples, which appear successively in the form of voltages across C_1, are designated U_k ($k = 1, 2, \ldots$). By adding a predetermined d.c. voltage to the a.c. signal to be delayed, it is ensured that all voltages U_k can be made positive but lower than a fixed reference voltage designated $+U$.

We start from the situation in fig. 2*a*, where C_i contains a signal sample U_k, while the voltage $+U$ appears across C_1 an C_2. Since the bases of Tr_1 and Tr_2 are grounded, these transistors do not conduct and the voltages on the capacitors remain stationary. To transfer the information from C_i to C_1 the base potential of Tr_1 is now raised to $+U$, while the base of Tr_2 remains grounded (fig. 2*b*). Tr_1 now starts to conduct, since its base-emitter voltage has become positive because

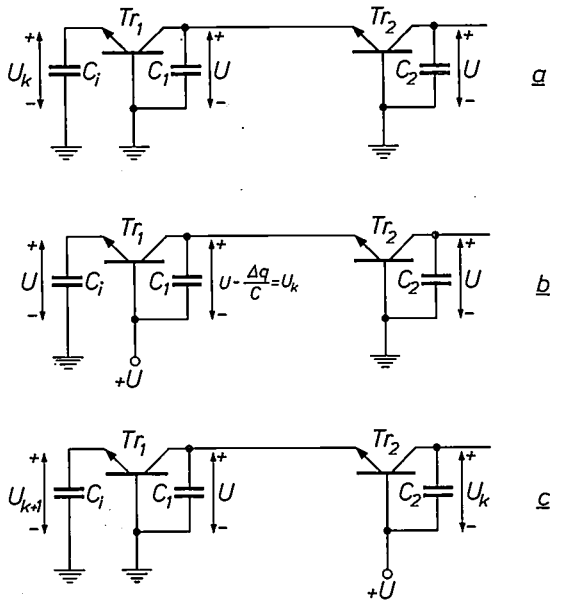


Fig. 2. The bucket-brigade shift register based on the new principle. Each stage or "bucket" consists of an *NPN* transistor and a capacitor. If a stage contains no signal sample, there is a voltage U across the storage capacitor. The signal samples, the voltages U_k ($k = 1, 2, \ldots$), are applied to the input capacitor C_i (*a*). To shift a sample to the next stage, the base of Tr_1 is brought to the potential $+U$. Tr_1 then passes current until a quantity of charge (Δq) has flowed from C_1 to C_i sufficient to bring the potential of C_i to $+U$ (*b*). If all the capacitances have the same value, the voltage on C_1 has thus dropped to $U - \Delta q/C = U_k$. In the same way U_k is shifted to C_2 by bringing the base of Tr_2 to the potential $+U$ (*c*). At the same time the next signal sample U_{k+1} is fed to C_1, thus completing one period of the shift process.

$U > U_k$. From C_1, whose upper terminal is now at the voltage $2U$, a positive charge can now flow to C_i until the voltage across C_i becomes equal to U (less a small residual voltage across the base-emitter junction of the transistor, which we shall neglect here). The base of Tr_1 is then returned to ground potential. Let Δq be the quantity of displaced charge, then: $\Delta q = C(U - U_k)$; the voltage now remaining across C_1 is $U - \Delta q/C$, which is equal to the signal sample U_k. (The effect of the base current, which is neglected here, will be dealt with later.)

In the second step the base of Tr_2 is brought to the potential U (fig. 2*c*), Tr_2 then starts to conduct and the charge Δq that had flowed away from C_1 is replaced from C_2. The capacitor C_1 is now at the potential U, while the voltage U_k appears across C_2. This completes one period of the shift process. Simultaneously with the second step another sample is taken, so that the next signal sample U_{k+1} appears across C_i.

When a storage capacitor is charged up, the voltage does not in practice quite reach the value $+U$. This is because there is always a transistor in series with the capacitor, and the non-linear resistance of the base-emitter junction of this transistor determines the charging process. As long as the voltage V_C on the capacitor is low, with the base-emitter voltage $V_{be} = U - V_C$ therefore high, the resistance of this junction is very low, so that the charging process begins with a short RC time constant. However, as V_C rises and V_{be} falls, the resistance now rapidly increases as a result of the exponential I_e-V_{be} characteristic. When V_{be} approaches the knee voltage that can be seen on the linear-scale characteristic (the threshold voltage of the transistor), the resistance becomes so high that there is effectively no further increase in V_C. If we now interrupt the charging process by returning the base of the transistor to ground potential, then the voltage on the capacitor is equal to $U - V_j$, in which the residual voltage V_j across the junction is somewhat higher than the threshold voltage. This means that when the next sample is taken over the reference voltage on the capacitor is not U, but $U - V_j$.

Although the charging current at the end of the charging process is small (1 μA to 1 nA) it is not quite equal to zero, and therefore the exact value of V_j (400 to 600 mV) depends on the time delay and hence on the shift frequency of the bucket-brigade circuit. Moreover, at high shift frequencies V_j is to some extent dependent on the voltage that appeared across the capacitor before charging up, and hence of the transferred sample. Because of this effect, a certain residual fraction of this sample remains behind to contribute to the reference level, giving an interfering effect between successive signal values which leads to attenuation at high frequencies. A similar effect arises because of the Early-effect interaction between the collector and emitter voltage in the transistor. We shall not discuss these effects further in this article.

In what follows we shall continue to assume, for simplicity, that all the storage capacitors have the same capacitance value, although this is not usually necessary in practice. For many applications, in fact, it is only the voltage on the last capacitor that is important, and not the intermediate values. Since the quantity of charge transferred is determined by the value of the input capacitance, and is independent of the capacitance values of the others, it is usually sufficient if just the output capacitor is equal to the input capacitor.

In the simplest version of the bucket-brigade shift register the transistor base voltages are supplied in the

[4] See C. E. Shannon, A mathematical theory of communication, Bell Syst. tech. J. 27, 379-423 and 623-656, 1948, and W. G. Tuller, Proc. I.R.E. 37, 468, 1949.

[5] F. L. J. Sangster and K. Teer, Bucket-brigade electronics — new possibilities for delay, time-axis conversion, and scanning, IEEE J. Solid-State Circuits SC-4, 131-136, 1969 (No. 3).

F. L. J. Sangster, Integrated MOS and bipolar analog delay lines using bucket-brigade capacitor storage, 1970 IEEE Int. Solid-State Circuits Conf., pp. 74, 75, 185.

form of two pulsed shift signals. *Fig. 3a* shows three stages of this circuit; fig. 3*b* gives the shift signals formed by the two complementary square-wave voltages Φ_1 and Φ_2. Also shown in this figure are the waveforms of the potentials V_1, V_2 and V_3 at the collectors of the transistors, in other words the voltages on the storage capacitors with respect to ground potential. In these graphs it can be seen that at the instant when Φ_2 goes positive the potential V_2 first rises rapidly to the value $+2U$, and then drops because of the flow of charge from C_2 to C_1. The final value of V_2 is then $U + U_k$, where U_k is the signal sample taken over from C_1. When Φ_2 goes to zero again, V_2 drops to U_k, but at the same time Φ_1 goes positive so that C_3 now transfers charge to C_2, causing V_2 to rise to the potential U again and V_3 in its turn to drop to the value $U + U_k$. Simultaneously C_1 takes over the next sample U_{k+1} from the stage preceding it, so that when Φ_1 goes to zero the sample U_k has been transferred from C_1 to C_3, and there is again a new value U_{k+1} present in C_1.

Compared with the circuits that have hitherto been available for analogue shift registers the circuit in fig. 3 is extremely simple. It is also easy to produce as an integrated circuit, and in fact when made as an integrated circuit such a shift register requires only one transistor per stage, since the capacitor can be obtained simply by giving the transistor a higher collector-base capacitance. As will be shown later, it is also easy to make the circuit with MOS transistors, and this has certain advantages over the bipolar-transistor version. In recent years several related circuits have been described [6], but these give more distortion than the circuit discussed here and are not so easy to integrate.

The sampling and output circuits are also driven by the shift signals Φ_1 and Φ_2. These circuits will be discussed later, but first we shall consider the signal distortion caused by the base current in the transistors.

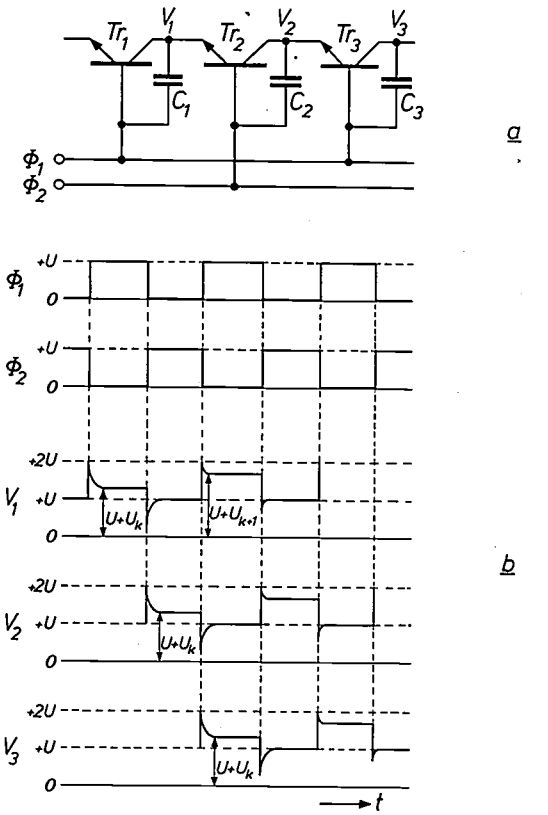


Fig. 3. *a*) Three stages of the simplest bucket-brigade circuit, using two shift signals. *b*) The shift signals Φ_1 and Φ_2, consisting of two complementary square-wave voltages, and the voltages V_1, V_2 and V_3 on the storage capacitors during the transfer of a signal sample U_k. When the base of a transistor goes to the voltage $+U$, the transistor starts to conduct and current flows from the associated capacitor to the preceding stage until this reaches the voltage $+U$. The potential on the stage itself rises momentarily to $2U$ and then drops, because of the loss of charge, to the value which the preceding stage had $(U + U_k)$, so that the information contained in this stage has now been transferred.

Effect of the base current

When the base of Tr_1 in fig. 2 is given the potential U, the charge that flows to capacitor C_1 is not derived from C_1 alone. Owing to the flow of base current in the transistor a small quantity of charge also flows from the base to C_1. Less charge is thus taken from C_1 than is supplied to C_1; because of this, Δq is attenuated slightly in each stage during the transfer of charge. This means that the signal sample $U_k = U - \Delta q/C$ approaches asymptotically to the value of U as it is transferred through the shift register. The a.c. component of U_k, which is of course the signal that is being delayed, becomes at the same time steadily smaller. We see this illustrated in *fig. 4*, which shows a signal before and after transfer through a number of stages. Eventually the signal can even disappear completely, if $U - U_k$ becomes smaller than the threshold voltage of the transistors; the transistors are then no longer brought into conduction and the voltage on the successive stages remains equal to U.

To give some idea of the magnitude of the signal attenuation we can express the attenuation of Δq per stage in terms of the current gain β of the transistors; the attenuation is then $i_e/i_c = 1 + 1/\beta$. After β stages this factor is $(1 + 1/\beta)^\beta$; this number of β stages has been chosen because $(1 + 1/\beta)^\beta$ for $\beta \to \infty$ has a known limit, which is e. Since in practice β will be at least 100, this limit is reasonably approximated here, and we can say that after β transfers Δq is attenuated by a factor e (approximately 9 dB).

In bucket-brigade shift registers with 10 to 20 stages it will not generally be necessary to compensate for this attenuation, but compensation is certainly necessary if there are large numbers of stages with the attendant risk of the signal eventually being lost. The compensation can be provided in a simple form by inserting at

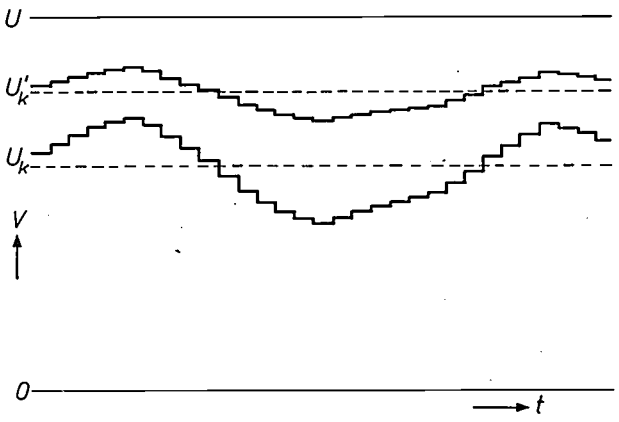


Fig. 4. Illustrating the effect of the base current in the transistors. The signal sample U_k fed to the bucket-brigade circuit consists of a d.c. voltage with the a.c. signal to be delayed superimposed on it. After passing through a large number of stages the signal has acquired the form U_k', since the quantity of charge transferred has become smaller because of the base current, giving a decrease of $U - U_k$. This causes an attenuation of the a.c. signal.

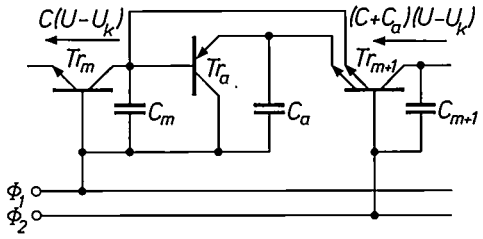


Fig. 5. Amplifier stage inserted at regular intervals in the line to compensate for the charge-transfer loss caused by the base current in the bucket stages. If Φ_1 is positive and the sample U_k appears in stage m, a quantity of charge $C(U - U_k)$ is transferred to stage $(m - 1)$. Since Tr_a is connected as an emitter follower with Φ_1 as supply voltage, C_a receives the same voltage as C_m. If Φ_2 now goes positive, both C_m and C_a are charged via the two emitters of Tr_{m+1} with charge from C_{m+1}, resulting in a charge transfer of $(C + C_a)(U - U_k)$. The displaced charge is thus multiplied by the factor $(C + C_a)/C$.

regular intervals an amplifier stage to restore the quantity of transferred charge to the correct value. *Fig. 5* shows a circuit of this type, inserted between the stages m and $m + 1$. When Φ_1 in this circuit goes positive, a signal sample of say U_k is applied to C_m. The voltage across C_m then drops from U to U_k, causing a charge $C(U - U_k)$ to flow from C_m to the preceding stage $(m - 1)$. The voltage across C_m is applied to the base of a *PNP* transistor Tr_a, which is connected as an emitter follower with Φ_1 as supply voltage. Consequently the same voltage U_k appears across capacitor C_a as across C_m (we again neglect the residual voltage across the base-emitter junction). Capacitor C_a is also connected to a second emitter of the transistor of stage $(m + 1)$. If now Φ_1 goes to zero and Φ_2 goes positive, this transistor starts to conduct and C_{m+1} discharges to take over the value U_k. Charge then flows, however, not only to C_m but also to C_a, and continues until these two capacitors are again at the voltage U [7]. As a result C_{m+1} loses a quantity of charge $(C + C_a)(U - U_k)$, so that Δq is multiplied by a factor $(C + C_a)/C$. If the amplifier stage is inserted after β stages, then $(C + C_a)/C$ must be equal to e, and therefore C_a must be equal to $(e - 1)C$.

The sampling circuit

The signal samples have to be supplied to the first stage of the bucket-brigade shift register. This is done by the circuit shown in *fig. 6a*. The signal is fed in at the left of the circuit and added to a positive bias voltage whose value is such that the voltage at point P is always between 0 and U volts. This voltage will from now on be taken as the signal. The combined action of the diodes D_1 and D_2 in conjunction with a kind of "buffer stage", consisting of the transistor Tr_b, the capacitor C_b and the diode D_b, puts samples U_k of this signal on to capacitor C_i, where they arrive at the instant when the first stage of the register can take them over in the normal way, i.e. at the instant when Φ_1 goes positive. Fig. 6*b* shows the shift signals Φ_1 and Φ_2, and the square-wave voltage Φ_2' used for driving the buffer stage; this voltage is obtained by subtracting the d.c. voltage U from Φ_2. The figure also shows the waveforms of the input voltage V_i on the capacitor C_i and the voltage V_1 on the first stage.

The sampling circuit works as follows. In the first half period Φ_1 is positive, Φ_2 is at zero level and Φ_2' is negative. Capacitor C_i is charged by transfer from C_1 to the voltage $+U$. The base of Tr_b is negative, so that this transistor does not conduct; diodes D_1 and D_2 now prevent current from flowing between C_i and the signal source, so that V_i remains constant. Since Φ_2' is negative, the buffer capacitor C_b discharges via D_b to the voltage $-U$. In the second half of the period, Φ_2' is at zero level. As the negative charge on C_b cannot leak away, the base-emitter voltage of Tr_b is now positive, so that Tr_b can conduct. Current now flows via D_1, D_2 and Tr_b until the voltage across C_b has again risen to near the zero level; there is then a small residual voltage across the base-emitter junction of Tr_b and the current has dropped to a negligible value.

The charging current for the buffer capacitor is derived in the first instance from C_i, since the voltage U across this capacitor is higher than the signal voltage.

---

[6] G. Krause, Analog-Speicherkette: eine neuartige Schaltung zum Speichern und Verzögern von Signalen, Electronics Letters **3**, 544-546, 1967.
R. A. Mao, K. R. Keller and R. W. Ahrons, Integrated MOS analog delay line, 1969 IEEE Int. Solid-State Circuits Conf., pp. 164-165.
W. S. Boyle and G. E. Smith, Charge coupled semiconductor devices, Bell Syst. tech. J. **49**, 587-593, 1970 (No. 4).

[7] The same result could be obtained by connecting C_a via a diode to the emitter output of Tr_{m+1}. In an integrated circuit it is easier, however, to make a second emitter contact in a transistor than to make a separate diode. Such "multi-emitter transistors" are therefore widely used in integrated circuits.

As a result of this loss of charge, V_i quickly drops to the value of the signal voltage at that moment, and further charging current is supplied by the signal source. V_i can now drop no further, since C_i is connected via the now conducting diodes to point P, and thus remains at the same potential. When C_b is charged up, so that the current no longer flows C_i is again separated from the signal source by D_1 and D_2, and V_i therefore remains identical with the sampled signal voltage. We have denoted this value by U_k ($k = 1, 2, 3, \ldots$).

Because there are no resistors in the circuit, the sampling takes place so rapidly (in a few nanoseconds) that it virtually coincides with the voltage jump of Φ_2'. This means that the circuit can be used up to very high shift-signal frequencies (e.g. 30 MHz). In fig. 6*b* it can be seen that the signal sample U_k is available at the right moment on C_i; if Φ_1 again goes positive at the beginning of the next period, the first stage takes over the value by charging C_i to the voltage $+U$.

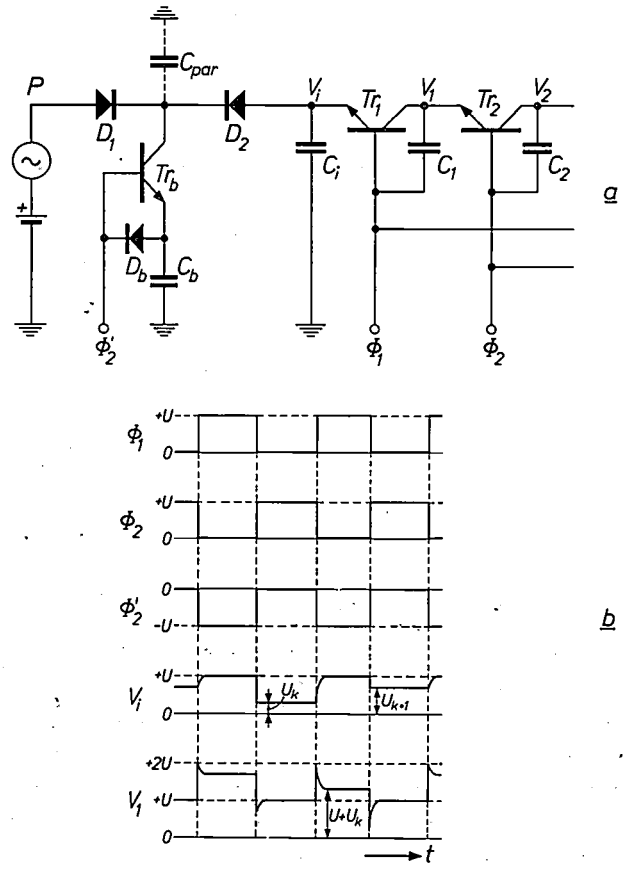


Fig. 6. *a*) Circuit which samples the signal and feeds the sampled values to the bucket-brigade circuit. *b*) The square-wave shift signals Φ_1, Φ_2 and Φ_2' which drive circuit (*a*), together with the waveform of the voltage V_i on capacitor C_i and the voltage V_1 on the first bucket stage. Whenever the signal Φ_2' goes from negative to zero level the transistor Tr_b conducts momentarily, so that V_i becomes equal to the value that the signal voltage V_P has at that moment. This value U_k remains stored in C_i until Φ_1 goes positive and the first bucket stage takes over the value.

The proper operation of the circuit depends on C_b being large enough to take up the maximum quantity of charge, not only from C_i but also from the stray capacitance C_{par} (see fig. 6*a*), which is present when the shift register is in the form of an integrated circuit. The maximum displaced charge $U(C_i + C_{par})$ occurs when the signal voltage is zero; C_b can then be charged from $-U$ to zero, so that C_b must be greater than $C_i + C_{par}$.

The circuit of fig. 6*a* is in fact half of the four-diode bridge circuit generally used for sampling. The resistor or pulse transformer in series with the source of the sampling pulses is replaced here by the buffer stage. This solution was chosen because there is then no need for the high-voltage pulses which are necessary for the bridge circuit.

The output circuit

If the voltage of the last stage of the bucket-brigade shift register is to be a useful output signal, the circuit must be capable of delivering a certain amount of power at the output. The storage capacitors are very

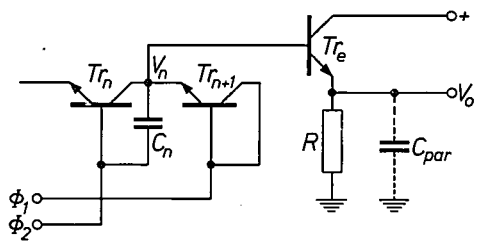


Fig. 7. Output circuit in which an emitter follower Tr_e is in series with the last stage. The output voltage V_o of this circuit is equal to the voltage V_n of the last stage (less a residual voltage V_j of 400 to 600 mV across the base-emitter junction). The emitter follower gives only a very small load on C_n, and a much higher current can be taken at the output. However, since the parasitic capacitance C_{par}, because of the decrease in V_o, has to discharge through R, this circuit can only be used in a limited frequency range. At high frequencies R must be small, otherwise C_{par} cannot discharge quickly enough, whereas at low frequencies R must be large to prevent C_n from discharging through R.

small, however — no more than a few pF in an integrated version — and current cannot therefore be drawn directly from the last capacitor. The amount of current that can be delivered can of course be increased by connecting an emitter follower in series with the last stage, as shown in *fig. 7*. Provided the resistor R is made sufficiently high, the output voltage V_o is equal to the voltage V_n of the last stage (less, of course, the residual voltage V_j across the junction). Owing to the high input impedance of the emitter follower, the load on C_n is minimal, while a much higher power can be delivered at the output of the emitter follower. (The transistor Tr_{n+1} is needed for repeatedly recharging C_n to the voltage U.)

Owing to the presence of a parasitic capacitance C_{par}

the circuit in fig. 7 cannot be used over the whole frequency range of interest for the bucket-brigade shift register. This is because the parasitic capacitance has to discharge via R when V_o drops, and the RC constant of this discharge may be so great that V_o cannot follow the voltage V_n, causing distortion. To avoid this effect at high shift frequencies, R must be given a low value. At low frequencies, on the other hand, R must have a high value, otherwise distortion results from too much current being drawn from C_n.

A circuit that can be used over a wide frequency range can be obtained by replacing the resistor of the emitter follower by two buffer stages of the type used in the sampling circuit. *Fig. 8a* shows the output circuit obtained in this way, which is used in the bucket-brigade shift register; fig. 8*b* gives the required voltages with the waveforms of V_n and V_o. We have already described the operation of the buffer stages in connection with the sampling circuit; the transistor of such a stage starts to conduct when the driving voltage (here Φ_1' and Φ_2') goes from negative to zero level, and it stops conducting (or the current becomes negligibly small) when the buffer capacitor is charged up. In the circuit of fig. 8*a* this means that at the beginning of each half-period of the shift voltages, i.e. at the moment when V_n changes value, one of the transistors Tr_{b1} and Tr_{b2} starts to conduct. This enables C_{par} to discharge, and consequently Tr_e starts to conduct and thus acts as an emitter follower, so that the output voltage V_o follows the voltage V_n. The values of the buffer capacitors are so chosen that the time required to charge them up is at least as great as the time required to charge the storage capacitor C_n, so that the current through Tr_e is not interrupted until V_n, and hence V_o, reach the new value. This output voltage is maintained until the end of this half-period, during which time no further current is drawn from C_n. The length of the period is of no importance here; this circuit will therefore also deliver the required output voltage at low shift frequencies without too heavy a load on C_n. The discharge of C_{par} and the re-establishment of V_o at a new value take place so quickly that the maximum shift frequency of the bucket-brigade shift register is not in practice determined by this output circuit. As a rule the upper limit of the frequency range is determined by the maximum frequency of the shift pulse source (e.g. 30 MHz).

We thus have a circuit in which the effect of C_{par} is neutralized over a wide frequency range without too heavy a load on C_n. Owing to the presence of buffer stages, however, the output impedance is still fairly high, and to reduce this to the required low value an ordinary emitter follower is connected in series with the circuit given in fig. 8*a*.

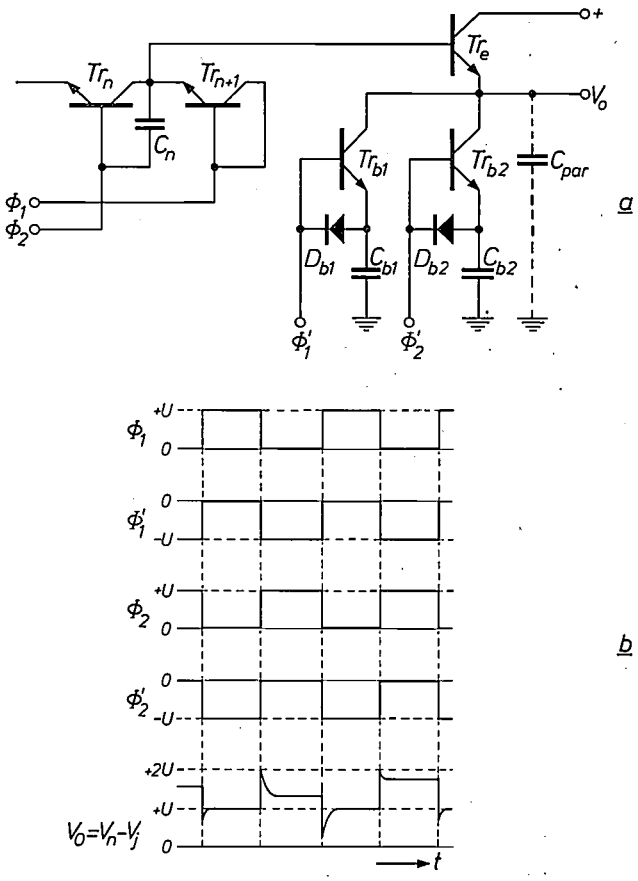


Fig. 8. *a*) Output circuit suitable for a wide range of frequencies. *b*) The driving voltages for (*a*) and the waveform of the output voltage V_o. In this circuit the resistor R of the emitter follower in fig. 7 is replaced by two buffer stages of the type used in fig. 6*a*. These are driven by the signals Φ_1' and Φ_2'. At the beginning of each half-period of the shift signals, at the moment when V_n acquires a new value, the transistor of these buffer stages starts to conduct. This enables C_{par} to discharge rapidly, while at the same time Tr_e starts to conduct; since this transistor acts as an emitter follower, the output voltage V_o follows the voltage V_n (where we again have $V_o = V_n - V_j$, of course). If the buffer capacitors are given appropriate values, the current is interrupted when V_o and V_n have reached the new constant value. During the first half-period no further current is drawn from C_n; this means that the requirement of a minimum load on C_n is also satisfied at low frequencies. The discharge of C_{par} and the re-establishment of a new value of V_o take place so quickly that the circuit can also be used for very high frequencies.

The information packing density

In the simplest form of bucket-brigade shift register, the type with two shift voltages as illustrated in fig. 3*a*, only half of the storage capacitors contain a signal sample at any given moment. The "packing density" is therefore no greater than $\frac{1}{2}$. A greater packing density can be achieved by modifying the circuit slightly and using a more complex pattern of shift pulses. *Fig. 9a* shows part of a bucket-brigade shift register that uses three shift signals, Φ_1, Φ_2 and Φ_3 (fig. 9*b*). This circuit operates in the same way as that in fig. 3, except that now three "shift pulses" are applied in each sampling cycle to groups of three stages. The voltages V_1, V_2 and V_3 on the stages are thus somewhat different in waveform from the corresponding voltages in fig. 3*b*.

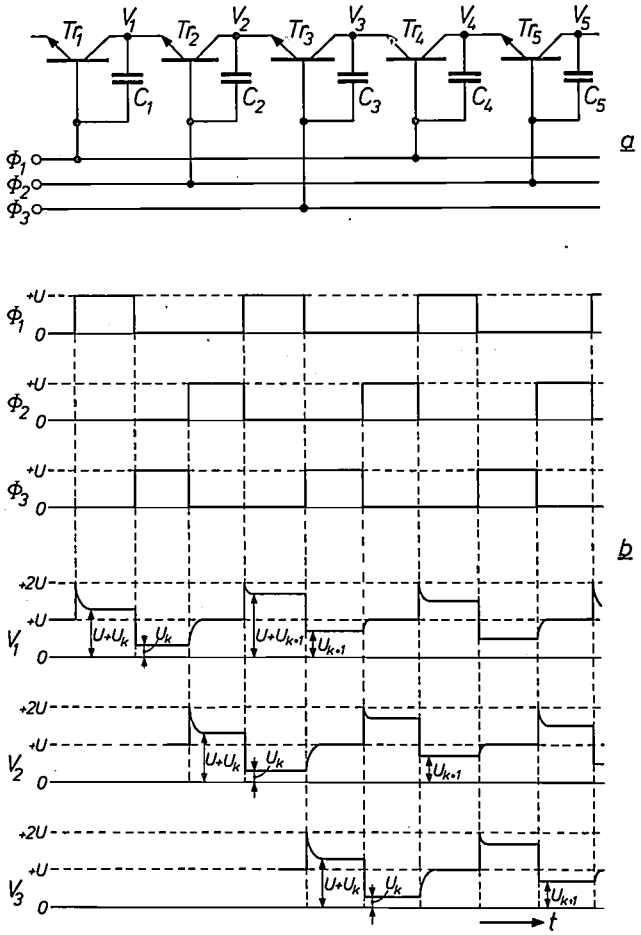


Fig. 9. *a*) Circuit of a bucket-brigade shift register with higher information packing density. *b*) The shift signals Φ_1, Φ_2 and Φ_3, and the voltages V_1, V_2 and V_3 of the first three stages. The curves show that in this circuit two out of every three stages contain a signal sample. The packing density is thus $\frac{2}{3}$, compared with $\frac{1}{2}$ for the circuit in fig. 3.

When Φ_1 becomes positive, the first stage acquires the signal sample, so that V_1 becomes equal to $U + U_k$. When Φ_1 returns to zero, C_1 is not immediately recharged to the voltage U, as in fig. 3, but V_1 goes first to the value U_k, and rises to the value U only when Φ_2 becomes positive and the second stage takes over the sample. The first stage thus contains a sample during two-thirds of the shift-pulse period. If we consider the three stages together, then at any given time two of the three always contain a sample, so that the packing density is $\frac{2}{3}$.

On this principle it is also possible to work with more than three shift voltages. With p voltages the information packing density is then $(p-1)/p$. We have already seen that to give a signal of bandwidth B a delay of τ_0 it is necessary to store $2B\tau_0$ values. If the packing density of the register is $(p-1)/p$, this means that $2B\tau_0 \cdot p/(p-1)$ stages are needed. In the circuit given in fig. 3, where $p = 2$, the number of stages needed is thus $4B\tau_0$; circuits with greater values of p require fewer stages.

There are also certain disadvantages in the use of a large number of shift signals. Since the time between two sampling operations remains the same (it is equal to $1/2B$), and since p shift pulses must appear during this time, a larger value of p implies shorter pulses and also proportionately less time for charge transfer between the capacitors, which must of course be completed within the period of one pulse. If the bandwidth of the signal is large, this may set a limit to p. Another disadvantage is that an additional shift register with p parallel outputs is required for producing the shift signals, whereas all that was needed for this in the basic circuit was a simple bistable circuit. Thirdly, the circuit requires more crossovers (see fig. 9*a*), which involves technological difficulties if a monolithic circuit is required. The choice of p will therefore depend on the nature of the application and on the technological possibilities.

The first disadvantage of a large value of p, that the time available for charge transfer is short, can be overcome by arranging the stages in p parallel rows instead of in one long row. The input stages are then all connected to the sampling circuit and the shift pulses are fed to the rows of stages in such a way that the successive samples are distributed among the parallel rows. Each row need therefore store only one of p samples; this means that the shift pulses can be p times longer, so that the time available for charge transfer is p times greater. An added advantage is that each sample need only pass $1/p$ times the original number of stages, giving a considerable reduction of distortion. Set against this is the disadvantage that any dissimilarity between the different rows will appear after combination of the samples in the form of a spurious signal.

The bucket-brigade shift register as an integrated circuit

The various types of bucket-brigade shift registers described in this article lend themselves very well to fabrication by integration techniques. We have already seen that only one transistor is needed for each stage of a bucket-brigade circuit; *fig. 10a* shows a cross-section of such a transistor [8]. The storage capacitor, located between base and collector, is formed by the capacitance of the junction between the *P*-type base and the enlarged N^+ collector contact layer, together with the "Miller capacitance" C_{cb} between collector and base. The total value of the capacitance is approximately 2.5 pF. Fig. 10*b* shows a plan view. The enlarged N^+ layer of the collector is clearly visible; it extends over a much wider area than can be seen in fig. 10*a*. *Fig. 11* shows a photograph of the first experimental circuit, where 16 of these stages are laid out on a silicon chip in an array of four rows of four stages (i.e. for operation with four shift signals).

Of the other parasitic capacitances the collector-

[8] The design for the integrated circuit was made in cooperation with Ir. C. Mulder of this laboratory.

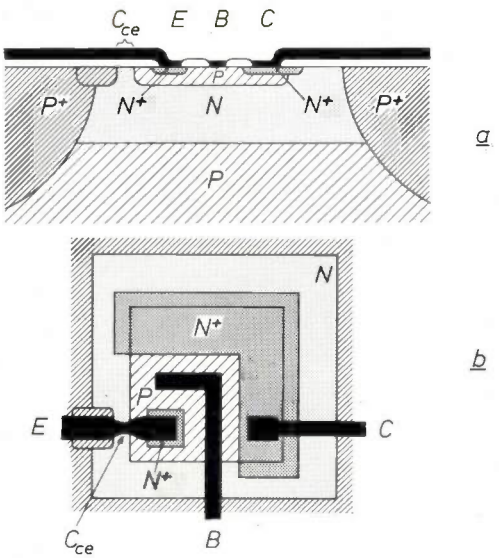


Fig. 10. Cross-section (*a*) and plan view (*b*) of a stage in a monolithic bucket-brigade shift register. In a *P*-type silicon substrate an island of epitaxial *N*-type silicon is formed by P^+ diffusions. In this *N*-type region an *NPN* transistor is formed, the N^+ diffusion of the collector contact being made great enough to cover a large part of the base diffusion. The junction between these diffused areas forms, together with the parasitic capacitance between collector and base, the storage capacitor of the stage. The aluminium connectors of base, emitter and collector are shown black in the figures, and the oxide white. The extra region of *P*-type material under the emitter connection serves to minimize the parasitic capacitance C_{ce}.

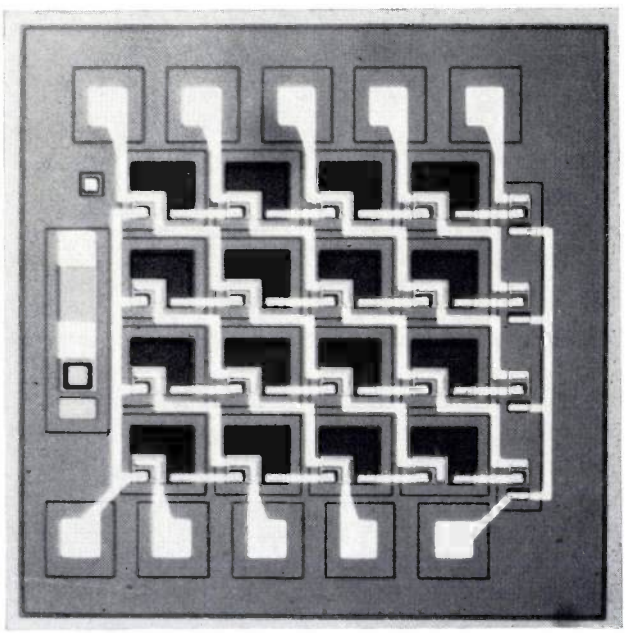


Fig. 11. First experimental version of an integrated bucket-brigade array with 16 stages, arranged in four rows of four stages. The chip size is 1.25×1.25 mm. The configuration of fig. 10*b* can clearly be recognized in this photograph. Successive transistors are directly interconnected by the aluminium strips of the collector and emitter (the aluminium looks white on the photograph). The base connections follow a particular pattern for the supply of the four shift signals.

substrate capacitance and the emitter-base capacitance act as voltage dividers for the shift signal and as extra parts of the storage capacitors, so that these capacitances present no difficulties at all. This is not the case for the collector-emitter capacitance C_{ce}; this bypasses the transistor and thus forms a direct coupling between successive signal samples, resulting in distortion of the signal steps. The collector-emitter capacitance arises because the aluminium strip which connects two transistors, and is connected with the emitter of one of these transistors, crosses the *N*-type collector area of this transistor (see C_{ce} in fig. 10). By making the aluminium connection and the underlying strip of *N*-type material as narrow as possible, and by constricting the *N*-type material at this point with an extra zone of *P*-type material, the area of the cross-over can be limited to 2.5×5 μm, giving a capacitance of less than 0.001 pF. The distortion caused by this capacitance then only becomes noticeable after a few hundred stages.

For the practical application of the bucket-brigade shift register it is of course desirable to accommodate a large number of stages together with all their ancillary circuits on a single chip. *Fig. 12* shows an integrated circuit of this type, consisting of 72 bucket stages (coupled in the manner illustrated in fig. 3), an amplifier stage, a sampling circuit and an output circuit. These integrated circuits can be connected in series to form an even longer delay line. A delay line has been built with four of these integrated circuits giving a total of 288 bucket stages. Such a delay line can give a signal of bandwidth B a delay τ_0 for which $B\tau_0 = 72$. The maximum delay time follows from the maximum shift frequency, which is about 30 MHz. The maximum sampling frequency is therefore 30 MHz as well, which corresponds to a maximum bandwidth of 15 MHz. The shortest delay is thus approximately 5 μs. The maximum delay time is determined by the leakage currents in the transistors, and this means that the shift frequency may not in general go below a few tens of kHz. The minimum bandwidth is in the region of 15 kHz and the maximum delay is 5 ms. Between 5 μs and 5 ms the delay can be continuously varied by varying the shift frequency.

To illustrate the low distortion of this delay line, *fig. 13* shows a television picture, the left-hand strip showing the original picture, and the following strips the picture after the passage of the signal through an increasing number of delay-line stages. In the second strip the signal has passed through a delay line with *four* integrated circuits (where $\tau \approx 14$ μs and $B \approx 5$ MHz), and in the third and fourth it has passed through *eight* and *twelve* circuits of 72 stages each. It can be seen that the picture has lost hardly any of its sharpness even after hundreds of stages.

The bucket-brigade shift register with MOS transistors

The first experiments with the bucket-brigade shift register were carried out using bipolar transistors. In some respects, however, the MOS transistor [9] seems to be better suited for this circuit than the bipolar transistor. Owing to the absence of d.c. gate currents in MOS transistors there is no loss during charge transfer, and therefore no amplifiers are necessary. In MOS bucket-brigade configurations the input and output circuits are also much simpler.

circuit a source follower (the equivalent of the emitter follower) is sufficient; here again, the absence of a gate current means that no current is drawn from C_n. The resistor R need not therefore be high, so that no trouble is experienced from any parasitic capacitance that may be present.

In integrated form the MOS bucket-brigade circuit can be a very simple configuration, since a source connection is only required at the input of a row of stages and a drain connection at the output, while only the

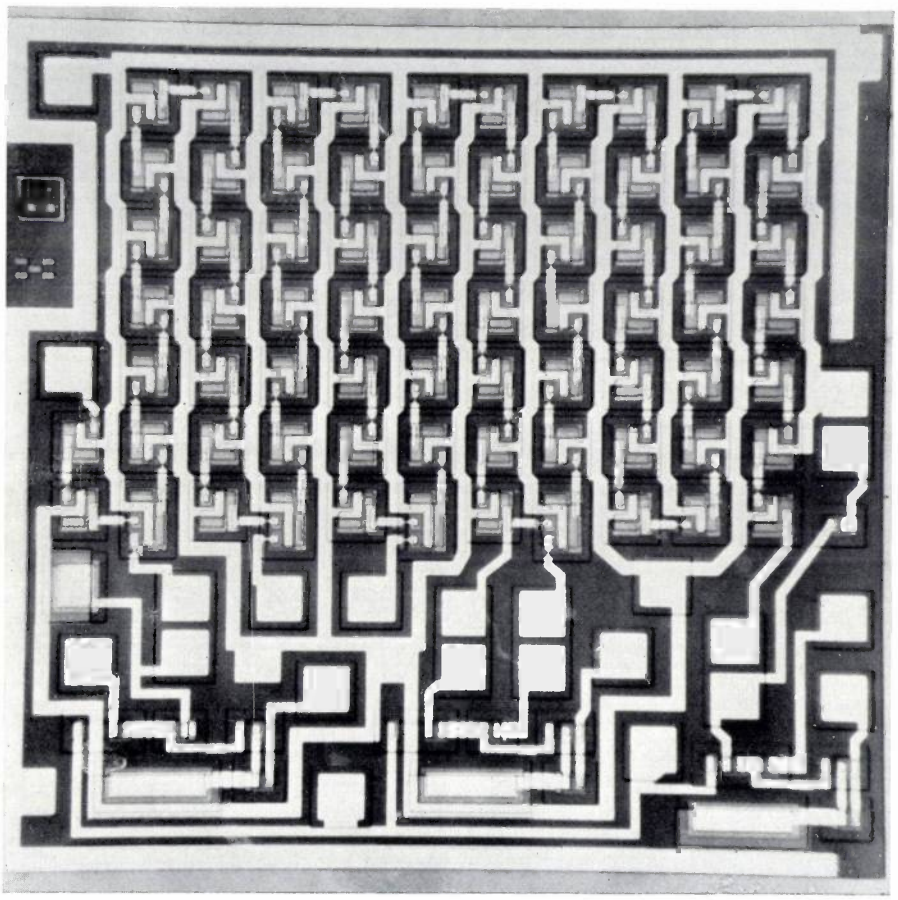


Fig. 12. Integrated bucket-brigade delay line with 72 stages, including sampling, amplifying and output stages. Chip size 2×2 mm.

Fig. 14 shows a diagram of a complete MOS bucket-brigade shift register. It is built up with *P*-channel MOS transistors, which means that the shift signals Φ_1 and Φ_2 must be negative. The transfer of information between the stages takes place in the same way as in the bipolar circuit. Sampling, however, takes place at a different point in time. In this case, when Φ_2 becomes negative, Tr_i starts to conduct, so that the voltage across C_i becomes equal to the signal and remains equal to it until Φ_2 has again gone to zero. The signal sample from that last instant thus remains stored in C_i, whereas in the bipolar case the sampling takes place at the instant when Φ_2 becomes positive. In the output gates of the intermediate stages have to be accessible. *Fig. 15* gives a cross-section of a circuit of this type [10]. In an *N*-type substrate *P*-type regions are diffused each of which acts as the drain of one MOS transistor and as the source of the next one. The storage capacitors are formed by situating the aluminium gate connections asymmetrically with respect to these *P*-regions, so that there is always a capacitance present between

[9] H. C. de Graaff and H. Koelmans, The thin-film transistor, Philips tech. Rev. **27**, 200-206, 1966. A forthcoming special issue of this journal will be entirely devoted to the MOS transistor.

[10] The design for the integrated circuit was made in cooperation with Ir. H. Heyns of this laboratory.

Fig. 13. Photograph of television picture illustrating the low distortion caused by the passage of the same signal through a bucket-brigade delay line. The original picture appears on the left; in the second strip it has been delayed by four integrated circuits of the type shown in fig. 12 (a total of 288 stages). In the third strip the signal has passed through eight of these circuits, and in the fourth strip twelve. It can be seen that there is very little degradation of the picture.

gate and drain. The part between the two dashed lines in fig. 15 thus comprises one bucket stage, with the MOST on the left and the capacitor on the right.

Fig. 16 is a photograph of a chip containing two integrated MOS bucket-brigade shift registers. The input and output circuits are not integrated here; each row contains 36 stages. Each row has contact points S and D for the input and output signals and points Φ_1 and Φ_2 where the shift signals are applied. The capacitance per stage is 8 pF in this version.

Compared with the bipolar bucket-brigade circuit the MOS version is much simpler and more elegant. MOS transistors do not switch as fast as bipolar transistors, however, and the maximum shift frequency of the MOS circuit is about 10 times lower; the frequency range of the circuit given in fig. 16 is between 100 Hz and 3 MHz. Moreover the MOS circuits need higher shift voltages, because an MOS transistor only starts to conduct when the gate voltage is higher than a threshold voltage of say 3 volts. Because of these higher voltages the dissipation of the MOS circuit is 10 to

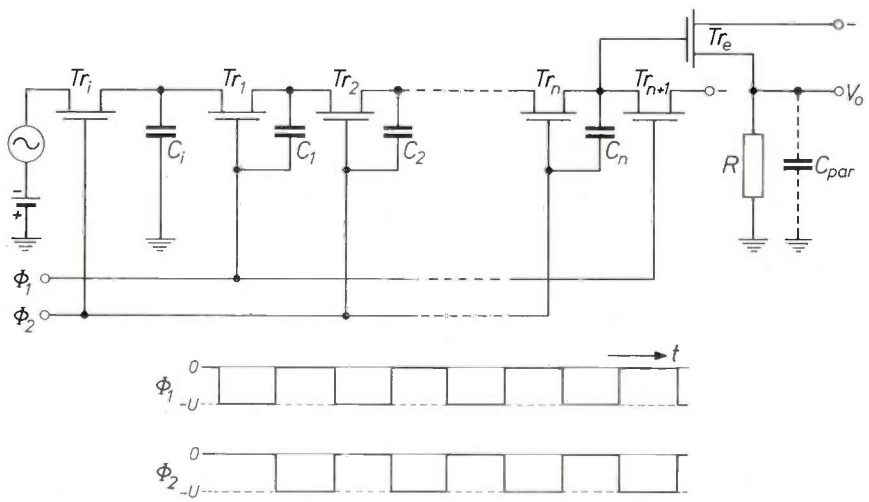


Fig. 14. Bucket-brigade circuit using P-channel MOS transistors. The absence of gate current in MOS transistors permits a much simpler input and output configuration than in circuits with bipolar transistors.

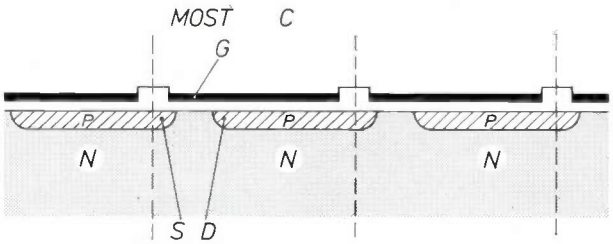


Fig. 15. Cross-section of an integrated bucket-brigade circuit using P-channel MOS transistors. The P-type regions diffused in an N-type substrate act at the same time as the drain of one transistor and the source of the next. The dashed lines comprise one bucket stage; the MOS transistor with source S, drain D and gate G is shown on the left; the capacitor C between the aluminium layer of the gate and the P^+ region of the drain is shown on the right.

20 times higher than that of the bipolar one. An advantage, on the other hand, is that higher signal voltages can be handled. Generally speaking, the MOS circuit is well suited to audio applications and the bipolar one more to video applications.

Some applications of the bucket-brigade shift register

The need for simple electronic delay lines to handle analogue signals in the audio and video frequency ranges has already been discussed in the introduction. We have explained there how the bucket-brigade shift register can meet this need over a very wide frequency range. In what follows we shall give some special examples of

electronic signal processing where the use of the bucket-brigade shift register results in particularly elegant methods and circuits.

The great virtue of the bucket-brigade circuit as a delay line is undoubtedly the facility of being able to continuously vary the delay time within a wide range by varying the shift frequency. This makes the bucket-brigade circuit particularly useful in systems where *timing correction* is required. A special case in point is the reproduction of audio and video signals recorded on magnetic tape, where timing errors are caused by tape stretch and by fluctuations in the tape speed [11]. These time errors can be corrected by passing the signal through a bucket-brigade delay line, using the errors to control the shift frequency. The time errors must then be measured in relation to a reference signal on the tape (for example a clock signal or a line synchronization signal).

Variation of the shift frequency can also be used for deliberately distorting the time axis of the signal, e.g. for expanding or contracting it. The television picture in *fig. 17* illustrates this facility. The left-hand part of the photograph shows the original picture. The signal for each line of this picture was fed into a bucket-brigade delay line at a shift frequency of 9 MHz. Immediately afterwards the signals were read out at the lower frequency of 3 MHz and fed back to the monitor, giving the picture displayed on the right of the photograph. The time axis has thus been expanded here by a factor of 3; this factor can be varied by means of the read-out frequency.

A practical application of this principle is to be found in telephony, where it is often desirable to transmit a number of narrow-band signals over one broad-band channel. For this purpose each signal can be fed into a separate bucket-brigade delay line at the low frequencies corresponding to their small bandwidth. The different signals are then collected by reading out the bucket-brigade delay lines in a fixed sequence at the high frequency corresponding to the bandwidth of the

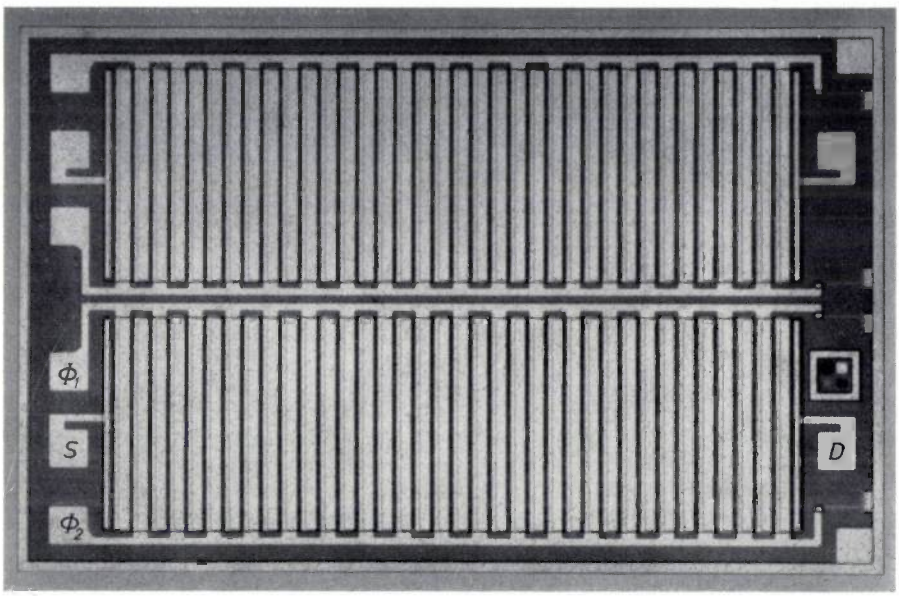


Fig. 16. MOS integrated bucket-brigade circuit with two rows of 36 stages. (The input and output circuits are not integrated.) Chip size 1.5×2.4 mm. The metallized (aluminium) areas on the photograph are light in colour, the regions where the oxide is visible are dark. In the lower row the connections are indicated for the source S of the first stage, the drain D of the last stage and the points where the shift signals Φ_1 and Φ_2 are applied.

Fig. 17. Television picture illustrating time-axis conversion. The original picture on the left was written line by line into a bucket-brigade shift register and read out at a lower frequency, giving the picture on the right.

channel. This results in a signal with the desired bandwidth, in which compressed portions of the signals follow each other at fixed intervals. At the receiving end the signals are separated again by repeating the process in the reverse order. This is known as a time-compression multiplex system.

The bucket-brigade circuit can also be used with advantage for making *transversal filters* [12]. The signal here is passed through a number of delay elements; after a delay τ it is multiplied by a certain weighting factor, and the products are summed to form the output signal. The filter can be given any desired frequency and phase characteristic by choosing appropriate values for τ and the weighting factors. As the signal in a bucket-brigade delay line is available after every stage, the circuit can very usefully serve as a basis for a filter of this type. The signal must of course pass through the line of delay elements unattenuated, and therefore only the MOS circuit enters into consideration for this purpose.

The weighting factors can be introduced into the bucket-brigade delay line in a very simple manner, since a measure of the signal sample transferred from one capacitor to the next is readily available. This is the quantity of charge that flows from the energized shift-signal source via the two capacitors and the MOS transistor to the other shift-signal source. The total current drawn from the shift-signal source during the transport is thus a measure of the sum of the displaced samples. The weighting factors are now introduced by dividing one of every two storage capacitors into two parts. One of these sub-capacitors is connected to the shift-signal source in the normal way; the other, which determines the weighting factor, is connected to the output of the filter. A photograph of such a filter can be seen in *fig. 18*.

An example of an application in which the bucket-brigade shift register is not used as a delay line is the scanning of an array of image-sensing devices. This can be done by using an integrated bucket-brigade circuit in which the charge pattern is applied not by electrical means but by illuminating photo-sensitive parts of the bucket stages [13]. In an MOS bucket-brigade circuit these are the parts of the *P-N* junctions not covered by aluminium, between the source and substrate. An array of $n \times n$ light-sensitive elements can then be made by forming an MOS bucket-brigade circuit on a silicon chip, with the buckets arranged in n rows of n buckets. Proceeding from an initial state in which all the capacitors are at a potential $+U$, illumination of the array causes a "charge pattern" to be formed as a result of

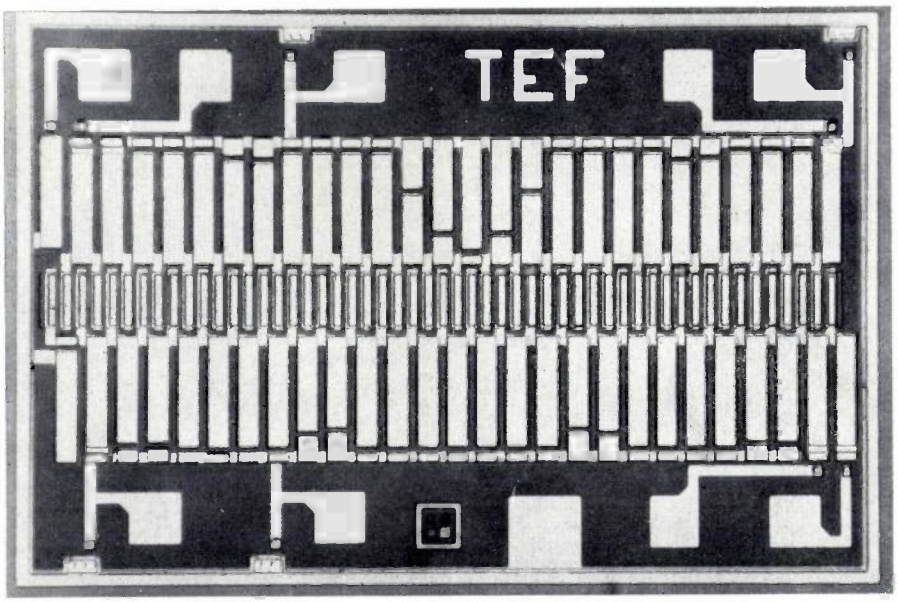


Fig. 18. Transversal low-pass filter consisting of an MOS bucket-brigade circuit with 52 stages and 23 taps. The MOS transistors can be seen in a horizontal row in the middle of the picture. Above and below are the even and odd storage capacitors; these are divided by oxide strips (dark grey) into two parts, the outside parts being connected to the output. Chip size 1.2×1.8 mm.

[11] W. J. Hannan, J. F. Schanne and D. J. Woywood, Automatic correction of timing errors in magnetic tape recorders, IEEE Trans. **MIL-9**, 246-254, 1965.

[12] H. E. Kallmann, Transversal filters, Proc. I.R.E. **28**, 302-310, 1940; see also p. 74 etc. in P. J. van Gerwen, The use of digital circuits in data transmission, Philips tech. Rev. **30**, 71-81, 1969 (No. 3).

[13] This application was proposed by Dr. Ir. K. Teer of this laboratory.

the varying extents to which the capacitors discharge during illumination. The information can be read out by rapid sequential scanning with shift pulses; the signal then obtained is rather like the signal obtained from the electron-beam scan in a television camera tube. Since no signal loss can be permitted in this application, even with correcting amplification at certain stages, only the MOS circuit is suitable here.

For the time being the number of image cells is limited. The minimum size of an image cell obtainable with present MOS technologies is about 45×60 μm; a matrix with the same resolution as a television camera tube, i.e. with 625×625 image cells, would occupy an area of 30×40 mm, which is not feasible at the present time. It is possible, however, to produce arrays of 50×50 image cells, which can be used for applications such as automatic character-recognition systems.

Although the bucket-brigade shift register was designed for analogue signals, this does not of course mean that it cannot be used to process digital signals. The bucket-brigade circuit can be used to good advantage in all cases where a binary shift register is employed. Since the bistable circuits used for making a binary shift register are fairly complex, the use of the bucket-brigade circuit with digital signals may well give a higher information packing density.

Summary. Since an analogue signal with a bandwidth B is completely determined by $2B$ samples per second, it can be stored in a shift register in which each cell or stage can contain an analogue signal sample. The bucket-brigade circuit is a register of this type; the stages consist of a storage capacitor whose voltage represents the signal sample, and a transistor which effects the transfer of the information. This is done by transferring the charge of the capacitors not in the direction of signal travel but in the opposite direction, the charge on the next capacitor in the line being transferred to the one preceding it until the voltage on that capacitor has reached a fixed reference level. The sampled signal value is then shifted to the next stage. In the simplest circuit two shift signals are needed, and every other stage contains a signal sample. The circuit may be used as a delay line, in which case, to give a signal of bandwidth B a delay of τ seconds, the number of stages required is $4B\tau$. An important advantage of the bucket-brigade delay line is that the same circuit can be used for large B and small τ (e.g. in the video band), and for small B and large τ (in the audio band). The time delay can also be continuously varied by varying the shift frequency. The information packing density of the circuit can be increased by using more shift signals. The circuit lends itself particularly well to integration. In integrated form it requires transistors only; the capacitance between the base and a specially enlarged collector contact layer serves as the storage capacitor. A large number of stages can be accommodated on one chip; the article describes an experimental monolithic circuit containing 72 bucket stages, an amplifier stage (needed for compensating losses due to the base current of the transistors), a sampling circuit and an output circuit. The shift register can be used at bandwidths ranging from about 10 kHz to 15 MHz. The bucket-brigade shift register can also be made with MOS transistors. Since MOS transistors have no d.c. gate current, the individual circuits are simpler than with bipolar transistors, but the maximum frequency is lower and the dissipation higher. The article mentions a number of applications: the delay of audio and video signals, the correction of timing errors in magnetic-tape playback, the use of bucket-brigade circuits for making transversal filters, and the scanning of arrays of image-sensing devices.

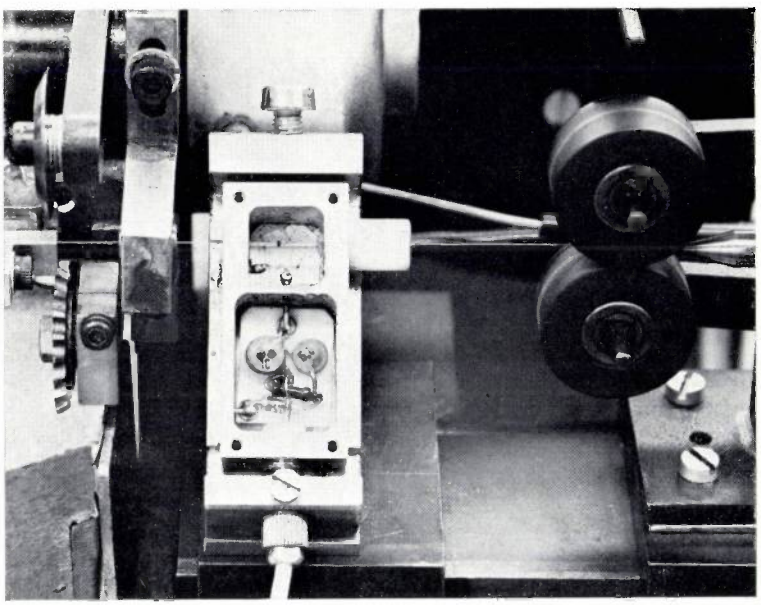
Measurement of wire-diameter variations

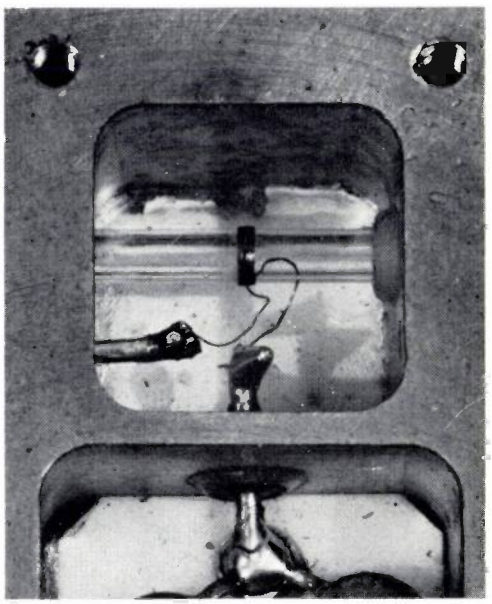
When small-diameter coils or filaments are wound in various forms on a winding machine, certain data is needed for designing, programming or setting up the machine for the winding process. The mechanical stress in the wire (e.g. copper wire with a thickness from ten to a few hundred microns) is an important factor and may vary considerably when winding coils of complicated shape. If this stress becomes too high, unacceptable variations occur in the diameter of the wire. These diameter variations can be continuously and very sensitively detected by passing the wire through a coil which, with stray capacitances present, forms a high-frequency oscillatory circuit. Variations in the diameter of the wire cause variations in the resonant frequency and damping of the oscillatory circuit.

The photographs show the measuring head, which contains a thin-walled quartz tube with an inside diameter of 1 to 3 mm — depending on the thickness of the wire to be measured. Around the tube a few turns of thin copper wire are wound. The measuring head also contains a circuit for rectifying and detecting the voltage across the oscillatory circuit, and the amplitude of this voltage varies as a function of the wire thickness.

The oscillatory circuit is fed by an oscillator which is connected to the head by a coaxial cable. The frequency of the oscillator is set to a point on the rising slope of the resonance curve of the oscillatory circuit. In this way, a slight change in tuning of the oscillatory circuit causes a relatively marked change in the amplitude of the voltage across the circuit.

The highest sensitivity is available when the coil around the quartz tube is dimensioned to resonate at about 300 MHz. A variation of $\frac{1}{2}\%$ in the thickness of copper wire is easily detected when the applied r.f. voltage is about 2 V rms.





The measuring head. The upper part contains a coil which is wound around a small quartz tube through which the wire to be measured is drawn. The lower part contains the rectifying and detection circuit. The front wall of the box, which is fitted with a glass window for observing the wire and coil, has been removed. The oscillator section of the measuring system is not shown.

The d.c. voltage signal which has been detected from the oscillator section is amplified and fed to a measuring or recording instrument. The signal can then be used for various purposes such as stopping the machine or regulating the winding stress during the coiling process.

P. G. Havas

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

An instrument for monitoring low oxygen pressures

N. M. Beekmans and L. Heyne

The use of stabilized zirconia as solid electrolyte in instruments for measuring low oxygen partial pressures has been common practice for some time. By increasing the number of contacts and improving their characteristics as reversible electrodes, the authors have developed an instrument which not only covers a very wide pressure range — approximately 10^{-30} to 1 bar — but which can also be used to control simultaneously the oxygen pressure in a gas system. The lower limit of the measuring range is so small that the instrument can also be used for measuring the partial pressure of gas-mixture components that react with oxygen.

In some solid oxides the oxygen ions become mobile at high temperatures, so that the material acts as a solid electrolyte. If gases of different oxygen partial pressure exist at opposite sides of a wall made of such an oxide, the diffusion of the negatively charged oxygen ions creates a potential difference just as in galvanic cells with liquid electrolytes. Fuel cells using this effect were proposed a long time ago [1], and the principle has also been applied for measuring oxygen partial pressures [2], using "stabilized zirconia" for example as electrolyte. At Philips Research Laboratories, Eindhoven, we have developed an instrument of this type that not only measures oxygen partial pressures down to very low values (about 10^{-30} bar at a working temperature of about 600 °C) but which also, *during measurement*, controls the concentration of oxygen in a gas or vacuum system by means of an extra electrode. The lower limit of its measuring range is so small that the instrument is also suitable for determining the partial pressure of those components in a gas mixture that can react with oxygen.

The construction of our instrument differs in quite a few respects from related instruments described elsewhere [3]. It consists of a stabilized-zirconia tube fitted with three cylindrical electrodes of porous platinum, two on the outside and one opposite them, on the inside; see *fig. 1a* and *fig. 2*. One of the outer electrodes (*M*) is used for measurement, and the other (*D*) is the dosing electrode. As the platinum is porous, the oxygen is able to pass through the electrodes. Moreover platinum acts as a catalyst in the dissocia-

N. M. Beekmans and Dr. L. Heyne are with Philips Research Laboratories, Eindhoven.

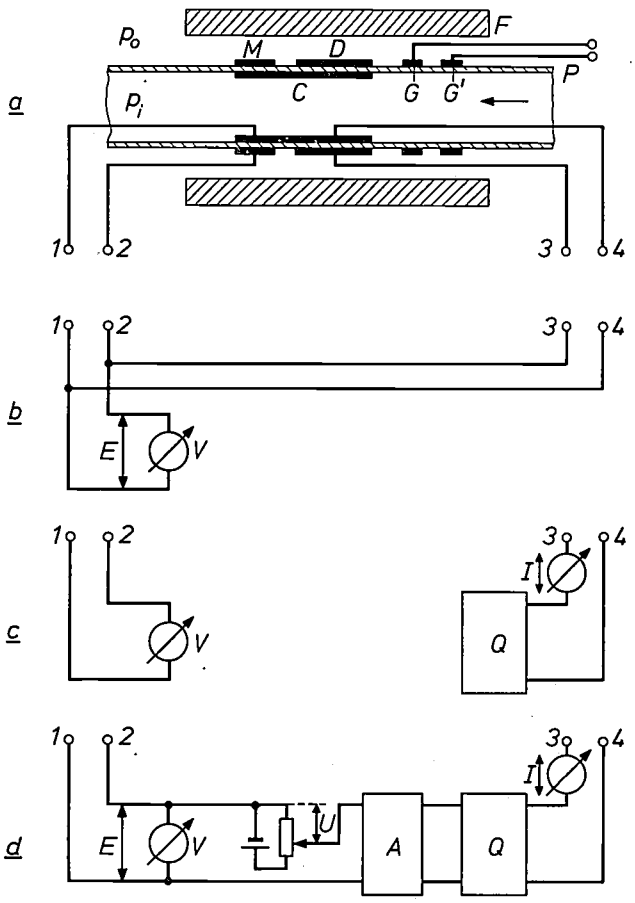


Fig. 1. *a*) Diagram of the instrument for monitoring the partial pressure of oxygen in a gas mixture. *P* tube of stabilized zirconia. *C* inner electrode, *M* measuring electrode, *D* dosing electrode, all made of porous platinum. *F* furnace. *G*, *G'* electrodes for temperature control. p_i, p_o oxygen partial pressures inside and outside the tube. The auxiliary equipment for various applications of the instrument is connected to terminals *1-4*: *b*) for measurement, *c*) for dosage, *d*) for automatically controlling the oxygen partial pressure. *V* voltmeter. *Q* current source. *I* ammeter. *A* amplifier. *U* adjustable auxiliary voltage.

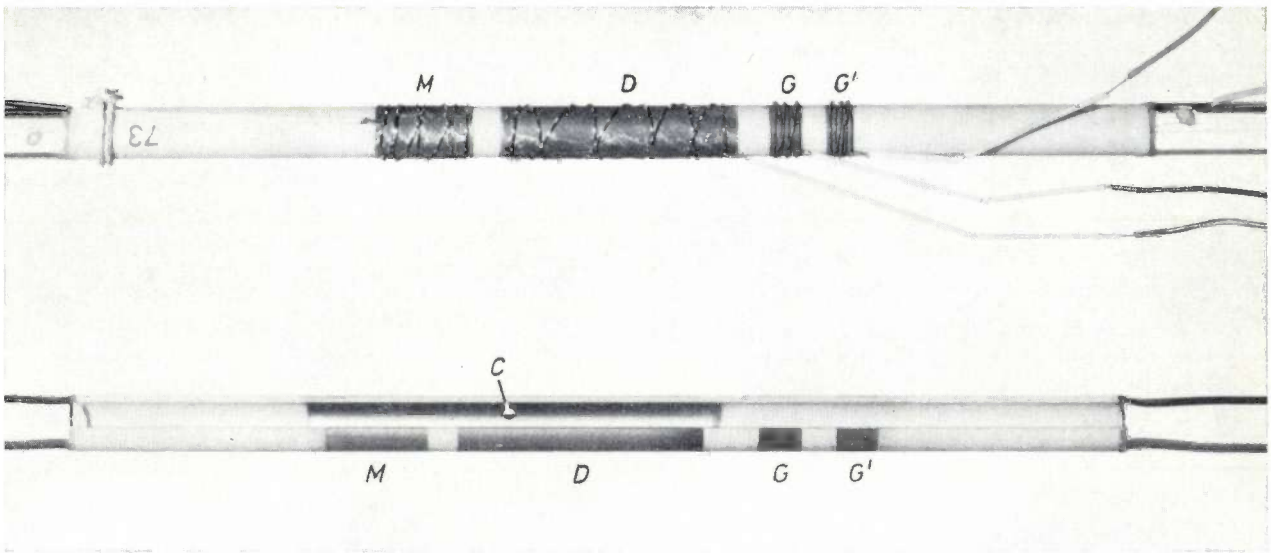


Fig. 2. The tube with electrodes. *Below*: a specimen where part of the wall has been removed to show the inner electrode. The letters have the same meaning as in fig. 1*a*.

tion and ionization of the oxygen and in the reverse processes, which is essential for the dosage control. (The function of electrodes G, G' will be discussed in a moment.)

The part of this measuring cell where the electrodes are situated is kept at a constant temperature by means of a furnace. If the oxygen partial pressures inside and outside the tube are p_i and p_o and the absolute temperature is T, there will be a potential difference E between the inner and outer electrodes (fig. 1*b*) that can be calculated with Nernst's relation for an electrochemical concentration cell. In our case this relation is:

$$E = \frac{kT}{2ze} \ln \frac{p_i}{p_o} . \quad \ldots\ldots \quad (1)$$

Here k is Boltzmann's constant, z the valence of the mobile ions (in our case 2) and e the elementary charge; the highest potential appears at the side of the highest partial pressure. The voltage E is thus a measure of the oxygen pressure p_i inside the measuring cell for a given p_o — in general the oxygen partial pressure of the atmosphere.

The voltage E, at a constant ratio of p_i and p_o, is proportional to the absolute temperature, and the temperature of the tube must therefore be very stable. Suitable temperature stabilization can be achieved in a straightforward way by making use of the high (negative) temperature coefficient of the electrical resistance of the zirconia. This is done by using the tube wall between two additional, narrow electrodes G, G' (see fig. 1*a* and fig. 2) as a sensing resistance in a circuit which controls the power supply to the furnace. In this configuration, the effects of the varying cooling rates, encountered when working with a variable gas-flow rate, are effectively compensated.

In practice the conduction of current in stabilized zirconia is not exclusively by mobile oxygen ions as suggested previously but also by electron flow. Although this contribution is often small under practical conditions, it may become important at low oxygen partial pressures and high temperatures when it tends to short-circuit the cell internally and therefore sets a low pressure limit to the measuring range below which equation (1) is no longer valid [4].

In less extreme conditions, where equation (1) still holds quite accurately, a second effect of the electron flow may become noticeable because it forms a load for the electrochemical cell. The resultant current flow causes the cell to behave as if oxygen is leaking through it. Although this leakage effect is usually small, it may become noticeable at low oxygen concentrations. The oxygen partial pressure at which the effect becomes significant depends on the system in which the cell is used and on the accuracy required. In vacuum systems where the pumping speeds are high a perceptible error will not appear above oxygen pressures of 10^{-10} to 10^{-14} bar, whereas in gas systems with low gas velocities the error may become apparent at 10^{-8} to 10^{-10} bar.

The electronic conduction can be decreased, and

[1] W. Schottky, Wiss. Veröff. Siemens-Werken **14**, No. 2, 1, 1935.
[2] J. Weissbart and R. Ruka, Rev. sci. Instr. **32**, 593, 1961.
[3] See, for example, H. S. Spacil, Metal Progress **96**, No. 5, 106, Nov. 1969, and R. G. H. Record, Instr. Practice **24**, 161, 1970 (No. 3).
[4] L. Heyne and N. M. Beekmans, to be published in Proc. Brit. Ceramic Soc., No. 19.

thus the leakage effect reduced, by working at a lower temperature. This entails a lower ionic conduction however, and also imposes stricter demands on the catalytic action of the electrodes, so that a compromise is necessary. Owing to the good catalytic properties and porosity of the platinum electrodes used, the operating temperature of the instrument could be chosen at about 600 °C. This is lower than usual for such instruments — typical figures quoted in technical literature are 850 °C and higher. It is even possible to use our instrument at a temperature appreciably lower than 600 °C but the response time, which is less than 1 second at the chosen temperature, will then increase somewhat.

Due to the suppression of electronic conduction the measuring range of our instrument is very wide — from approximately 10^{-30} bar to more than 1 bar, and this is several decades greater than that of many other systems used for determining oxygen partial pressures. But in many cases, other advantages are of more importance for the user, such as the ease with which the instrument can be handled, the absence of liquids and vulnerable parts, the excellent stability and simple calibration (see below), and also the fact that the oxygen content is not perceptibly affected by the instrument itself during the measurement. The logarithmic relationship makes it possible to cover the whole measuring range on one scale of the voltmeter, if required. This scale can then be calibrated either in pressure or in oxygen content, and the relative accuracy will be constant over the whole range. *Fig. 3* shows the instrument being used for measurement.

To provide *dosage* of oxygen in a gas system, a current source Q is connected between the inner electrode C and the dosing electrode D (fig. 1*c*). The current flowing in the electrolyte wall will be carried by oxygen ions, which are discharged at the positive electrode resulting in the formation of very pure, dry oxygen gas. If required, a voltmeter can be connected as before between C and the measuring electrode M to determine the effect of the dosing current. Depending on the direction of the current, oxygen is fed into or extracted from the system, and the current intensity is directly proportional to the quantity of oxygen: 1 mA for 1 minute corresponds to 3.5 mm^3 of oxygen at standard temperature and pressure. Since the quantity of oxygen thus supplied or extracted per unit time can be accurately determined by measuring the current, a gas flow can be provided with an exactly known oxygen content. This is the principle used for calibrating the instrument.

Since equation (1) is obeyed very accurately, all instruments can be calibrated to one and the same calibration curve by choosing the value for T. This is done by simply adjusting the temperature-control circuit for each instrument such that one decade change in p_1 (equation 1) corresponds to the same difference in E, e.g. the convenient value of 45 mV (corresponding to $T \approx 900$ °K). To realize this pressure change, a second identical instrument is connected for dosing and positioned upstream in the gas-flow system.

The calibration is now carried out as follows. The temperature of the measuring cell is generally such that the measuring error in the ppm range does not exceed a factor of 2. Helium is passed through the two series-connected instruments, and the p_1 of the oxygen already present in the helium without dosing is measured with the *non-calibrated* instrument. This will give a reading which could be 2 ppm for example. The current source of the dosing instrument is now switched on (3 mA) and thus an *exactly* known quantity of oxygen is introduced into the helium flow. For example this could correspond to 100 ppm exactly. The total oxygen concentration is then approximately 102 ppm which, according to equation (1), should correspond to 148.8 mV at the chosen potential difference of 45 mV per decade of oxygen pressure. The temperature-control circuit is now adjusted to give this reading of 148.8 mV on the voltmeter. If the current source is now switched off, the instrument measures the "basic" calibration gas again. Being better calibrated however, the instrument now indicates a more accurate value for p_1 of say 2.4 ppm. When the current source is switched on again the new reading must be 102.4 ppm of oxygen, corresponding to 148.7 mV. It only remains to make a very slight correction for the temperature. This procedure is repeated until all combinations of concentration and potential difference are consistent. In practice the required accuracy (0.1 mV at 100 ppm) is usually reached after one attempt.

In the above procedure the value of p_1 is dependent on the gas flow, and the latter is therefore measured accurately by means of a soap-film gas-flow meter. Corrections are also made for variations in atmospheric pressure and room temperature.

The current source Q can also be controlled by the voltage E across the measuring electrodes C and M via an amplifier A (fig. 1*d*): this results in *automatic control* of the oxygen partial pressure, which can be set to a given value by applying the difference between the voltage E and an adjustable voltage U to the amplifier input.

A few examples of the many applications of the instrument will be given below. Some of these might not occur immediately, for example measuring partial pressures of known components of a gas mixture that react with oxygen (such as H_2, CO, CH_4), and use as a detector in gas chromatography.

The instrument does not have to be modified in any way to measure partial pressures of gases reacting with oxygen, provided that certain other gases are either absent or known in quantity. As an example we will describe the important case of measuring the CO content in flue gases after incomplete combustion of fuel oil [5].

The main constituents of flue gases after incomplete combustion of fuel oil in air are water vapour, carbon dioxide, carbon monoxide and nitrogen. To determine the CO content a small fraction of the flue gas is passed through the instrument. In the state of thermal equi-

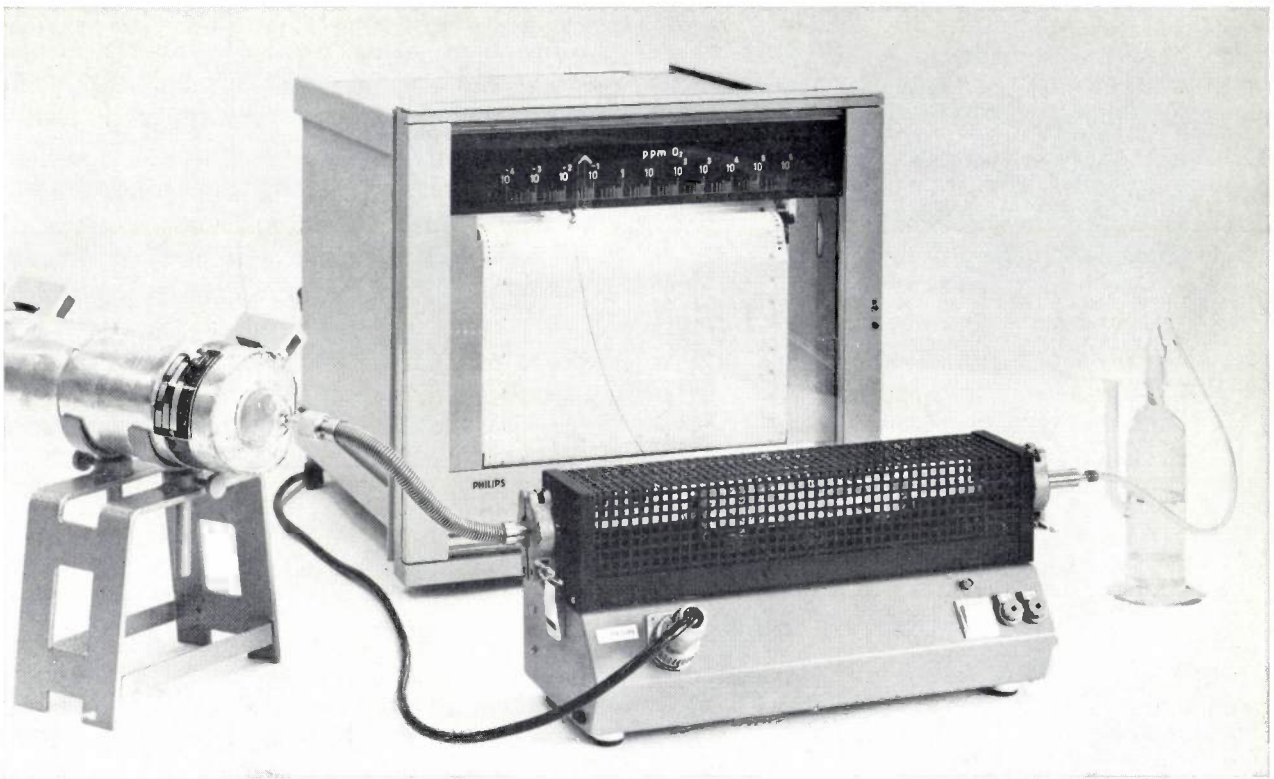


Fig. 3. Arrangement for recording the oxygen concentration in nitrogen during conduction measurements on oxides. The measuring cell, surrounded by the furnace, is in the upper part of the cabinet shown in the foreground; the lower part houses the furnace control system.

librium the oxygen partial pressure in this gas mixture is determined by the chemical equilibrium between CO and CO_2:

$$p(O_2) = \frac{p(CO_2)^2}{p(CO)^2} \cdot K_p. \qquad (2)$$

This equilibrium is established rapidly inside the measuring tube owing to the catalytic action of the porous platinum of the electrodes. The value of the equilibrium constant K_p is known, and using equations (1) and (2) we now find the carbon-monoxide partial pressure $p(CO)$ from:

$$E = -\frac{kT}{4e}\left[2 \ln \frac{p(CO)}{p(CO_2)} + \ln \frac{p_o}{K_p}\right]. \qquad (3)$$

For small CO contents, e.g. 10^{-4} to 10% by volume, the carbon-dioxide partial pressure $p(CO_2)$ can be assumed to be constant. The value of $p(CO_2)$ may then be taken as that which follows from the reaction equation for the *complete* combustion of the fuel oil. From equation (3) we see that for small CO contents, a logarithmic CO scale corresponds to the linear scale of the voltmeter. The leakage effect, mentioned previously as the factor that limits the measuring range, has scarcely any influence on measurements of this type in spite of the very low value of $p(O_2)$. This is because the leakage effect can only give rise to an extremely small relative change in the carbon-monoxide and carbon-dioxide partial pressures, and it is these pressures which determine the oxygen partial pressure as shown in equation (2).

This application demonstrates clearly the exceptionally wide measuring range of the instrument. For instance, a content of 1% by volume of carbon monoxide in the flue gases corresponds to an oxygen partial pressure in the measuring cell of approximately 10^{-20} bar, and this means that a transition from full combustion with 1% excess oxygen to incomplete combustion with 1% carbon monoxide causes a change of 18 decades in the oxygen content. This method of measurement is therefore particularly useful for optimizing combustion processes, e.g. for adjusting central-heating installations or for the monitoring of flue gases in boiler houses.

In contrast with this application using the very small lower limit of the measuring range, we will now give an example of the use of the instrument for keeping a check on an oxygen content which is of the order of 1%. In certain types of ferroxcube, the ratio Fe^{2+}/Fe^{3+} is an important parameter for certain physical properties. This ratio must be controlled during

[5] See also the article by Record [3].

manufacture by carrying out the sintering process in a protective gas containing oxygen. However, the required oxygen content is a function of temperature and the sintering furnaces, which are about 20 metres long, contain compartments at different temperatures. These are now separated by gas-tight partitions, and the oxygen concentrations in the three most critical compartments are controlled by means of the oxygen meter described here.

An application where oxygen content has to be measured between the values of the previous examples, is found in the manufacture of high-voltage transistors. A small amount of oxygen inside the metal encapsulation of these transistors can be detrimental to the breakdown voltage. The encapsulation must therefore be sealed in a nitrogen atmosphere containing no more than a few parts per million of oxygen. This operation takes place in a "glove box", where the oxygen content of the shielding gas is determined with our zirconia oxygen meter.

In the field of medicine the instrument can be useful for respiration studies [6]. The very fast response of the meter makes it possible to record the variation of the oxygen concentration in air exhaled during the normal breathing process.

There are many applications where the instrument can be used for monitoring a dose of oxygen. One example concerns the quantitative analysis of gas mixtures based on the automatic control of the oxygen partial pressure (see page 114). For example, to determine the unknown content of a reducing gas in an inert gas, the instrument can be set for a low oxygen pressure in the measuring cell, e.g. 10^{-10} bar. The amplifier ensures that this pressure remains constant and that during the passage of the gas mixture, the amount of oxygen introduced into the gas flow is exactly the amount needed for complete combustion of the reducing gas. The combustion equation gives the relation between the measured dosing current and the quantity of reducing gas introduced per unit time into the cell. For example the arrangement can be used for detecting the combustible components emerging from a gas-chromatograph column [7]. The sensitivity of this system compares well with that of the conventional system of flame-ionization detection, but the ionization yield is 100%, which is 10^5 times greater than that of the flame-ionization detector so that a sensitive current amplifier is not required [8]. Furthermore, the peaks in the gas chromatogram can be evaluated in terms of component mass without the need of calibration.

Although the applications discussed so far relate to gas systems, the instrument can equally well be used in vacuum systems. One example is the dosage of oxygen during reactive sputtering of silicon dioxide in a glow discharge, where the oxygen pressure of the sputtering gas influences the speed of deposition of the cathode material. Another is in the vacuum evaporation of nickel-chrome alloys such as those used in the manufacture of resistors. During this process a very low oxygen pressure is needed to obtain the appropriate resistivity and temperature coefficient. The use of an automatic control system for stabilizing the oxygen pressure in the low-pressure system of a thermobalance has previously been mentioned in this journal [9].

The use of the solid electrolyte has evident advantages over the use of a liquid electrolyte: in the first place only oxygen is transported and nothing else — not even water vapour; further, the tube acts as a partition between the gas system and the atmosphere, and finally it is only with this type of electrolyte that the oxygen can be supplied to a system as well as extracted from it without intermediate steps.

---

[6] I. E. Sodal, R. R. Bowman and G. F. Filley, J. appl. Physiol. **25**, 181, 1968.

[7] This application is at present being developed by Ir. B. Jansen of these Laboratories.

[8] For the definitions of the terms used here, see: J. Krugers, Ionization detectors in gas chromatography (Thesis, Eindhoven 1964), Philips Res. Repts. Suppl. 1965, No. 1, p. 50.

[9] P. J. L. Reijnen, Philips tech. Rev. **31**, 24, 1970 (No. 1).

Summary. In some solid oxides, e.g. stabilized zirconia, the oxygen ions become mobile at high temperature, so that such a material acts as a solid electrolyte. If gases of different oxygen partial pressure exist at opposite sides of a wall consisting of such an oxide, a potential difference is created by the diffusion of oxygen ions. This effect has been utilized in a simple, robust instrument which not only measures oxygen partial pressures — over 30 decades — but can also simultaneously be used for dosing oxygen in the gas or vacuum system by the passage of an electric current. The electrodes are made of platinum, whose good catalytic and porosity characteristics make possible an operating temperature of about 600 °C. As a result, the electronic conduction remains so small compared with ionic conduction that the lower end of the measuring range for oxygen partial pressures is extremely low (about 10^{-30} bar). The instrument can also be used to measure partial pressures of components in a gas mixture that react with oxygen. The oxygen pressure can be controlled automatically by using the resultant potential difference to control the oxygen dosage by means of a current. Some of the numerous applications of the instrument are described briefly or indicated. One of these is the measurement of the CO content in flue gases, where the highest sensitivity of the instrument occurs in the transition region from weak O_2 content to weak CO content. Another is the control of the O_2 content in the shielding gas during the manufacture of ferroxcube components and in the N_2 atmosphere used during the encapsulation of high-voltage transistors. Other applications are in respiratory examinations, the quantitative determination of reducing gas in gas chromatography, and the dosage of oxygen in low-pressure systems.

A method of determining noise in X-ray films

C. Albrecht and J. Proper

In photography, as in electronics, the signal-to-noise ratio is an important quantity; together with the contrast and sharpness of the image, it determines the perceptibility of details. The noise in radiographs is attributable both to the graininess of the film and to statistical fluctuations in the intensity of the X-ray beam. A method for measuring the transmission noise of X-ray films has been developed in which the film sample is scanned by a beam of light that is made narrow because of the relatively great thickness of these films. With this method several film samples can be rapidly investigated one after the other.

Introduction

To make a correct diagnosis a radiologist sometimes has to be able to observe details in a radiograph which are as fine as about 40 microns in diameter. Detail perceptibility may be limited not only by lack of contrast or by blurring but also by noise, caused on one hand by the grainy structure of the film emulsion (film noise) and on the other by the statistical fluctuations in the intensity of the X-ray beam used in making the radiograph (shot noise from the stream of X-ray quanta, briefly referred to as quantum noise).

Instruments known as microdensitometers have long been used for studying fine details in photographs made by visible light or infra-red. The noise is measured by scanning a film specimen with a beam of light and detecting the fluctuations in the quantity of transmitted light by means of a photomultiplier tube or photoelectric cell [1]. The operation of such an instrument is illustrated in *fig. 1*. An image of the diaphragm D_1 illuminated by the lamp S and the condenser lens C_1 is formed on the film F by the objective O_1, and the objective O_2 forms an image of the illuminated patch of film on the measuring diaphragm D_2, whose aper-

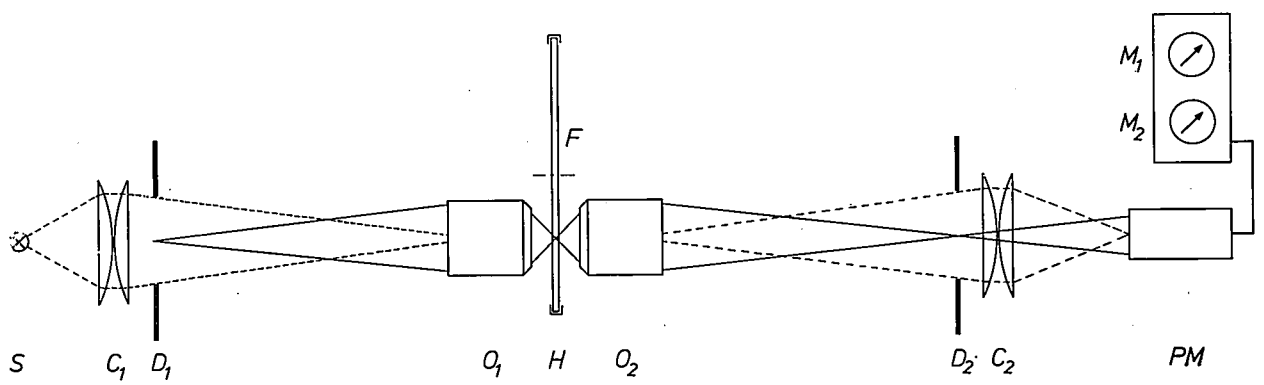


Fig. 1. Schematic arrangement of a conventional microdensitometer. S lamp. C_1 condenser. D_1 diaphragm which collimates the light beam to minimize scatter. O_1 and O_2 microscope objectives. F film. H eccentrically rotating film holder. D_2 measuring diaphragm. C_2 condenser. PM photomultiplier tube. M_1 linear instrument for measuring the average transmission T. M_2 square-law instrument for measuring the rms value of the fluctuations in the transmission (noise). The objective O_1 forms an image of D_1 on the film. An image of the resultant light spot is formed on D_2 by objective O_2. The transmitted light is directed by the condenser C_2 on to the photomultiplier tube, whose output signal is measured with M_1 and M_2.

Dr. C. Albrecht is with Philips Research Laboratories, Eindhoven; J. Proper is with Philips Medical Systems Division, Eindhoven.

[1] See for example G. C. Higgins and L. A. Jones, Photogr. Engng. 6, 20, 1955.

ture can be changed by changing the diaphragm. To avoid scattered light, D_1 is always made greater than D_2. The light transmitted by D_2 is focused on to the cathode of the photomultiplier tube PM by the condenser lens C_2.

If the film holder H is now made to rotate eccentrically with respect to the optical axis, the light spot will describe a circular path on the film and the photomultiplier tube will give an output signal which is proportional at every instant to the local transmission T of the film, averaged over the area of the light spot. The transmission, averaged over time, through the area of the film scanned by the light spot is measured with a linear instrument M_1, and the effective value of the transmission fluctuations is measured with an instrument M_2 which has a square-law characteristic. This effective value is equal to the standard deviation $\sigma(T)$ of the transmission T [2] and is a measure of the noise intensity. The ratio $\bar{T}/\sigma(T)$ then gives the signal-to-noise ratio of the transmission and can be used, at a given contrast, as the starting point for determining detail perceptibility.

As the layer of emulsion on ordinary photographic films has a thickness of the order of 10 μm, the light beam used for scanning such films can have a large angular aperture. In this way greatly reduced images can be obtained (a reduction factor of 200 is no exception). With a measuring diaphragm of reasonable dimensions it is therefore still possible to use a scanning beam whose cross-section at the film is only a few microns. This can be a great advantage in measuring films of very fine structure, e.g. films for cartography and microphotography.

In radiography maximum absorption of X-ray quanta is required, and for this purpose the film base is coated on both sides with a layer of emulsion, giving such a film a total thickness of 200 to 300 μm. The scanning beam used here therefore has to have a much smaller angular aperture. The fact that this entails less reduction is not a disadvantage, since the smallest details of interest in an X-ray film are generally larger than in the other photographic films mentioned.

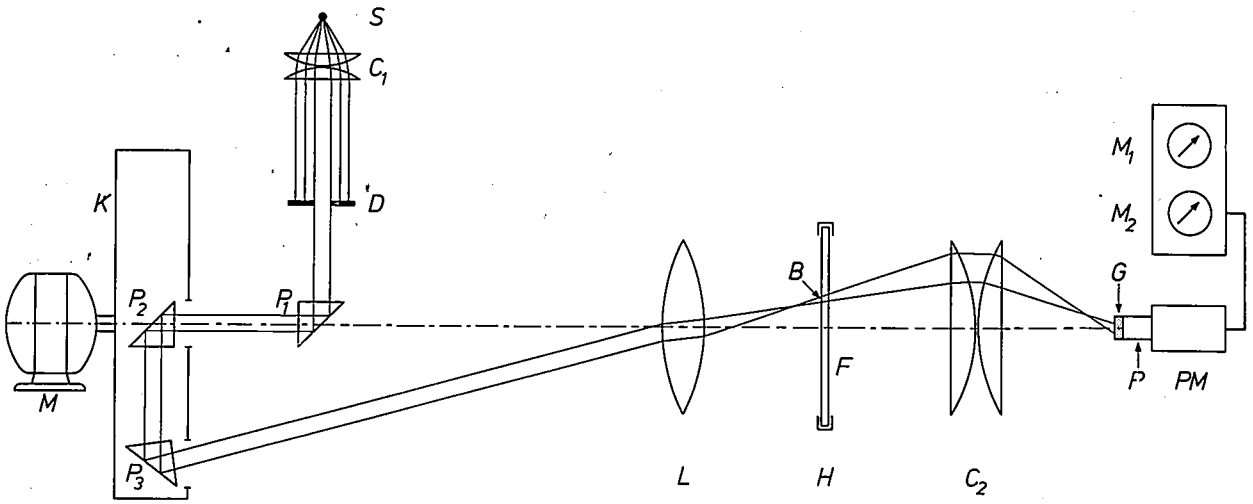


Fig. 2. New arrangement for measuring noise in X-ray films. S mercury-vapour lamp with a light-emitting area of only 300 μm cross-section (Philips type CS, 100 W). C_1 condenser. D diaphragm. P_1 adjustable prism which causes the axis of the parallel light beam to coincide with the optical axis of L and C_2. P_2 and P_3 prisms, mounted in the sleeve K fixed to the rotating shaft of M, the motor. L objective lens ("Xenotar" $f/2.8$; 150 mm). F film. H stationary, adjustable film holder. B image of the diaphragm D. The image B can be varied in diameter steps by changing D, the minimum diameter is 40 μm, the maximum diameter 1 mm. C_2 condenser. G opal-glass disc, cemented to a rod of transparent plastic coated with Al_2O_3. PM photomultiplier tube. M_1 linear-law meter, and M_2 root-mean-square meter. Some dimensional data: S-C_1 50 mm. C_1-D 350 mm. Optical path D-L 800 mm. L-F about 150 mm.

Another difference between the method of measurement we have developed and the conventional microdensitometer is that the light beam rotates and the film holder is kept stationary. This facilitates quick changing of the film specimens and also makes it possible to investigate different zones on one specimen in rapid succession.

The measuring arrangement for X-ray films

The arrangement we have designed is shown schematically in *fig. 2*. The light source S is a Philips type CS 100 W mercury-vapour lamp. This lamp was specially designed for measurement and projection purposes, and has a light-emitting area with a diameter of

only 0.3 mm. The condenser C_1 forms the light from the mercury lamp into a practically parallel beam, which is directed on to the interchangeable round diaphragm D. The aperture of this diaphragm can be varied in steps from 0.2 mm to 6 mm. The transmitted light beam is reflected by the adjustable prism P_1 in such a way that the beam axis coincides with the optical axis of the lens system L-C_2. From the prisms P_2 and P_3 the beam then goes to a lens L, an $f/2.8$ "Xenotar" objective with a focal length of 150 mm, which produces a 6 times reduced image of the diaphragm on the film F. The prisms P_2 and P_3 are mounted in a sleeve K connected to the shaft of the motor M. When the shaft rotates, the light spot describes a circular path on the film F. The light beam rotates at a speed of 25 revolutions per second, the diameter of the circular path is 35 mm. By changing D the diameter of the light spot on the film can be varied from 40 μm to 1 mm. The specimens of the film under investigation can simply be inserted in the stationary film holder H, without the motor having to be stopped, which is necessary in the microdensitometer in fig. 1. Moreover the film holder can be adjusted in three directions, lengthwise for good focusing, and horizontally or vertically for examining different parts of the film.

The light passing through the film is directed by the condenser lens C_2 on to an opal glass disc. This is connected to the window of the photomultiplier tube PM by a plastic rod coated with Al_2O_3, so that the light is diffusely distributed over the photocathode of the photomultiplier tube. This ensures that changes in the diameter and point of incidence of the beam have no effect on the output signal, which would otherwise be affected by local differences in the sensitivity of the photocathode. As in the microdensitometer of fig. 1, $\bar{T}$ and $\sigma(T)$ are measured by means of meters M_1 and M_2. It has been found that an axial displacement of the film holder by 1 mm in either direction has hardly any effect on the results, indicating that the depth of focus is sufficient to measure relatively thick X-ray films accurately.

Comparison of the measured transmission noise with the quantum noise

The noise measured by the method described is a combination of film noise and quantum noise. The share of the quantum noise can be determined separately in the following way. A measurement is made of the exposure [3] needed to produce a predetermined blackening of the film. This exposure and the quantum energy give the number of quanta incident on unit area of the film; this number is called the fluence, Φ. The fluence and the absorption coefficient of the film, α, which is determined from the mass-absorption coefficient of the silver bromide in the film emulsion and the mass of AgBr per m^2, give the number of quanta absorbed per unit area, $\alpha\Phi$. Since the quantum noise follows a Poisson distribution, the standard deviation or noise related to a surface area of diameter d is:

$$\sigma(\alpha\Phi) = \tfrac{1}{2}d\sqrt{\alpha\pi\Phi}, \quad \ldots \quad (1)$$

and the signal-to-noise ratio is:

$$\frac{\frac{1}{4}\pi d^2\alpha\Phi}{\frac{1}{2}d\sqrt{\alpha\pi\Phi}} = \tfrac{1}{2}d\sqrt{\alpha\pi\Phi}. \quad \ldots \quad (2)$$

Both quantities are in this case equal.

In order to compare the measured signal-to-noise ratio of the transmission with that which is caused by the quantum noise, it is desirable to express the transmission fluctuations in terms of equivalent fluence fluctuations.

To do this we must take the roundabout route of using the characteristic curve of the film. This gives the relationship between the density D, i.e. the blackening of the film, and the logarithm of the fluence Φ (see *fig. 3*). The density is defined as the negative logarithm of the transmission T:

$$D = -\log_{10} T. \quad \ldots \quad (3)$$

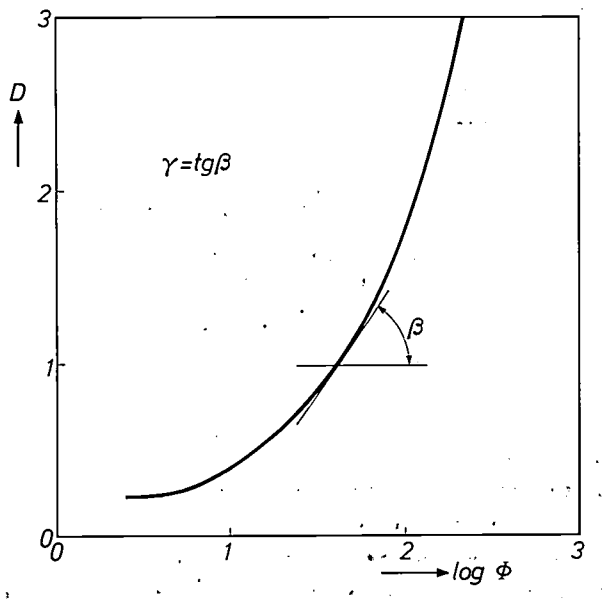


Fig. 3. Typical characteristic curve of an X-ray film, showing the density D plotted against the logarithm of the fluence Φ. The gradation γ of the film is the tangent of the slope of the curve β. Unlike the characteristic curve obtained on irradiation with visible light, the curve shown in this figure has no linear region. The gradation increases with increasing density.

[2] For continuously varying transmission the standard deviation $\sigma(T)$ is equal to $\{\int_0^{2\pi r}(T-\bar{T})^2\,dx/2\pi r\}^{1/2}$, where $2r$ is the diameter of the scanning track. This expression is equivalent to the rms value of T, measured over the circular path.

[3] This quantity used to be called the exposure dose; see, for example, J. Hesselink and K. Reinsma, Philips tech. Rev. 23, 56, 1961/62.

The gradation of the film, i.e. the steepness of its characteristic curve, is given by $dD/d \log_{10}\Phi = \tan\beta$ and is denoted by the symbol γ [4]. From (3) it can be shown that

$$\gamma = \frac{dD}{d \log_{10} \Phi} = -\frac{dT}{d\Phi} \cdot \frac{\Phi}{T}. \quad \ldots \quad (4)$$

In all parts of the film, therefore, a relative increase $\Delta T/\bar{T}$ of the transmission is associated with a relative decrease $\Delta\Phi/\bar{\Phi}$ in the fluence, which is given by:

$$\frac{\Delta\Phi}{\bar{\Phi}} = -\frac{1}{\gamma}\frac{\Delta T}{\bar{T}}. \quad \ldots \quad (5)$$

Here $\bar{T}$ and $\bar{\Phi}$ represent the average values of the transmission and fluence that relate to the background. We may therefore write:

$$\left(\frac{\sigma_{eq}(\Phi)}{\bar{\Phi}}\right)_d = \frac{1}{\gamma}\left(\frac{\sigma(T)}{\bar{T}}\right)_d, \quad \ldots \quad (6)$$

where the subscript d indicates that the measurements were made for surface area of diameter d. The quantity $\sigma_{eq}(\Phi)$ is the "equivalent" standard deviation of the fluence that the measured standard deviation of the transmission would cause if all the noise were due to the X-ray beam. We can now compare quantum noise and film noise in quantitative terms.

If a particular detail with the relative contrast $\Delta T/\bar{T}$ is to be perceptible against a background of transmission $\bar{T}$, then $\Delta T/\bar{T}$ must be at least three times the value of $\sigma T/\bar{T}$ [5]. Together with equation (6) this gives:

$$\{\bar{\Phi}/\sigma_{eq}(\Phi)\}_d \geqq 3\gamma\bar{T}/\Delta T. \quad \ldots \quad (7)$$

We have thus found an expression for the minimum value of the equivalent signal-to-noise ratio of the fluence at which a detail with a particular transmission contrast is still perceptible.

In order to be able to calculate the signal-to-noise ratio with the aid of (7) we still have to determine the value of γ at the given density. To measure $\tan\beta$ from the density increment accompanying a slight increase in $\log_{10}\Phi$ would not be accurate enough. For this reason, and also to avoid errors due to the Callier effect, we have devised a method of measurement in which the noise meter itself is used for determining γ. An underlying principle of this method is that the accuracy of a measurement is increased by frequent repetition of the observation.

For the determination of γ we start by making a test pattern. A lead wheel with about thirty spokes and an inside diameter a little greater than the diameter of the track scanned by the noise meter is placed on a strip of film. After the piece of film has been exposed to X-rays and developed, a high-contrast image of the spoked wheel is obtained. When the strip is scanned with the noise meter the light spot illuminates in turn the exposed parts of the film and the unexposed parts where the spokes were, giving a measurement of the root-mean-square (rms) variation of the transmission. A static measurement at and between the "spokes" gives the peak-to-peak value. From this the calibration constant of the noise meter is found, and with this constant we can thus translate a root-mean-square value into a contrast.

To measure the noise and determine the γ of a film the spoked wheel is now placed on a test strip of the film and a short exposure is made. The wheel is then removed and the test strip is exposed again until the parts of the film outside the spokes of the wheel reach the required density on developing. The initial exposure time is chosen so as to give a relatively small contrast between the spokes and the rest of the film (5 to 10% difference in transmission). When the image of the wheel is scanned with the noise meter in the way described above, the rms value of the transmission variations and the knowledge of the calibration constant give an accurate indication of the difference in contrast caused by a known small change in exposure. From this the value of γ can be calculated. The noise measurement is made on another part of the test strip.

Some results obtained with the new method

Fig. 4 shows the results of measurements on two types of X-ray film, one for industrial use (Gevaert type D4) and one for medical use (Gevaert type D10). The equivalent signal-to-noise ratio of the fluence $\{\bar{\Phi}/\sigma_{eq}(\Phi)\}_d$ is plotted as a function of the diameter of the scanning spot, both on a logarithmic scale for a density $D = 1.75$ (solid curves). The measurements were done with X-rays with rms quantum energies of 32 keV and 105 keV. The radiation was so closely filtered that for our purposes it may be regarded as monochromatic. The dashed curves give the value of the signal-to-noise ratio expressed as the number of absorbed X-ray quanta, calculated from (2). This quantity is equal to the value of $\{\bar{\Phi}/\sigma_{eq}(\Phi)\}_d$ that would be found if the transmission fluctuations were caused solely by the spatial statistical fluctuations in the distribution of the incident radiation and by the absorption processes in the film.

As can be seen in fig. 4, the equivalent signal-to-noise ratio of the transmission is virtually equal to that of the quantum noise for both films. Moreover the difference lies within the margin of error in the calculation of the quantum noise. One may therefore conclude that the limiting factor for both films is the quantum noise, since no higher value of the signal-to-noise ratio can be measured on the film than that determined by the quantum noise. From what we have said it follows that the resolution of our noise meter is more than sufficient to determine the noise in the very fine-grained D4 film. The instrument is therefore certainly suitable for measuring the usually rather faster and coarser films for medical applications.

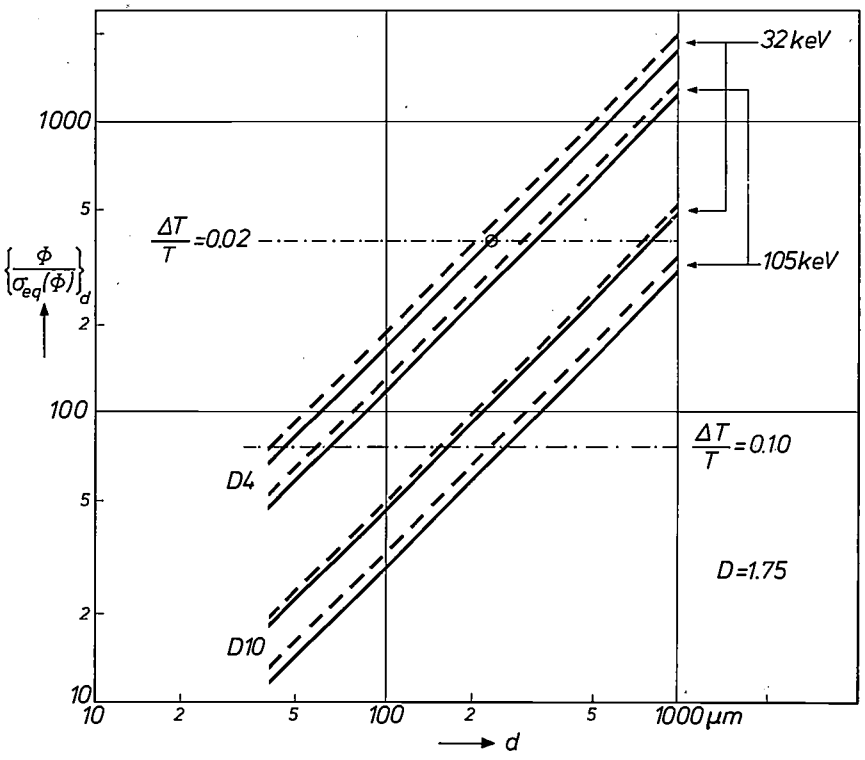


Fig. 4. Signal-to-noise ratio as a function of the diameter d of the scanning spot for Gevaert films D4 and D10 at two quantum-energy values. The film density D was in all cases 1.75. Solid lines: signal-to-noise ratio $\{\bar{\Phi}/\sigma_{eq}(\Phi)\}_d$ of the fluence obtained from transmission measurements. Dashed curves: signal-to-noise ratio calculated from d, the fluence Φ and the absorption coefficient α. Apart from slight differences probably due to inaccuracy in the data used for calculating the quantum noise, the measured signal-to-noise ratio is identical with the calculated values in all cases. This implies that in this case the resolution is determined by the quantum noise.

The horizontal chain-dotted lines give the value of the signal-to-noise ratio required for details with a relative transmission contrast of 2% and 10% to be perceived. For example, the circle indicates that for a D4 film with X-radiation of 32 keV and a contrast of 2%, details with a diameter of 250 μm are still perceptible.

It might appear that, as long as the film noise is lower than the quantum noise, the film could be made coarser to increase its speed. In this case, however, the number of quanta required for a particular density will be less, so that the signal-to-noise ratio will be lower. This can be seen clearly for the Gevaert D10 film, which is 30 times faster than the D4 film and has emulsion layers twice as thick, thus requiring 15 times as many quanta to produce the same density. The signal-to-noise ratio would thus have to be a factor of $\sqrt{15}$ smaller, and this is confirmed by the measurements.

Another thing shown by fig. 4 is that the signal-to-noise ratio is proportional to the diameter of the scanning spot: this also follows from equation (2).

For radiation with a quantum energy of 105 keV the signal-to-noise ratio is smaller than for radiation of 32 keV, because fewer quanta of 105 keV are needed to produce a given density.

One further and final conclusion can be drawn from fig. 4. The smallest transmission contrast $\Delta T/\bar{T}$ that is still perceptible to the eye is about 2%. It follows then from (7) that for $\gamma = 2.5$, the value of $\{\bar{\Phi}/\sigma_{eq}(\Phi)\}_d$ must be at least 375 if details with a diameter of about 250 μm on the D4 film and about 800 μm on the D10 film are to be perceived at an energy of 32 keV. For a contrast of 10% the D4 film gives a minimum detail size of about 50 μm and the D10 film one of about 150 μm at this quantum energy (lower horizontal chain-dotted line).

This method of measurement has also been used for X-ray films that are used with a fluorescent layer pressed against each side during the exposure, a technique used to increase the sensitivity of the emulsion layers [6] and reduce the X-ray dose. Our method can also be used for determining noise in radiographs obtained with the aid of an X-ray image intensifier.

Summary. The noise in a radiograph is due not only to the quantum noise of the X-ray beam but also to the grainy structure of the film emulsion. This film noise can be a limit to the quality of the radiograph and thus reduce its usefulness for diagnosis. A simple method has been developed for measuring the signal-to-noise ratio of X-ray films. It uses a rotating beam of light of small angular aperture which scans a stationary film specimen along a circular path. The small angular aperture gives a sufficient depth of focus for the measurement of relatively thick X-ray films (thickness about 0.3 mm). The use of a stationary film holder makes it possible to change film specimens quickly and to measure different parts of films in rapid succession. One interesting result of measurements by this method is that the signal-to-noise ratio of type D4 and D10 Gevaert films has been found to be limited by quantum noise. The method can also be used for measuring the noise characteristic of radiographs taken with the aid of an X-ray image intensifier.

[4] For a more detailed treatment of the properties of X-ray films, see D. H. O. John, Radiographic processing in medicine and industry, The Focal Press, London 1967.

[5] T. Tol, W. J. Oosterkamp and J. Proper, Limits of detail perceptibility in radiology particularly when using the image intensifier, Philips Res. Repts. **10**, 141-157, 1955.

[6] C. Albrecht and J. Proper, Medicamundi **11**, 44, 1965, particularly fig. 1.

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.

W. Albers & J. Verberkt: The SnSe-$SnSe_2$ eutectic; a *P-N* multilayer structure.
J. Mat. Sci. **5**, 24-28, 1970 (No. 1). *E*

D. Andrew, J. P. Gowers, J. A. Henderson, M. J. Plummer, B. J. Stocker & A. A. Turnbull: A GaAs-Cs-O transmission photocathode.
J. Physics D **3**, 320-326, 1970 (No. 3). *M*

F. G. M. Bax: Analysis of the FM receiver with frequency feedback.
Thesis, Nijmegen 1970. *E*

V. Belevitch: Interpolation matrices.
Philips Res. Repts. **25**, 337-369, 1970 (No. 5). *B*

F. Berz: Theory of low frequency noise in Si MOST's.
Solid-State Electronics **13**, 631-647, 1970 (No. 5). *M*

G. Blasse: Structure and luminescence of $MgGaBO_4$.
J. inorg. nucl. Chem. **32**, 700, 1970 (No. 2). *E*

G. Blasse: Luminescence of the tungstate group in scheelite and fergusonite.
Philips Res. Repts. **25**, 231-236, 1970 (No. 4). *E*

G. Blasse & A. Bril: Some observations on the luminescence of β-Ga_2O_3.
J. Phys. Chem. Solids **31**, 707-711, 1970 (No. 4). *E*

M. T. Borchert & J. S. C. Wessels: Combined preparation of ferredoxin, ferredoxin-$NADP^+$ reductase and plastocyanin from spinach leaves.
Biochim. biophys. Acta **197**, 78-83, 1970 (No. 1). *E*

C. J. Bouwkamp: Determination of the characteristic of a non-linear resistor by harmonic excitation.
Ingenieur **82**, ET 1-2, 1970 (No. 5). *E*

C. J. Bouwkamp & D. A. Klarner (Technical University of Eindhoven): Packing a box with Y-pentacubes.
J. recreat. Math. **3**, 10-26, 1970 (No. 1). *E*

K. H. J. Buschow: The erbium-copper system.
Philips Res. Repts. **25**, 227-230, 1970 (No. 4). *E*

K. H. J. Buschow & H. J. van Daal: Investigations on the resistivity of the compound $CeAl_3$.
Solid State Comm. **8**, 363-365, 1970 (No. 5). *E*

K. H. J. Buschow & A. S. van der Goot: The crystal structure of some copper compounds of the type RCu_6.
J. less-common Met. **20**, 309-313, 1970 (No. 4). *E*

H. B. G. Casimir: On supergain antennae.
Philips Res. Repts. **25**, 237-243, 1970 (No. 4). *E*

M. C. Collet: Recombination-generation centers caused by 60°-dislocations in silicon.
J. Electrochem. Soc. **117**, 259-261, 1970 (No. 2). *E*

C. Crevecoeur & H. J. de Wit: Electrical conductivity of Li doped MnO.
J. Phys. Chem. Solids **31**, 783-791, 1970 (No. 4). *E*

H. J. van Daal & K. H. J. Buschow: Anomalous behaviour of the electrical resistivity in the compound Ce_3Al_{11}.
Physics Letters **31A**, 103-104, 1970 (No. 3). *E*

H. Dammann: Blazed synthetic phase-only holograms.
Optik **31**, 95-104, 1970 (No. 1). *H*

R. Dändliker & K. Weiss: Reconstruction of the three-dimensional refractive index from scattered waves.
Optics Comm. **1**, 323-328, 1970 (No. 7). *E*

M. Davio: Extremal solutions of unate Boolean equations.
Philips Res. Repts. **25**, 201-206, 1970 (No. 4). *B*

M. Davio & G. Bioul: Representation of lattice functions.
Philips Res. Repts. **25**, 370-388, 1970 (No. 5). *B*

P. Delsarte: Automorphisms of abelian codes.
Philips Res. Repts. **25**, 389-403, 1970 (No. 5). *B*

J. C. Diels: Light pulse propagation in homogeneously broadened amplifiers.
Physics Letters **31A**, 26-27, 1970 (No. 1). *E*

J. C. Diels: Self induced transparency in near resonant media.
Physics Letters **31A**, 111-112, 1970 (No. 3). *E*

A. M. van Diepen (Natuurkundig Laboratorium der Universiteit van Amsterdam), **K. H. J. Buschow & H. W. de Wijn** (Natuurk. Lab. Univ. Amst.): Nuclear magnetic resonance and magnetic susceptibility of Pr_3Al_{11}, Nd_3Al_{11}, and $EuAl_4$.
J. chem. Phys. **51**, 5259-5263, 1969 (No. 12). *E*

P. Dolizy & R. Legoux: A new technology for transferring photocathodes.
Adv. in Electronics and Electron Phys. **28A**, 367-373, 1969. *L*

H. C. Donkersloot & J. H. N. van Vucht: Martensitic transformations in gold-titanium, palladium-titanium and platinum-titanium alloys near the equiatomic composition.
J. less-common Met. **20**, 83-91, 1970 (No. 2). *E*

G. Eschard & J. Graf: Quelques problèmes concernant les multiplicateurs canalisés pour intensificateur d'image.
Adv. in Electronics and Electron Phys. **28A**, 499-506, 1969. *L*

G. Eschard & R. Polaert: Tubes obturateurs pour photographie ultra-rapide au temps de pose d'une nanoseconde.
Adv. in Electronics and Electron Phys. **28B**, 989-998, 1969. *L*

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.
Proc. IEE **117**, 515-520, 1970 (No. 3). *M*

A. A. van der Giessen: De hydrothermale bereiding van keramische poeders.
Klei en Keramiek **20**, 30-38, 1970 (No. 2). *E*

W. J. A. Goossens & D. Polder: Size effects in the resonances of nonlocal helicon waves.
Phys. Rev. **187**, 943-950, 1969 (No. 3). *E*

A. D. C. Grassie (School of Math. and Phys. Sci., Univ. of Sussex, Brighton, UK) **& D. B. Green:** Transition anomalies of disordered aluminium films.
Physics Letters **31A**, 135-136, 1970 (No. 3). *E*

C. A. A. J. Greebe: Enhancement of the interaction between an acoustic surface wave in a piezoelectric and a drifted electron gas by means of a magnetic field.
Physics Letters **31A**, 16-17, 1970 (No. 1). *E*

G. Groh & G. W. Stroke (State University of New York, Stony Brook, N.Y.): Information retrieval from coded images formed by generalized imaging systems.
Optics Comm. **1**, 339-340, 1970 (No. 7). *H*

E. E. Havinga & M. H. van Maaren: Superconductivity and band structure.
Physics Letters **31A**, 167-168, 1970 (No. 4). *E*

E. E. Havinga, J. H. N. van Vucht & K. H. J. Buschow: Reply to the comments of K. A. Gschneidner *(pp. 255-256, on the paper "Relative stability of various stacking orders in close-packed metal structures").*
Philips Res. Repts. **25**, 257-258, 1970 (No. 4). *E*

H. F. van Heek: Emission profile of the Mg resonance line by Zeeman scanning.
Spectrochim. Acta **25B**, 107-109, 1970 (No. 2). *E*

B. J. Hoetink (Philips Glass Division, Eindhoven): Process dynamics of a glass furnace following a step change of one of the batch components.
Glass Technol. **10**, 84-89, 1969 (No. 3).

E. P. Honig & J. H. Th. Hengst: Current/voltage curves of $BaSO_4$ membranes.
Electrochim. Acta **15**, 491-499, 1970 (No. 3). *E*

J. T. Klomp & Th. P. J. Botden: Sealing pure alumina ceramics to metals.
Amer. Ceramic Soc. Bull. **49**, 204-211, 1970 (No. 2). *E*

H. Levine (Dept. of Mathematics, Stanford Univ., Cal.): Some problems in potential theory.
Philips Res. Repts. **25**, 207-222, 1970 (No. 4).

F. A. Lootsma & J. D. Pearson (Philips Information Systems and Automation Division, Eindhoven): An indefinite-quadratic-programming model for a continuous-production problem.
Philips Res. Repts. **25**, 244-254, 1970 (No. 4). *E*

J. Monin (Conservatoire National des Arts et Métiers, Paris), **J. Houdard & G.-A. Boutry** (Cons. Nat. A. et M.): Détermination des constantes optiques du césium dans le visible et dans le proche ultraviolet.
C.R. Acad. Sci. Paris **270B**, 279-282, 1970 (No. 4). *L*

K. Mouthaan: Niet-lineariteit van de lawine-looptijd-oscillator.
Ingenieur **82**, ET 4-7, 1970 (No. 5). *E*

M. Noé: Problèmes d'interconnexion optimale (2e partie).
Rev. MBLE **12**, 117-132, 1969 (No. 4). *B*

B. Paternostre & P. Wodon: Description et applications d'un générateur de macros.
Rev. MBLE **12**, 133-151, 1969 (No. 4). *B*

R. Plumier (Centre d'Etudes Nucléaires de Saclay, France) **& F. K. Lotgering:** Antiferromagnetic interactions between Fe^{3+} ions at a large distance in $Fe_{1/2}Cu_{1/2}Rh_2S_4$.
Solid State Comm. **8**, 477-480, 1970 (No. 6). *E*

A. Rabenau, H. Rau & G. Rosenstein: Über Chalkogenidhalogenide des Kupfers.
Z. anorg. allgem. Chemie **374**, 43-53, 1970 (No. 1). *A*

J. E. Ralph: Optical transitions in Tm^{3+} doped $Y_3Ga_5O_{12}$ near 2.0 μ.
Solid State Comm. **7**, 1065-1067, 1969 (No. 15). *M*

J. E. Ralph & T. L. Tansley: Photoelectric effects in cadmium sulphide MIS devices.
J. Physics D **3**, 620-623, 1970 (No. 4). *M*

H. D. Rüpke: Magnetodynamic modes in ferrite spheres for microwave filter application.
IEEE Trans. **MAG-6**, 80-84, 1970 (No. 1). *H*

U. J. Schmidt & W. Thust: Temperature stabilisation of the deflection pattern of a digital light deflector containing single prisms.
Opto-electronics **2**, 29-35, 1970 (No. 1). *H*

J. Schröder & F. J. Grewe: Darstellung und Eigenschaften von Wolframpentafluorid.
Chem. Berichte **103**, 1536-1546, 1970 (No. 5). *A*

E. Schwartz: Die Bandbreite von Anpassungsvierpolen mit zwei Reaktanzen.
Archiv elektr. Übertr. **24**, 179-186, 1970 (No. 4). *A*

H. Schweppe: Excitation of two adjacent resonances with a chosen frequency separation in a ceramic piezoelectric resonator.
IEEE Trans. **SU-17**, 12-17, 1970 (No. 1). *A*

J. M. Shannon, R. Tree & G. A. Gard (Atomic Energy Research Establishment, Harwell, England): Electrical characteristics of ion implanted boron layers in silicon.
Can. J. Phys. **48**, 229-235, 1970 (No. 2). *M*

L. A. Æ. Sluyterman & M. J. M. de Graaf: The fluorescence of papain.
Biochim. biophys. Acta **200**, 595-597, 1970 (No. 3). *E*

L. A. Æ. Sluyterman & J. Wijdenes: An agarose mercurial column for the separation of mercaptopapain and nonmercaptopapain.
Biochim. biophys. Acta **200**, 593-595, 1970 (No. 3). *E*

T. L. Tansley & J. E. Ralph: Photoeffects in metal-insulator-gallium arsenide diodes.
J. Physics D **3**, 807-811, 1970 (No. 5). *M*

A. Thayse: Transient analysis of logical networks applied to hazard detection.
Philips Res. Repts. **25**, 261-336, 1970 (No. 5). *B*

K. Timling: Scaled-particle theory of two-dimensional anisotropic fluids.
Philips Res. Repts. **25**, 223-226, 1970 (No. 4). *E*

H. J. L. Trap: Les effets de quelques traitements conventionnels sur la conductibilité superficielle du verre.
Verres et Réfr. **23**, 28-37, 1969 (No. 1). *E*

T. S. te Velde: Mono-grain layer solar cells.
Performance forecast of selected static energy conversion devices, 29th Meeting of AGARD Propulsion and Energetics Panel, Liège 1967, pp. 927-941. *E*

F. F. Westendorp: On the coercivity of $SmCo_5$.
Solid State Comm. **8**, 139-141, 1970 (No. 3). *E*

M. V. Whelan: Electrical behaviour of defects at a thermally oxidized silicon surface.
Thesis, Eindhoven 1970. *E*

Contents of Mullard Technical Communications **11**, No. 108, 1970:

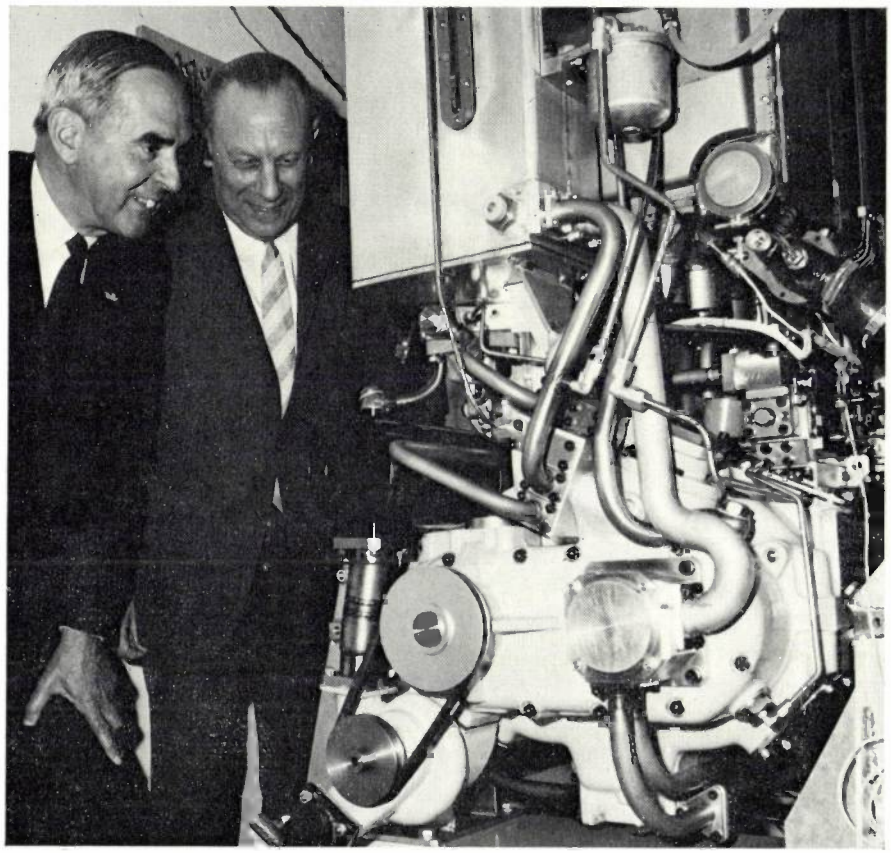
A. J. Guest: Channel multiplier plates for image intensification (pp. 170-176).
J. Wickens: 15 W class A audio amplifier (pp. 177-178).
R. F. Mitchell, R. G. Pratt, J. S. Singleton & W. Willis: Surface wave filters (pp. 179-181).
M. H. Jervis & F. D. Morten: Mercury cadmium telluride infrared detectors, at 5 μm and normal ambient temperature (pp. 182-184).

Contents of Valvo Berichte **16**, No. 1, 1970:

F. Weitzsch: Farbabweichungen als Folge von Exemplarstreuungen im Farbart-Teil von Farbfernseh-Empfängern (pp. 1-12).
B. **J. M. Overgoor:** Ein Operationsverstärker mit Feldeffekt-Transistoren in der Eingangsstufe (pp. 13-33).
Schliffbild eines Sperrschicht-Feldeffekt-Transistors (pp. 34-35).

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No. 5/6



One of our most loyal readers has recently celebrated the 40th anniversary of his joining Philips on 1st December 1930. Anniversaries do not usually receive any attention in our pages, but we feel that this one is rather special, since the reader we have in mind, one of the founders of our Review, is the President of the group of companies that bears his name. The photograph above shows our President (on the left) and Dr. Rinia, the former director of Philips Research Laboratories, looking at an experimental Stirling motor for heavy traction work. We join the authors in presenting to Ir. F. J. Philips this issue, whose articles are all on subjects that have attracted his special interest.

Hydraulically driven precision lathe with hydrostatic bearings for the main spindle and for the carriages, designed in the Philips Research Laboratories. The two numeric displays above the lathe show the coordinates of the cutting tool in ten-thousandths of a millimetre. Machine tools are the subject of the article on the adjoining page.

Forty years of workshop technology

B. L. ten Horn

Introduction

In many kinds of industry workshop technology is a vital and basic activity. Yet it is not so easy to define exactly what "workshop technology" implies, for there are so many kinds of workshop using such a wide variety of techniques.

In the Netherlands, long a sea-faring nation, engineering works and boiler-makers' shops first grew up around the shipyards, and they probably represent the oldest form of workshop technology in this country. It was a natural step for these shops and works, which originated around the ship's engine and its boiler, to help to provide for the needs of the power and chemical industries.

Construction yards for building cranes, bridges, roofing structures, etc., form another and entirely different category of workshops. In the mass production of consumer goods workshop technology plays an important part in the manufacture of production tools, such as special machine tools, punches and dies. Workshop techniques are also indispensable to research.

Scientific and medical equipment, telecommunication systems and equipment used for national defence and transport, are all examples of products that often require the highest possible degree of mechanical refinement. It is no exaggeration to say that many of the new fields opened up to science, such as nuclear technology, space travel and advanced medical methods, are accessible to science only provided workshop technology is able to supply the necessary equipment.

The vital importance of workshop technology to very many industries stems from its role in the production of the capital goods that are essential for the running of those industries.

A new viewpoint

In the last two decades numerical control has led to a new, refreshing approach to workshop technology, which has also shed light on problems outside the direct field of application of this new method of processing. In the jargon of cybernetics, which includes numerical control as one of its specializations, there is much to say that is highly relevant to workshop technology.

Every form of mechanical production involves a large quantity of mainly geometrical information. This information originates in the conception and development of a product. In the specification and detailing of a design in the drawing office an enormous amount of information is added, which is set down, in all its multitudinous ramifications, in engineering drawings [1]. Behind the indications on the parts lists of these drawings there is a wealth of information packed in standards sheets.

The "workshop" in its various forms mentioned above always contains the interface between the world of the designer, dealing in abstract ideas, and the world of the actual manufacture of the product by production machinery.

The skilled workshop technician is perhaps best characterized as the man who can read a drawing and who, with his machines and tools, can make whatever is described in the drawing. We invariably find him at the interface between design and production, usually operating a machine-tool or lathe, or perhaps in the assembly shop, using the drawings to assemble the parts that make a product.

The processing of the flow of information referred to above is time-consuming and is moreover a potential source of human error. The double delay involved in thinking out the right way of setting the machine by reference to the drawing, and then checking and often rechecking the setting decided upon — for which the Americans have coined the apt term "ear-scratching time" — makes productivity at the interface of design and production unavoidably low.

In the one-off manufacture of capital goods, little can be done about this state of affairs. As soon as the manufacture becomes repetitive, however, it becomes possible to make the effort required for the information processing just once for the manufacture of a whole batch of products. The information must then be programmed into the setting of the machine, into the jigs and fixtures and other "hardware" in such a way that the product can be made with no need for a drawing to be consulted. How far one can go in this direction depends on the size of the production run. In mass production it always pays to record the entire design

Prof. Ir. B. L. ten Horn is a Scientific Adviser with the Philips Machine Workshops Division and is a Professor Extraordinary of Delft Technical University. He is also a member of the Board of Management of the Philips Centre for Manufacturing Techniques, with special responsibility for Mechanical Process Operations.

[1] H. Huizing, Numerieke besturing: integratie van construeren en produceren?, Ingenieur 80, W 197-203, 1968.

information in punches, dies, special machines, test and inspection devices, etc., so that the reading of drawings is no longer necessary in the actual production and the rate of production can be stepped up.

The design/production interface has now been shifted to the toolmaking shop and the specialized machine shop, where we again find the skilled operator with his drawings, his universal tools and his flexible methods of working.

The position of workshop technology between design and production was undoubtedly recognized before the advent of numerical control. Numerical control, and particularly the integration of design and production with the aid of modern information technology — computer-aided design (CAD) and computer-aided manufacture (CAM) — have forced engineers into more advanced thinking in this field and have provided us with the terminology to define our ideas on this subject [2].

The characteristic trends of the forty years reviewed in this article are rationalization and increase of scale, which have frequently turned one-off production into batch production and batch production into mass production. Here the skilled operator has moved more and more towards the preparation for production.

In the last fifteen years we have seen the emergence of an entirely new development. The higher degree of automation brought about by numerical control is no longer directed towards a shift in the labour-intensive link in the production process from production itself to the manufacture of the means of production, but towards automation of the information processing on the spot, in order to rationalize this difficult production phase as well.

Some important techniques

Metal-cutting or machining is still one of the principal subjects of workshop technology. The combination of skilled operator and a universal metal-cutting machine remains the best means of making components of mechanisms direct from the information in the drawing. The interplay of rotary and linear movements is applied in such a way as to generate the shapes required for the components of mechanisms that move in a related way. There are very few other shaping techniques that give sufficient accuracy for making properly mating machine parts.

Mechanical metrology, including the theory of fits and tolerances, is rightly regarded as belonging to workshop technology. Its connection with the manufacture of components and with assembly is obvious.

In sheet-metal work the distinction between parts manufacture and assembly is often of little significance. The product is made by joining together relatively simple parts and jointing techniques such as resistance and fusion welding, brazing, bonding, etc., are of vital importance.

In small-batch production by sheet-metal fabrication it is usually preferable to make parts by bending flat pieces of sheet, since special tools are required to form compound-curved surfaces. In the aircraft industry it is absolutely essential to be able to make surfaces with compound curvature. This industry has given us a number of workshop methods that can be used to produce small batches of products with compound-curved surfaces from sheet metal.

Toolmaking has always been a skill on its own. Toolmakers have learnt to think in shapes "the other way round": a plastic product, for example, is for them a cavity in a mould. In machining operations this reversal of the shape of a product often creates intractable problems of accessibility. To solve these problems toolmakers have developed a number of reversal techniques such as hobbing, electroforming and spark machining.

Finally, the whole of the data-processing activity in planning, tooling up, work study, etc. also constitutes an essential part of workshop technology. However, in this historical survey we shall limit ourselves to the purely engineering aspects of data-processing seen in numerical control.

Forty years of machining technology

The thirties were important years for machining. High-speed steel was well established as a cutting material. The cumbersome overhead drive systems (*fig. 1*) had disappeared or were disappearing from most workshops. This led not only to better siting of the machines but also made it possible to supply as much power to the machine as was considered necessary.

In 1927 the firm of Krupp demonstrated for the first time the new "Widia" cutting material at the Leipzig Fair [3]. This was what we now know as cemented carbide or "hard metal". This event proved to be just as significant as the demonstration by F. W. Taylor and M. White of high-speed steel at the Paris World Exhibition in 1900. The new material once again meant an increase in cutting speed by a factor of about five. Nevertheless, for reasons which we shall explain later on, the new cutting material made disappointing progress in the thirties.

Meanwhile the development of the older cutting material, high-speed steel, received fresh incentive because of the competition of hard metal. The original invention by Taylor and White related rather to a new heat treatment for the already familiar tungsten-alloyed "Mushet steel" than to a new type of steel, even though high-speed steel proper originated in its turn from the adaptation of tungsten steel to the new heat treatment.

The further developments in the consolidation of high-speed steel in the thirties were therefore mainly directed more towards perfecting the heat treatment than the alloy itself. These activities were made possible by improvements in temperature control and the introduction of salt-bath hardening. The results of the heat treatment were verified by life tests, which could now be speeded up through the knowledge of the Taylor relation between cutting speed and tool life. W. F. Brandsma's disc test, developed at Philips Research Laboratories, is a good example of this approach [4].

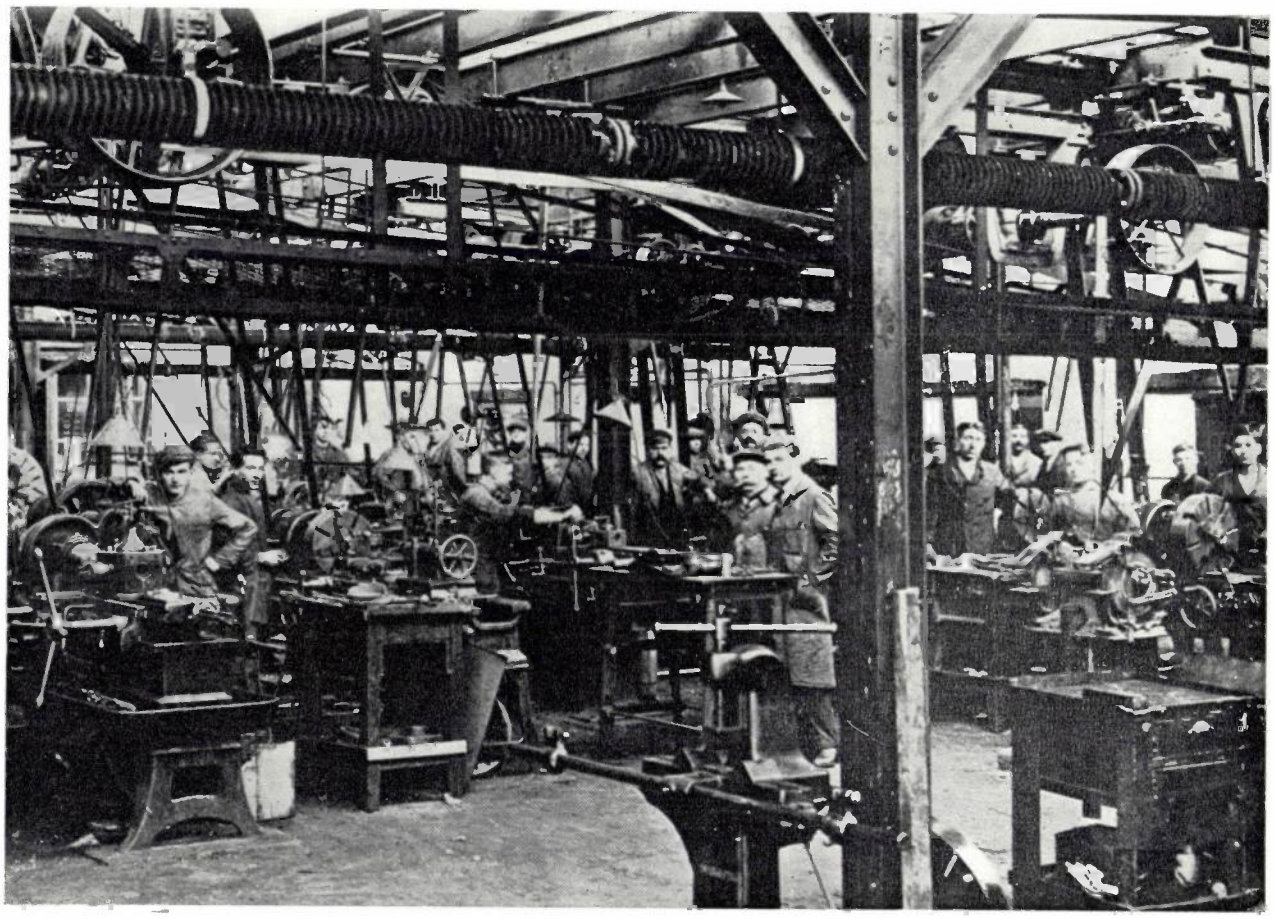


Fig. 1. The Philips Engineering Workshop in 1917. Groups of machine tools were driven with the aid of belts by overhead driving shafts.

The new hard metal, which was originally developed to make dies for drawing tungsten wire, was a sintered product consisting of tungsten carbide and cobalt. Now at the temperature of more than 1000 °C which is reached on the rake face of the tool at the cutting speeds used in carbide turning, tungsten carbide possesses a solubility in iron which is by no means always negligible [5]. Because of this solubility the earliest carbide cutting tools gave rapid crater wear at the rake surface, which meant that hard metal was not very suitable for machining steel. This kind of wear was not found in the machining of non-ferrous metals and cast iron.

An even more troublesome property of the new cutting material was its brittleness. The pressure of the chip causes tensile stresses in the tool material immediately behind the cutting edge. With a cutting material as brittle as hard metal these stresses can lead to chipping or complete breakage of the tool. To keep these stresses low the wedge angle (see *fig. 2*) must be increased at the expense of the rake angle. This was not serious, however, since at the high cutting speeds permissible with hard metal there is good chip flow even when the rake angle is small.

The practical experience gained with high-speed steel

[2] J. L. Remmerswaal, De machinefabriek van morgen, Ingenieur **82**, A 626-631, 1970 (No. 33).

[3] (*Anonymous*) Widia, Metaalbewerking **1**, 20-21, 1934/35.

[4] See: A speed-increment test as a short-time testing method for estimating the machinability of steels, Philips tech. Rev. **1**, 183-187 and 200-204, 1936 (compiled by P. Clausing).

[5] E. M. Trent, Some factors affecting the wear of cemented carbide tools, Machinery (Eng.) **79**, 823-828 and 865-869, 1951.

had indicated that rake angles should be between 20° and 30° and the machine operator found it difficult to accept the new rake angles of 0° to 6°, which were the optimum values for carbide. Moreover, tungsten carbide could only be ground with diamond. Since the diamond grinding wheels were expensive and easily damaged, the operator was not usually allowed to resharpen the tools himself. The trouble and loss of time which this involved did not endear the new cutting material to the operator, who of course could do the job himself when high-speed steel was used. What is more, the machine tools were often not powerful enough to achieve the proper cutting speeds for hard metal. They were not sufficiently stable, and unwanted vibrations led to early failure of the cutting edge of the tool.

was when it was found possible to sinter alumina with such purity and freedom from internal stresses and defects that it could suitably be used for machining most metals. The experience meanwhile gained with carbide paved the way for the successful application of this even more brittle cutting material. Aluminium oxide could not, however, be joined to the metal shank by the usual method of brazing and it now became absolutely essential to clamp the tool tips in the same way as had already been used here and there for cemented carbides. To avoid tensile stresses near the cutting edge of the ceramic tool, a wedge angle of 90° was used and in addition the cutting edge was chamfered (fig. 2).

All the measures necessary for making successful use of the ceramic cutting materials were also found to be

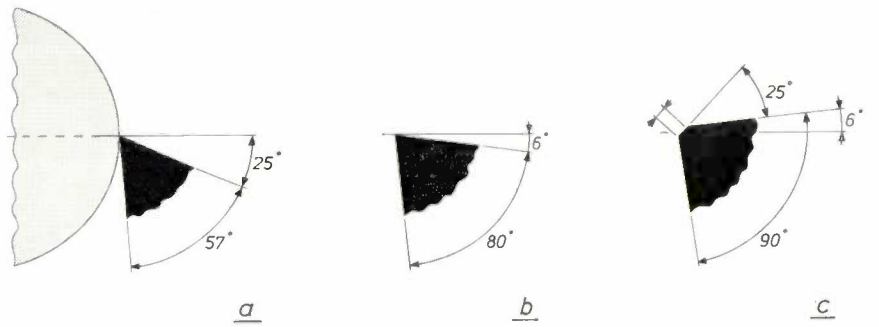


Fig. 2. The rake angle and the wedge angle of lathe tools made from high-speed steel (*a*) cemented carbide (hard metal) (*b*) and ceramic (*c*). The rake angles are 25°, 6° and −6° respectively, the wedge angles 57°, 80° and 90°. The cutting edge of the ceramic tool is bevelled to reduce the tensile stresses that occur during machining.

All these reasons taken together meant that it was not really until after the Second World War, when mass destruction followed by Marshall Aid led to a large-scale renewal of the stock of plant and machinery, that a real breakthrough came in the application of hard metal.

Even in the early years, experiments were being carried out in the United States of America (Fansteel "Ramet" 1930) and elsewhere with sintered products of the carbides of titanium and tantalum. Krupp had great success with mixed carbides of tungsten and titanium containing about 10% of TiC [6] (Widia X, 1931). The lower solubility of titanium carbide in the material of the hot chip gave much less crater wear than with the tungsten-carbide hard metal. The mixed carbide of titanium and tungsten was much easier to sinter to a non-brittle product than the pure titanium carbide. Moreover, with the titanium carbides it was very difficult to braze the tool tip to the shank unless the titanium-carbide content was kept below a critical value.

The story of titanium-carbide hard metal was to have a remarkable sequel, closely bound up with the development of ceramic cutting materials in the late fifties. This the very measures to get the best out of the titanium carbides that had been prematurely born in the thirties. The titanium-carbide hard metal, with nickel used as a binder and molybdenum carbide as an additive, and bearing a close resemblance to the material of the thirties, was successfully used by the Ford Motor Company in 1959 [7] [8].

The use of titanium-carbide hard metal offered many of the advantages of the ceramic cutting material. Since it was less brittle, however, it was much easier to put to practical use. Ford's example found so many followers that it definitely slowed down the advance of the ceramic cutting materials in the sixties.

At the present time it looks as if this phase has run its course. Ceramic cutting materials are coming into wider use again, largely because of the pioneering work with titanium carbide, which is rather more brittle than the older hard metal but not so brittle as ceramic. Titanium carbide has recently gained yet another field of application as a wear-resistant coating on ordinary cemented carbide [9].

The clamp holder (*fig. 3*) did more than solve the problem of brazing the ceramic material and the tita-

nium carbide. Even with the older hard metal, which could be easily soldered, the absence of brazing stresses improved the durability of the hard-metal tip. What is even more important, however, is that with the latest clamping devices the operator has a number of sharp cutting edges at his disposal. Freeing the carbide tip (or insert), indexing and reclamping it to bring a new cutting edge into play are jobs that require some care but which can quite safely be left to the skilled operator. When the new cutting edge is turned into position the setting of the tool remains unchanged, which is not the case when a tool has to be completely replaced. Once again, the skilled operator can do the job himself, just as when a high-speed steel cutting tool was used.

The use of hard-metal inserts, also known as "throw-away" tool tips, has not been confined to lathe tools. Mechanically clamped inserts are also being used with success in hard-metal milling cutters.

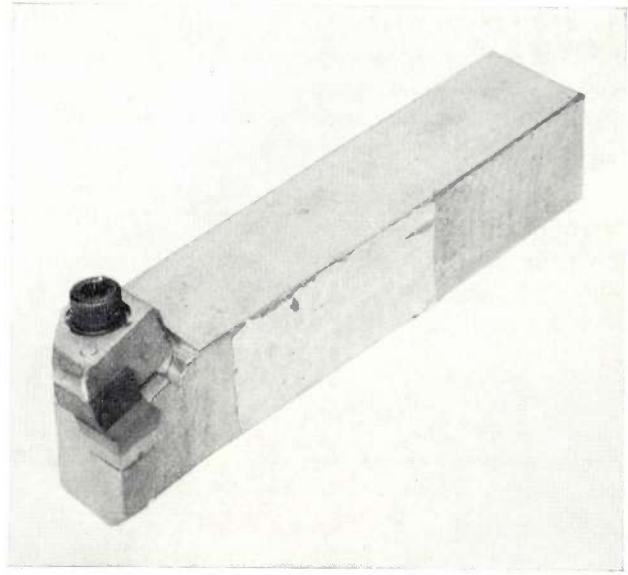
Machine tools

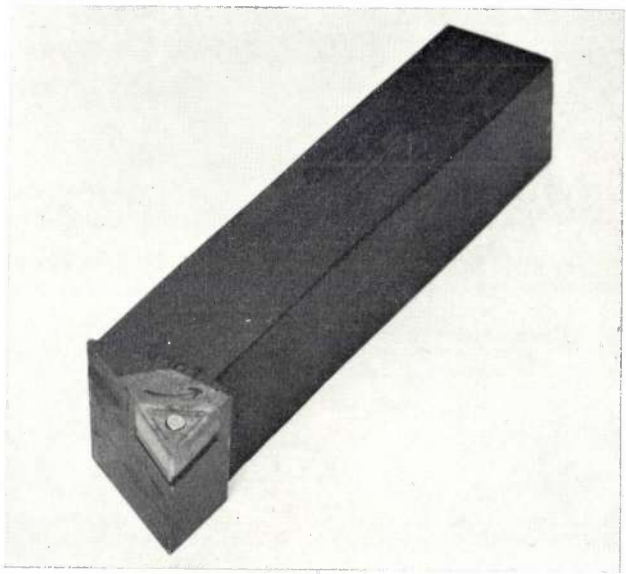
Machine tools are machines which shape a workpiece by the kinematic generation obtained through the interplay of rotary and linear movements. The very high accuracy that can be achieved is the most characteristic feature of this method of shaping.

In 1927 G. Schlesinger published his "Prüfbuch für Werkzeugmaschinen", which described test methods and standards for determining the accuracy of machine-tool movements. This book ushered in a new era. Before that time evaluation of a machine was usually left to the experience of the skilled operator. Now a foundation had been laid for an objective evaluation based on scientific measurement.

A shortcoming of Schlesinger's test methods and standards, important though they were, was that they had to be applied to idle machines. They did not take into account the stringent requirement for the movement of the machine to be accurate in spite of appreciable dynamic stress that appears when metal is being cut. In fact from 1930 to 1950 these problems became even more important, because the power applied to machines was increased while machining with the new high-grade but brittle cutting materials required stable and vibration-free operation. These matters however fall outside the scope of Schlesinger's "Prüfbuch".

The movements of a machine tool can be divided





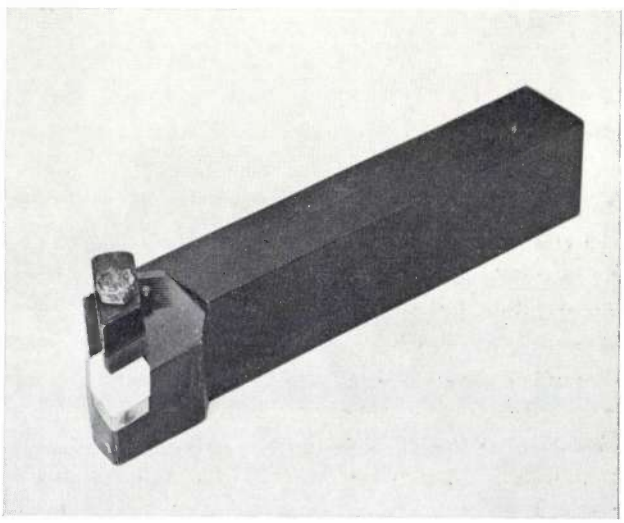


Fig. 3. Cutting tools with interchangeable tip ("throw-away" tip) that has more than one cutting edge. As soon as a cutting edge becomes blunt the next one is brought into use. The cutters showed here are clamped to the shank in three different ways.

[6] E. Ammann and J. Hinnüber, Die Entwicklung der Hartmetallegierungen in Deutschland, Stahl und Eisen **71**, 1081-1090, 1951.
[7] New titanium carbide tools outlast ceramics, Amer. Machinist **103**, No. 6, 125-126, 1959.
[8] B. L. ten Horn and R. A. Schürmann, Beproeving van titaancarbide hardmetaal, Metaalbewerking **26**, 1-3, 1960/61.
[9] P. Groen and C. A. van Luttervelt, Hardmetalen wisselplaten met titaancarbidelaag, Metaalbewerking **35**, 495-500, 1969/70 (No. 20).

into the cutting movement, the feed movement, and positioning. These movements together encompass a considerable speed range, which increased with the introduction of high-speed steel, and yet again with the introduction of carbide.

Up to the thirties most main spindle bearings of centre-lathes and milling machines were plain bearings. With the higher spindle speeds needed for machining with hard-metal cutters there was a marked increase in the top limit of the speed range that the machine had to cover. A speed range of 1 : 40 to 1 : 60 was fairly normal for a lathe. It is very difficult, however, to design a plain main bearing that gives adequate hydrodynamic stiffness at the lowest speeds in such a range and does not get too hot at the highest speeds.

Towards the end of the thirties roller and ball bearings began to come into use for the main bearing of lathes and milling machines, particularly in the United States. The fact that good results could be achieved with these much less speed-sensitive bearings, provided they are of high quality and correctly mounted, was proved by the AI-DR 1 instrument lathe (*fig. 4*). In this lathe, designed in 1940 and the first machine tool to be made in any quantity in the Netherlands, the headstock spindle was mounted on ball bearings.

After the war the use of pre-stressed ball and roller bearings made good headway. The Monarch lathe, supplied to Europe in large numbers under Marshall Aid, and also the highly successful Cazeneuve lathe both had roller bearings in the main spindle.

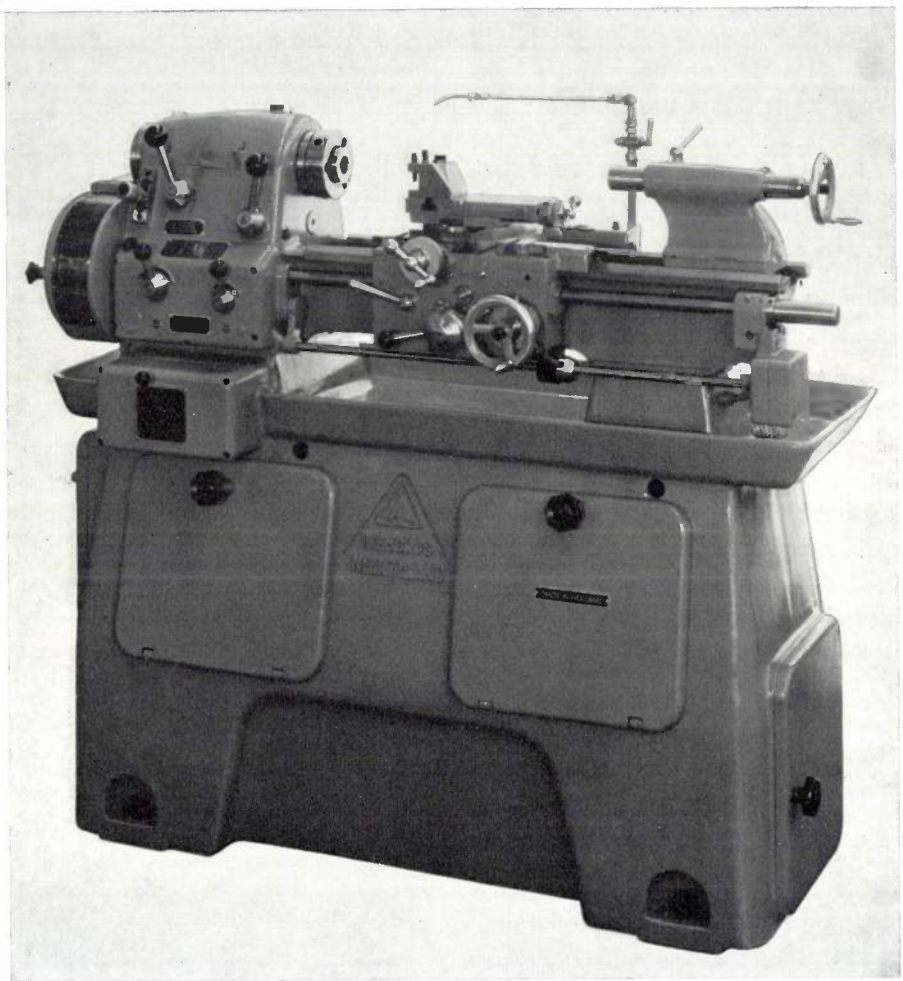


Fig. 4. The Dutch AI-DR 1 instrument lathe, designed in 1940.

The recirculating-ball leadscrew and nut can be made to give virtually absolute accuracy in location; this is vital in numerical control. The use of balls and rollers in the elements responsible for the linear movement of machine tools came about twenty-five years later, however, than their use in the main spindles of lathes, milling machines and grinding machines.

The bearings based on rolling friction have certain disadvantages which are not unimportant. One of them is the virtual absence of damping in both journal and guide bearings. Moreover, a roller or ball bearing is no more accurate than the elements it is made from. Any

slight defect in the rollers, balls or races causes a disturbance in the movement controlled by the bearing. The oil film of a plain bearing, on the other hand, is able to smooth out small, unsystematic defects of this type, so that the accuracy of the movement obtained is better than that of the components of the bearing. For the very highest accuracy of rotation, therefore, the plain bearing comes back into consideration. A noteworthy bearing of this type is the plain bearing that was developed in Philips Research Laboratories in connection with the manufacture of the Schmidt optical correction plates for television projectors. This bearing was mounted in the headstock of a lathe that was used, with specially ground diamond tools, to turn methacrylate material to a mirror finish [10].

The hydrostatic bearing, which came into wide use in the sixties, combines to a certain extent the advantages of plain and ball bearings. Since the lubricating film is maintained by an external pump, and is therefore independent of the sliding-speed, this bearing is ideally suited for very slow movements such as those occurring in positioning the critical moment of stopping at the right position. For this reason hydrostatic bearings are very suitable for use in the guideways of high-precision machines for performing measurements and special machining operations. Good examples of these are the step-and-repeat cameras used for the reproduction of the photomasks used in microelectronics. Another application can be seen in the title photograph [11].

The hydrostatic bearing also has important advantages at the high end of the speed range. The high stiffness required at low speed does not have to be achieved by choosing a bearing clearance which is too small at the highest speed. Much of the bearing surface consists of recesses which are very deep compared with the normal clearance, thus considerably reducing the viscous friction. Furthermore the heat generated in the bearing is removed by the oil flow that circulates in the bearing and can be cooled externally.

A hydrostatic bearing has a very high stiffness, which can be raised to any required value by means of a control system incorporated in the oil circuit. With the membrane restrictors developed at Philips Research Laboratories the stiffness can be made infinite.

The advent in the thirties of hard metal necessitated the design of more stable machines and tools. Strangely enough, new cutting materials like hard metal and ceramics caused a decrease rather than an increase in the forces that arise during machining. There was a dramatic increase, however, in the power entering the machine via the drive motor, to be dissipated entirely in and around the machine. "Scholls Führer des Maschinisten", which was widely used on the Continent in the latter half of the 19th century, quotes the maximum powers that were needed for the heaviest machining in those days: "a heavy-duty lathe, which cuts large chips, uses up to $2\frac{1}{2}$ hp", and "it is reckoned that 2 to 3 hp is used in boring out cylinders for steam tools, blowers, etc.". Modern lathes comparable with the first machine would probably take some 30 or 40 hp.

Many of the problems which faced the designer of machine tools after the introduction of carbide tools relate to the control of the greatly increased flow of power through the machine. Part of this power may in certain circumstances be tapped off from the main flow and give rise to vibrations in the machine, tool or workpiece. These self-exited vibrations, known as "tool chatter", must of course be avoided. The limited success achieved in the early applications of hard metal and the poor results later obtained with the even more brittle cutting materials were largely due to vibrations of this kind.

Tool chatter had not been unknown with high-speed steel, and the experienced operator was usually able, after a certain amount of trial and error, to find the conditions for stable machining. Owing to the greater power available and the brittleness of the cutting material, the consequences of tool chatter when hard metal was used were almost invariably fatal to the tool before anything could be done.

In the early days the designer set about designing machines of higher rigidity in a rather intuitive fashion, by increasing the stiffness of all the vital parts of the machine. This usually increased the mass of these parts which meant that the effect of such measures on the dynamic stability of the machine was rather uncertain.

In the last decade the dynamic behaviour of machine tools has been the subject of intensive study (*fig. 5*). The vibrational behaviour of the machine in its many degrees of freedom has been calculated and measured, and the mechanism underlying the transfer of power from the machining process to the vibrating system has also been carefully investigated [12]. The knowledge thus acquired is difficult to apply in practice because the workpieces and tools form an integral part of the vibrating system yet are frequently changed for different ones. The problem is an urgent one, however, as the unpredictability of machining vibrations has proved to be one of the greatest problems in the introduction of numerical control.

Apart from tool chatter, forced vibrations or shocks

---

[10] L. M. Leblans, A high-precision lathe headstock, Philips tech. Rev. 19, 68-69, 1957/58.

[11] H. J. J. Kraakman and J. G. C. de Gast, A precision lathe with hydrostatic bearings and drive, Philips tech. Rev. 30, 117-133, 1969 (No. 5).

[12] J. Peters and P. Vanherck, Ein Kriterium für die dynamische Stabilität von Werkzeugmaschinen, Industrie-Anzeiger 85, 168-174 and 342-346, 1963.

can also upset the process of metal removal. The shocks arising from the gear teeth in the transmission as they engage may completely ruin a precision-turned surface if they penetrate to the jaws of a chuck. For this reason the designer of a precision lathe will attempt to interpose a transmission element of low stiffness between the gearbox and the main spindle. In the renowned Kärger lathe (*fig. 6*), built in the thirties, this was done by connecting an endless flat belt between the gearbox and the spindle, so that the gearwheels and workpiece were entirely isolated from each other, at least at the spindle speeds used for finish-turning. The machine had such a good reputation as a precision lathe that its disappearance as a result of the post-war situation in Germany was widely regretted.

The low torque that can be transmitted by a flat belt under acceptable tension makes it less suitable for incorporation in the transmission behind the gearbox. Later successful lathes like the AI-DR 1 and the Cazeneuve lathe mentioned earlier (*fig. 7*), dating from the fifties and sixties, used V belts, with the same basic system as the Kärger lathe, for transmitting the power from the gearbox to the main spindle.

All the power supplied to the machine by the motor is ultimately converted into heat. Because of the friction losses in belts, bearings and gearwheels sufficient heat is developed to bring these parts some tens of degrees Celsius above the ambient temperature. The power used up in the actual machining is also converted into heat, and is distributed among the workpieces, the tool and the chip. A substantial part of this heat is transferred to the machine by conduction. The quite appreciable thermal deformations of the machine which this can give rise to received surprisingly little attention for many years. This was because the operator was well capable of making the necessary corrections to the machine setting by regularly measuring his workpieces. In numerically controlled machine tools workpiece

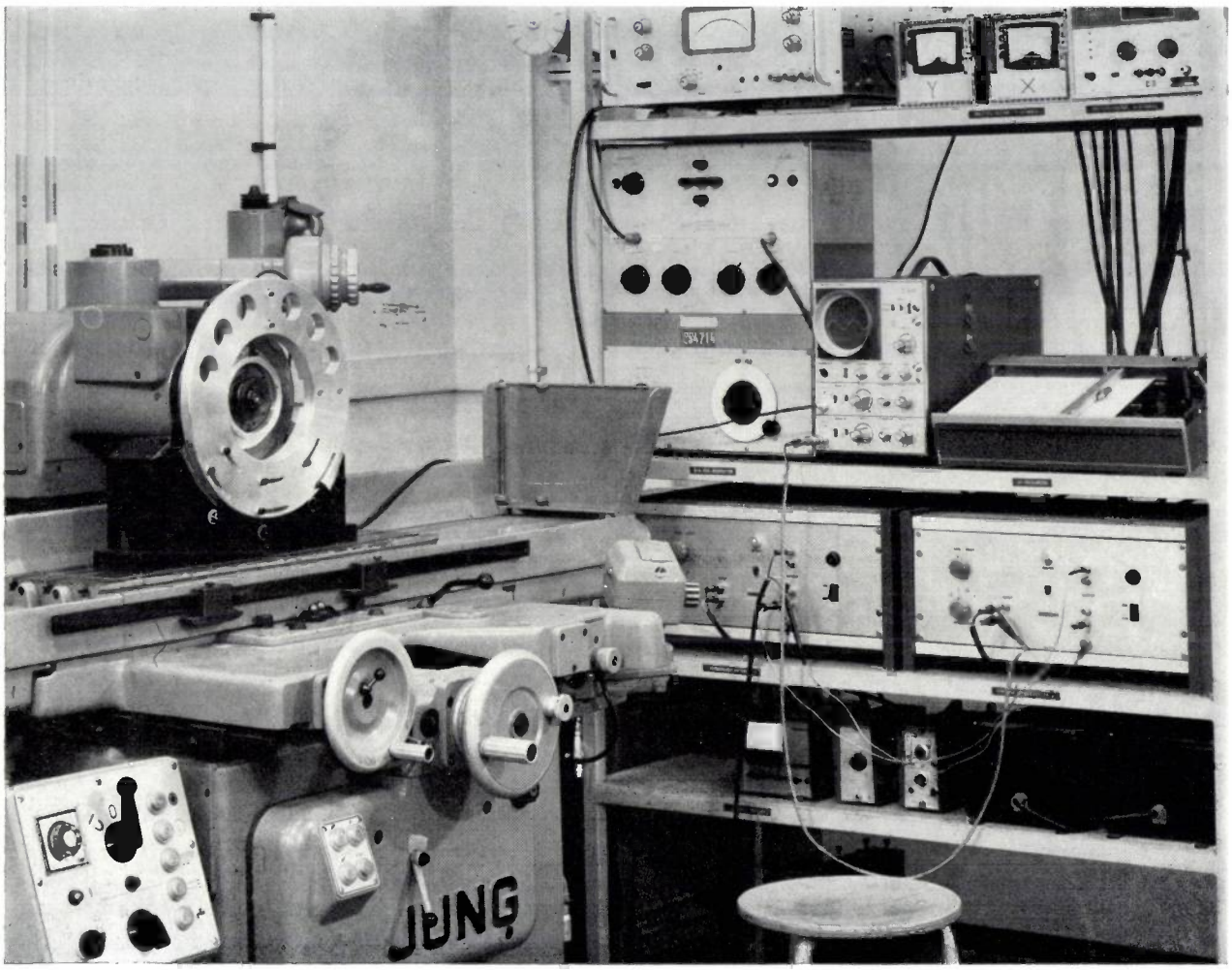


Fig. 5. Investigation of vibration on a surface grinder in the Philips Centre for Manufacturing Techniques. The machine is made to vibrate by electrodynamic transducers.

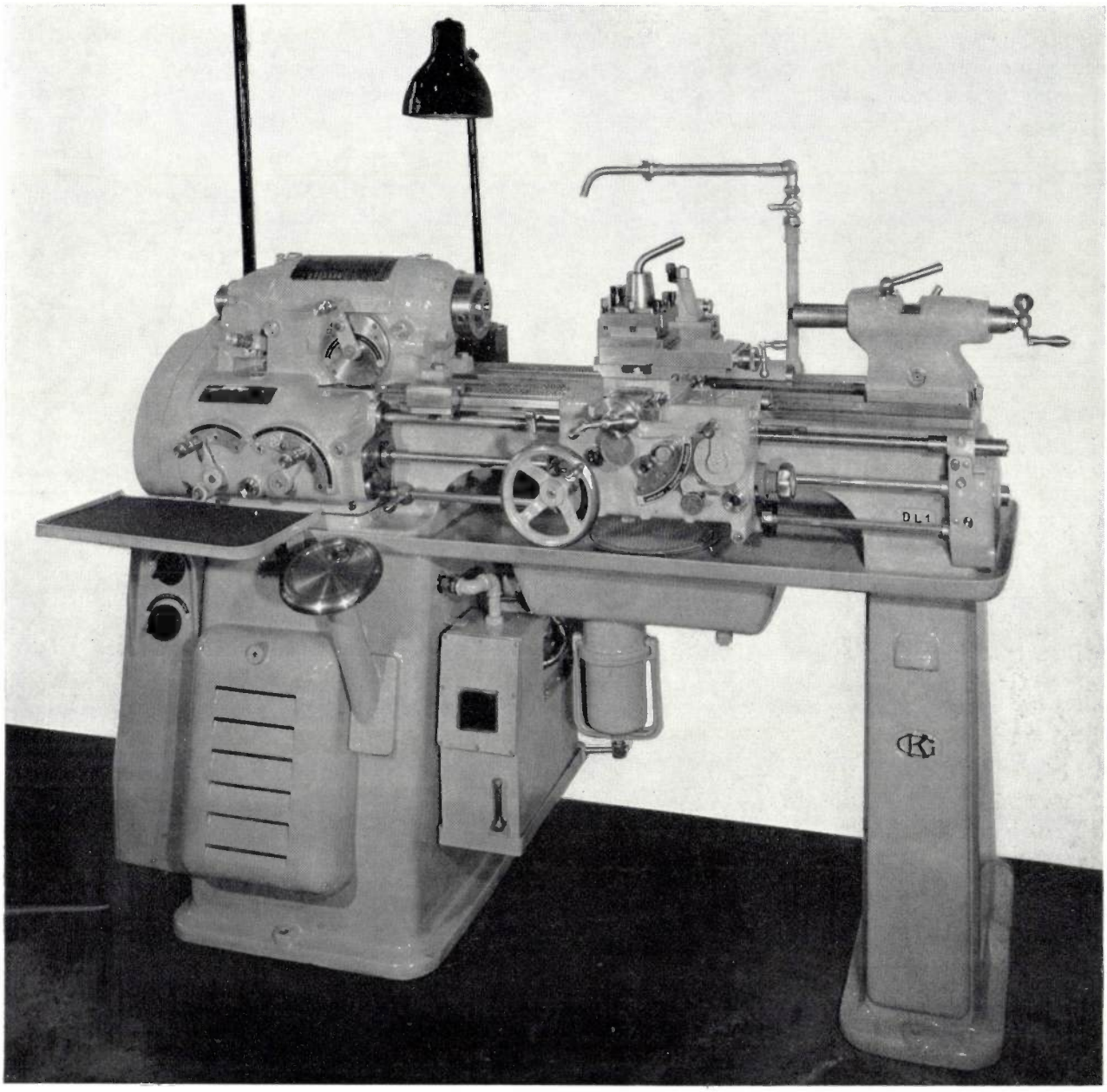


Fig. 6. Kärger DL 1 lathe. This lathe had a very good reputation before the Second World War as a high-precision machine.

dimensions are checked by sensors mounted on the machine slideways. In this case thermal deformation of the machine affects the resulting workpiece dimensions and presents an undesirable factor in machining accuracy. Towards the end of the sixties there was therefore a considerable growth of interest in thermal deformation in machine tools.

The magnitude of the power flow that appears as heat in and around the machine makes it seem unlikely that troublesome heat sources will ever be completely avoided. A more likely solution is an adaptive-control method in which temperatures are measured and suitable corrections are derived from these measurements.

Forty years of shop metrology

The manufacture of interchangeable components, whose origin is to be found at the very beginning of the nineteenth century in Eli Whitney's arms factory in Massachusetts, U.S.A., was crowned in the thirties of the twentieth century by the work of ISA (International Standards Association, later ISO, International Standardization Organization) recommending a system of fits for international use.

The interchangeability of components is an absolute necessity in mass production, and it was therefore the automobile industry that gave the great impetus for its penetration into ordinary non-military production. The

year 1908 was a great year for two reasons. This was the year in which Henry Ford introduced his Model T, of which 15 million were to be manufactured; and it was also the year in which Henry M. Leland had three Cadillacs completely dismantled at the Brooklands circuit of the British Royal Automobile Club, and their components shuffled like a pack of cards. After 89 of these parts had been replaced by randomly chosen spares, the three Cadillacs were put together again, and then proceeded to complete a faultless trial run of 500 miles.

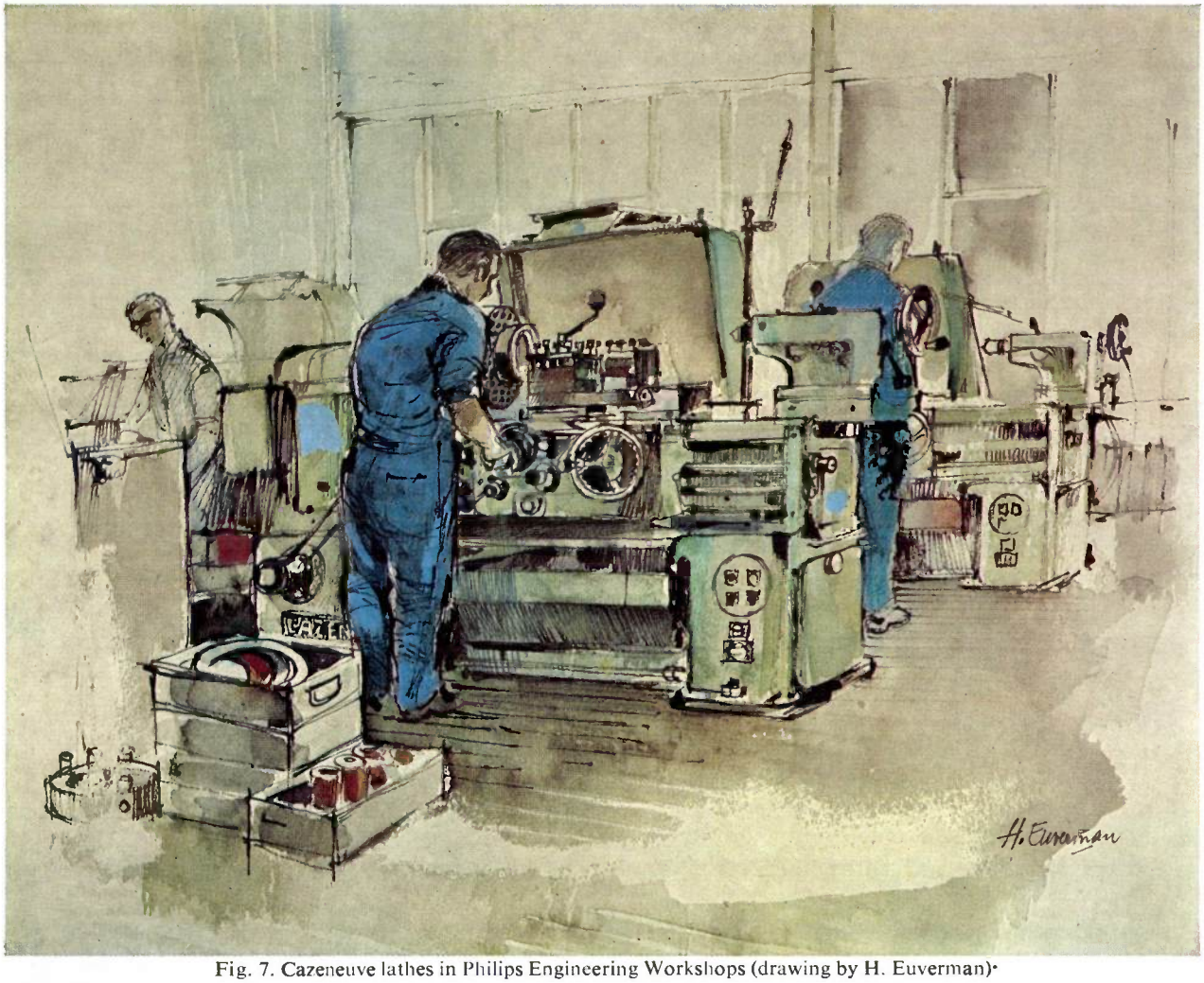


Fig. 7. Cazeneuve lathes in Philips Engineering Workshops (drawing by H. Euverman).

In spite of all this, manufacture still remained strongly dependent on components "made to measure". In the typical workshop in the thirties the only instrument available for checking dimensional accuracy was the gap gauge.

Gauges are excellent aids in inspection work, and the term "Go and Not Go" gives a good indication of what they were useful for; they cannot, however, be called proper measuring instruments. The operator at his machine continued to rely on all the old skills and tricks developed in the shop for producing the required fit for mating parts. By mating all the holes of a batch with one and the same plug gauge and all the shafts with a ring gauge a reasonably good interchangeability was ensured. Since, however, the gap gauge did not in fact provide the operator with much more information than whether the machining operation was a success or a failure, many a high-precision piece of work had to be made a number of times before it could be "passed" by the gap gauge. This situation obviously imposed a cautious and scarcely productive style of working on the operator when he was producing workpieces requiring high dimensional accuracy. Not until the fifties did instruments of a truly scientific type begin to find their way into the workshop. Only instruments giving an output signal are capable of providing the quantitative dimensional information that can be fed back to control the operation.

There were, of course, good reasons for this state of

affairs. For reliable measurements the accuracy of the measuring instruments should be an order of magnitude better than the dimensional accuracy required of the workpieces. However, instruments as accurate as this are all too easily damaged or put out of adjustment. If instruments in a workshop give incorrect readings, the results can be disastrous.

A cylindrical shaft or hole can be out of shape in many ways ("errors of form"). It is difficult to interpret a measurement made on such a workpiece. Moreover, the relation between the dimensioning of mating parts and the functional quality of their fit is not so straightforward as might be suggested by a system of fits and limits. The combination of materials, the geometrical configuration of the mating surfaces and the character of errors of form that may be present all have their effect on the quality of the fit. Plug and ring gauges used for inspection simulate the constructive function of a fit, and, provided "Taylor's principle" is observed, proper allowance is automatically made for errors of shape.

In the fifties Solex, the carburettor and moped manufacturers, used air for measuring the apertures of carburettor jets. By passing the air through two constrictions connected in series, one of which is known and the other unknown, the size of the unknown constriction can be determined with great accuracy. In pneumatic gauging the unknown constriction consists of an aperture that approaches the surface of a workpiece more or less closely. The measuring method combines very high sensitivity with low vulnerability. In addition to air gauging, many other mechanical and electronic measuring devices were brought out in the fifties and sixties, all of which have contributed to the development of workshop metrology.

At least as important as the measuring devices, however, is the knowledge needed to use them. This knowledge applies not only to the handling of expensive and sensitive instruments, but above all to the choice of the right instrument and of the right method for the measurement to be performed (*fig. 8*). A good solution to this problem is to have a measurement specialist in the workshop, who is responsible for issuing the instruments and at the same time for giving the instructions for setting up an appropriate measuring arrangement.

Metrology and quality control are always closely interrelated in mechanical engineering. In the development outlined here, in which measuring techniques have progressed from inspection to an aid in the control of manufacturing processes, it is appropriate to have a method of ensuring quality which is directed more towards control than to inspection.

In shops employing skilled labour it has always been good practice to make the operator responsible for the quality of the work he produces. In this connection the quality-control department must take responsibility for the means, methods and organizational procedures needed for effective inspection, leaving the actual inspection of parts to the shop personnel.

The information which still has to be obtained by the quality inspectors from the traditional final inspections is used for supervising and controlling the inspection procedures of the production departments. The information gathered should be used for tracking down

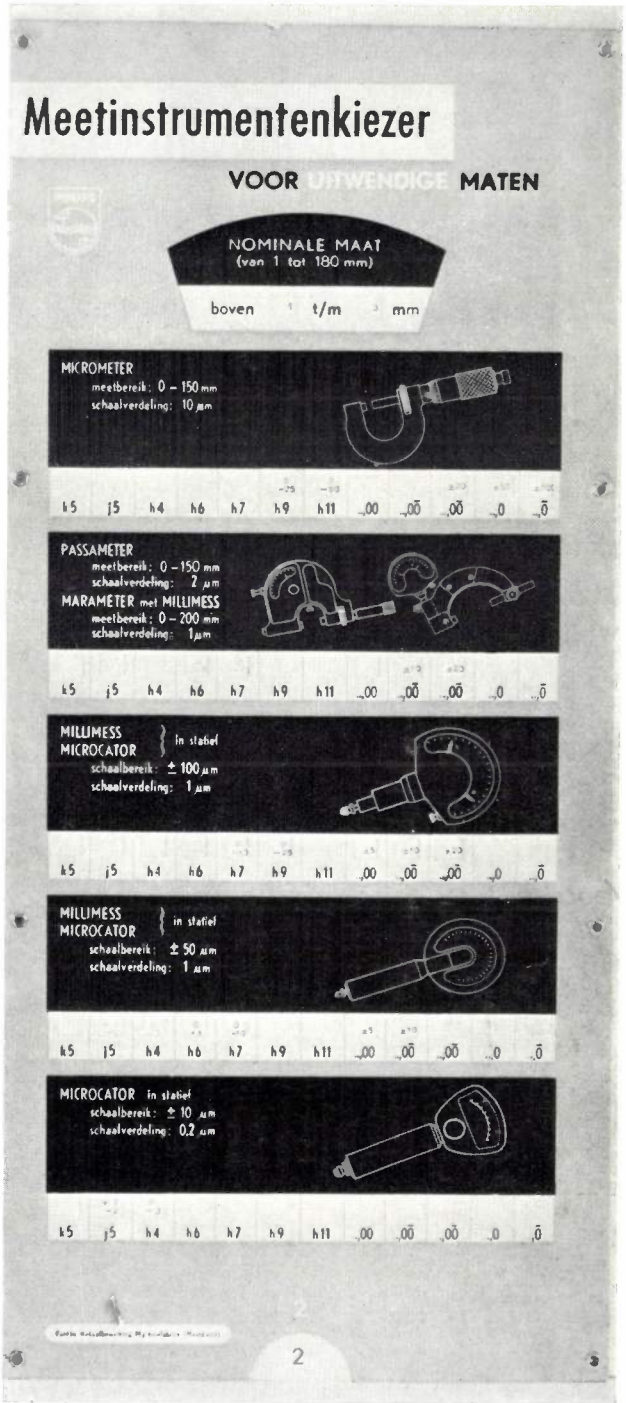


Fig. 8. Measuring-instrument selector in the form a of slide. The windows are located above the horizontal white strips.

critical points in manufacture to determine whether the working methods and measuring instruments used at those critical points are capable of guaranteeing the required quality.

The most outstanding and fundamental development in linear measuring techniques in the last forty years was the advent of interferometric methods of measurement. The platinum metre of the International Bureau of Weights and Measures was replaced as a standard of length by the wavelength of a particular line in the spectrum of crypton 86.

The potential of interferometry was enormously widened by the invention of the laser. The extremely intense monochromatic and strictly coherent light emitted by the laser made interferometric measurements possible over distances with light-path differences of tens of metres (*fig. 9*). In such measurements enormous numbers of interference fringes have to be counted, and it is only through the speed and reliability of electronics that this is practical at all.

The laser interferometer of fig. 9 also gives a good illustration of another important recent development in metrology. Numerical control of machine tools and the related development of measuring instruments with digital read-out have created a need for displacement-measuring devices that often have to satisfy very difficult requirements for accuracy of the absolute value at total displacements that may be quite large. Most of these devices are based on a linear scale provided with a grating whose displacement can be read off optically or by an electronic system. From its very nature such a measuring system cannot be more accurate than the grating it is based on, and the making of accurate gratings is becoming an important exercise in metrology. The laser interferometer is a particularly useful device in this activity of making accurate gratings.

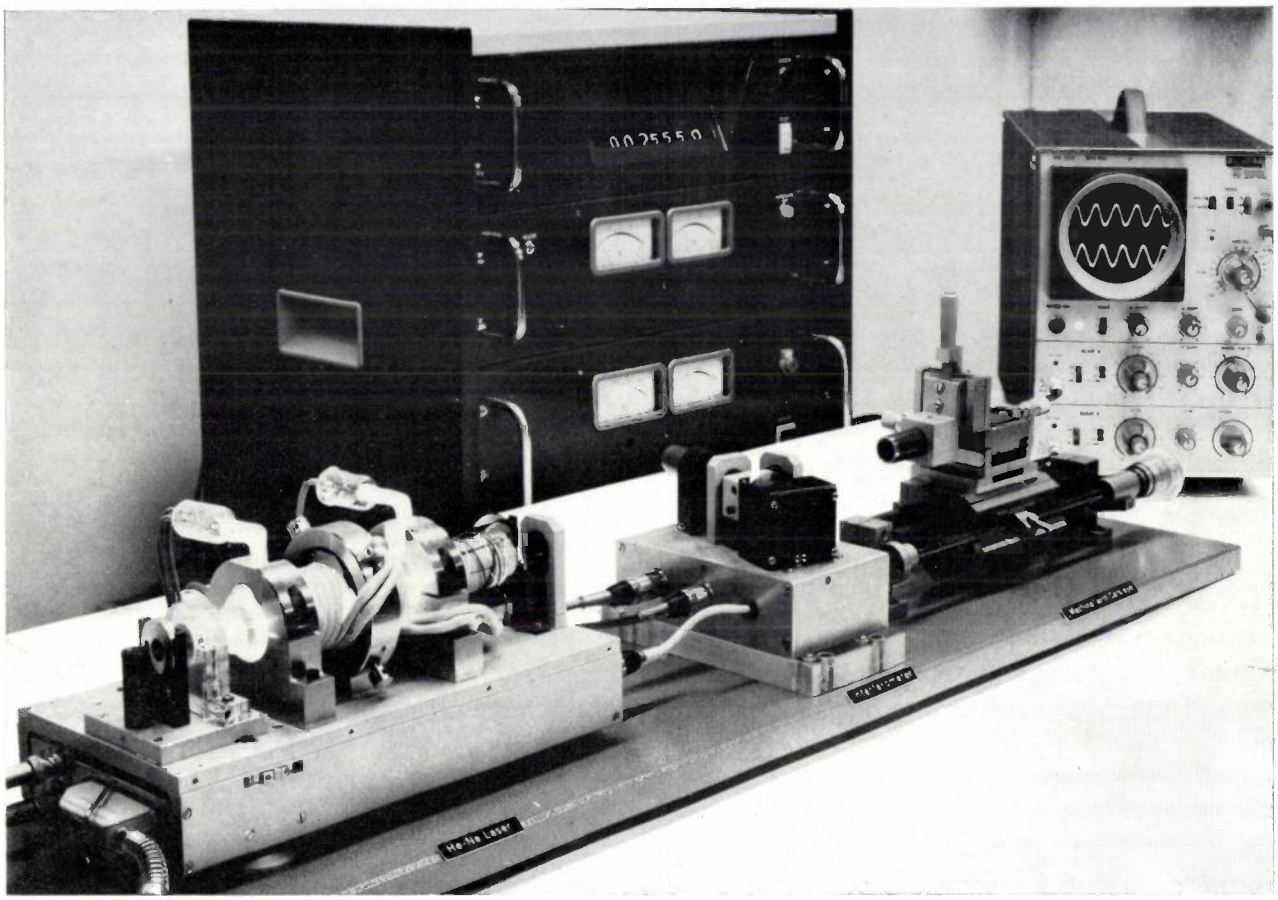


Fig. 9. Arrangement for measuring displacements by means of the interference of laser light in a modified Michelson interferometer [13]. On the left: a stabilized helium-neon laser. Three of the four interferometer mirrors are mounted on the carriage in the centre, the fourth on the workpiece on the right. The counter counts the number of quarter wavelengths over which the workpiece is displaced.

At Philips Research Laboratories and in the Industrial Applications Division ("PIT") measuring systems based on optical gratings have been developed. The PIT system has an accuracy that is suited to the more usual machine tools, and is particularly insensitive to contamination. The system developed at the Research Laboratories, based on a grating [14], is extremely accurate and is therefore suitable for applica-

tions in instruments and in machine tools of exceptionally high accuracy. The digital unit is 0.5 μm; by means of analogous interpolation and with the help of an analogue-digital converter one can easily have an accuracy which is ten times better. The super-precision lathe in the title photograph makes use of displacement-measuring devices of this type.

The laser beam has also enabled the shape of a workpiece to be recorded in a hologram. Since the familiar interference patterns can be made by putting together a series of holograms, it is possible to make extremely accurate comparisons of workpieces before and after small changes of shape.

Laser interferometry is so young that its range of applications has probably not yet been fully discovered and applied.

Forty years of production with sheet metal

In sheet-metal work there is no clear-cut distinction between parts manufacture and assembly. Indeed, in production methods based on sheet metal the actual joining together of the different parts is usually the most important operation. This is why jointing techniques such as riveting, welding and, in special cases, bonding, have been so much in the forefront of developments in the past forty years. In heavier constructions hot-riveting has been almost totally superseded by electric-arc welding.

In fine sheet-metal work resistance welding has been extensively developed and in aircraft engineering, where the special properties of duraluminium sheet covered with a layer of pure aluminium rules out the use of thermal jointing methods, metal bonding is becoming steadily more important.

In the thirties electric-arc welding was regarded as a jointing method that could not be used where strength and reliability were of prime importance. Boilers and ships, for example, still had to be riveted for the sake of safety, and the construction of a welded bridge had already proved a failure in at least one case.

The coating of welding electrodes in the thirties protected the transferred metal droplets by a slag formed from the coating and also by means of gas that contained a great deal of hydrogen because of the reaction between the iron and the water vapour. In certain cases this gave rise to serious porosity in the weld, and where this was not the case the hydrogen dissolved in the metal often gave rise to microcracks under the bead of the weld or next to it. This occurred particularly in carbon steel, which can acquire a martensite type of structure at these locations [15].

Virtually the only protection provided for the melting metal by the coating of the electrode, which had made welding with an electric arc possible, was against the nitrogen in the air. Efforts to reduce the effect of oxygen by the addition of reducing agents to the coating led to a problem with hydrogen as long as the gas from the coating still contained a great deal of water vapour [16]. Only by welding under a powder shield, where the slag protected the welding metal from contact with gases, was it possible to produce welds that did not have dangerous weaknesses due to the presence of hydrogen in the arc plasma.

It was not until the fifties that an electrode coating was brought out that produced a gas shield free from water vapour and could give completely reliable electric welds.

Another important development, which did not emerge until after the Second World War, is the use of inert protective gases such as helium and argon. Argon-arc (TIG) welding yielded excellent results in the welding of thin aluminium sheet and stainless steel. Unfortunately the method was not suitable for the most important of all structural materials, ordinary mild steel.

The "low-hydrogen" electrode, whose gas shield consisted of a mixture of CO and CO_2 after elimination of all components containing hydrogen, prompted the use of a corresponding gas mixture for welding with a gas shield. Obviously, there could be no question of using large quantities of poisonous CO gas in the workshop, and welds were therefore made in pure CO_2 gas (*fig. 10*). The non-melting tungsten electrode could not be used with this protective gas. The electrode had to be a consumable wire, which also contains the deoxidizing agents needed to compensate for the oxidizing action of the CO_2. With the CO_2 process it was readily possible to weld mild steel [17].

The ideal protection for a weld is obtained by completely eliminating the gas atmosphere around the joint to be welded. Modern reactive metals like titanium and zirconium can only be welded in an extremely pure inert gas or in a vacuum. A vacuum represents the easiest way of limiting the concentration of unwanted gases to an extremely low value. The plasma arc cannot serve as a heating source *in vacuo* and the heating has

---

[13] H. de Lang and G. Bouwhuis, Accurate digital measurement of displacements by optical means, II. Displacement measurement with a laser interferometer, Philips tech. Rev. **30**, 160-165, 1969 (No. 6/7).

[14] H. de Lang, E. T. Ferguson and G. C. M. Schoenaker, Accurate digital measurement of displacements by optical means, I. Displacement measurement with a reflection phase grating, Philips tech. Rev. **30**, 149-160, 1969 (No. 6/7).

[15] J. D. Fast, Causes of porosity in welds, Philips tech. Rev. **11**, 101-110, 1949/50.

[16] P. C. van der Willigen, Booglasmethodes voor staal en de rol die de waterstof daarbij speelt, Chem. Weekblad **57**, 170-176, 1961.

[17] P. C. van der Willigen, Some modern methods of arc-welding steel, Philips tech. Rev. **24**, 14-26, 1962/63.

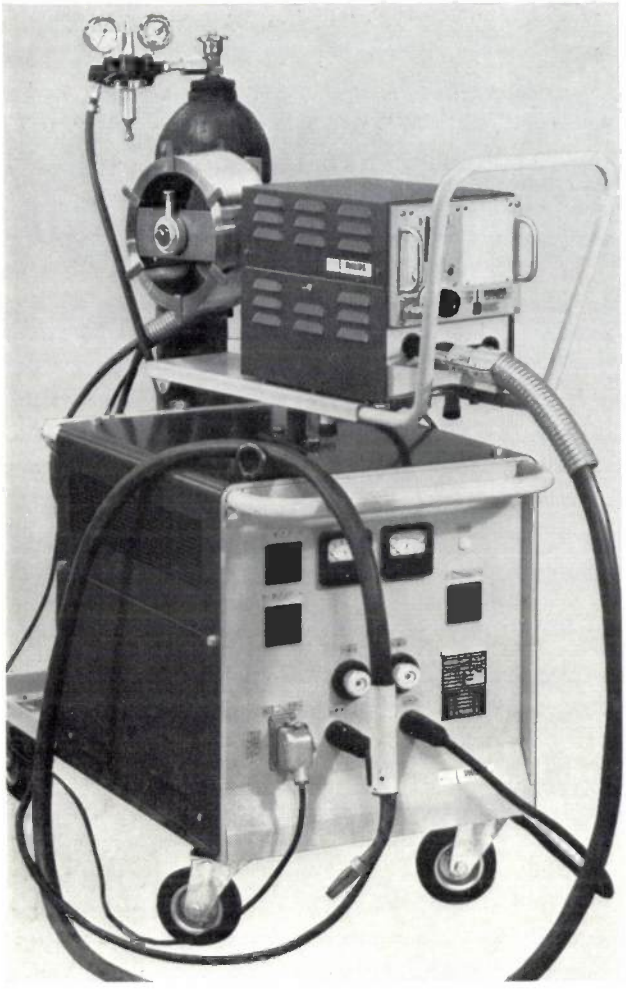


Fig. 10. Philips CO_2 welding unit.

to be done by the beam of fast electrons. The electron beam also has certain other very attractive features for welding.

If we consider the development of fusion welding it is noticeable that there has been a continuous increase in one particular quantity, which is the power supplied per unit area, referred to below as the *power density*. The oxy-acetylene flame is the only open flame hot enough to melt low-carbon iron, but it has little heat in reserve and the power density of the heat flow to the steel heated to the melting point is not very high. The plasma of the electric arc has much more to offer, and in the development of electric-arc welding we see that the power density is gradually increasing. The welding methods using a shielding gas permit a more intensive heat flow than those using coated electrodes, and the further development of argon-arc welding is in the direction of "constricted-arc" or plasma welding. Various means are used to concentrate the arc plasma more intensely on the part to be heated in order to increase still further the power density of the heat source. In this respect the beam of fast electrons is a virtually ideal heat source. The power density is extremely high and the beam can be readily focused and controlled by electronic means. Only the laser exceeds it in power density.

In welding metalwork together, heat is a necessary evil. When a desired shape is obtained by assembling various separate parts together, the relative location of the parts must be very accurately held. Distortion of the assembly due to heat from the weld must as far as possible be avoided. The method that gives the absolute minimum of such "thermal contamination", coupled with the absolute minimum of chemical contamination, is electron-beam welding.

Another welding method that is noted for giving little warping due to heating is resistance welding. Here again, the good performance is achieved because of the very high power density of the energy supplied. An added advantage of resistance or spot welding is that the spot-welding electrodes also act as a clamp which keeps the parts to be joined in the correct relative locations during welding. The welding rate, compared with the fusion method, is extremely high, so that this method of welding is economically very attractive.

Against all this is the virtual impossibility of inspecting the joint, hidden between the plates, other than by destroying it. Because of this spot welding was regarded in the thirties as a cheap and not very reliable jointing method, which could in any case only be used on the easily welded mild steel.

Spot welding, more especially since the war, has developed in the direction of greater reproducibility of the process to produce a more reliable weld. This can largely be attributed to the use of electronic devices, which control the primary current in the welding transformer by means of thyratrons and ignitrons. The same equipment controlled the electrode pressure synchronously with the power supply, where necessary in accordance with a complex pressure-time programme. Accurate switching of the applied power in synchronism with the mains enabled the welding time to be markedly reduced, with a corresponding inversely proportional increase in the power. The resultant steep increase in power density meant that metals like aluminium and aluminium alloys, which are difficult to weld owing to their high thermal and electrical conductivities, could now be joined by spot welding.

The control of the spot-welding process has developed into what is called "predictive control". The process parameters are controlled with such high precision that the result can be reliably predicted from experience with welds previously produced in an identical way. In spot welding, again because of the inaccessibility of the welded joint, no generally applicable method of "adaptive control" has yet been found.

Integral construction with sheet metal

Although the basic idea of integral construction with sheet metal is not new, this method of construction gained rapid ground in the thirties. The majority of large objects made from sheet metal have two functions: to "carry" their contents and to "enclose" them. In older methods of construction the carrying function was fulfilled by a chassis or framework of ribs and trusses, etc., covered with some material for the enclosing function. The two functions can be integrated if the shell is designed in such a way that it carries the load.

Notable examples of integral construction in the thirties were the Douglas DC 2 built in 1934 and the bodywork of the Citroën car built in 1937. Each in its own way ushered in a new era in aircraft engineering and in automobile manufacture.

Integral construction also made headway in less spectacular fields of sheet-metal applications, such as the building of cabinets, panels and control desks for electronic and engineering products and medical equipment. The availability of wide strip, which was superior in surface quality to the separately rolled sheets used in the past for covering purposes, and which moreover could be bent easily, gave a considerable boost to the use of integral construction, in which the steel framework is eliminated.

There are undoubted advantages in building up integral-construction sheet-metal designs from sheet-metal parts that owe their inherent stiffness to compound curvature. Such parts can only be manufactured by deep-drawing methods that require special tools and expensive presses. This is no problem in the automobile industry, with the enormous quantity it produces, but in the aircraft industry it does present difficulties. Since sheet-metal parts with compound curvature are indispensable in aerodynamic design, this branch of industry struck out along other lines.

The methods in which only the most easily manufactured press tool is made, the male part, and in which the female die is replaced by a liquid or by rubber, were developed by the aircraft industry but have also found important applications outside it [18] [19]. With these methods, sometimes known as "hydroforming", "rubber-die pressing" etc., a single draw can produce shapes that would require several draws with the traditional steel dies.

A drawback of the hydroforming method is that it requires a large and costly press. Explosive forming makes such a press completely superfluous; all that is needed here is half a tool, but here it is the female part. The role of the punch is taken over by a shock wave generated by means of a chemical explosion or electrical discharge in water, imparting high velocity to the blank. The shaping of the metal is effected by the kinetic energy transferred to the sheet metal by the shock wave.

Both methods have so far been used only on a limited scale and usually confined to specialized applications. In the sixties the fashion in design moved towards austere shapes with sharp edges, which could readily be made by bending operations. The danger and the noise involved by the use of explosives — or of their equally dangerous and noisy electrical counterpart — have tended to discourage the wider application of explosive forming.

Though less spectacular, the development of the traditional methods of cutting and shaping sheet metal — shearing, punching and bending — has been no less significant. A substantial improvement in accuracy and also in mechanization and rationalization, coupled with the improvements mentioned earlier in basic materials and jointing techniques, have raised sheet-metal work to a higher standard and in many cases made it an interesting alternative to other, more traditional methods of construction. Punching, for long a particularly efficient operation in mass production, also came into its own in batch production with the application of jig punching machines with facilities for the rapid exchange of tools, in some cases effected by a tool turret [20].

The transformation of industry by electrical and electronic methods has of course given great impetus to the development of the small-batch production of the sheet-metal cabinets that are used to support and enclose electrical or electronic equipment.

Toolmaking

In the years between the First World War and 1930 the growth of the radio industry caused the demand for punches and dies to increase by leaps and bounds. In the Netherlands, German toolmakers had to be brought in to help meet this demand. The German master toolmakers were very skilled in their craft, which consisted in a number of difficult manual operations demanding a high degree of mechanical knowledge, but which in particular called for a great deal of patience and devotion to their work. The productivity of these methods of working was low, but as toolmaking represents no more than a few per cent of the cost price of the end product, this was not so important. Since it determined the quality of the product, the quality of the tool almost completely overruled considerations of cost.

[18] H. Hermans and F. R. H. F. Vermeulen, Metaalvervorming met behulp van rubber, I, II, III, IV, Metaalbewerking **25**, 299-303, 320-324, 335-338, 361-366, 1959/60.

[19] W. van der Meulen, De technische mogelijkheden van hydroform, I, II, Metaalbewerking **29**, 232-235, 239, 253-258, 1963/64.

[20] J. W. Stemerdink, Toepassingen van revolverkopieerpersen, Metaalbewerking **31**, 21-28, 1965/66.

In the Netherlands the greatly increased demand for tools to make thermionic valves and radios was originally met by employing German toolmakers, and soon afterwards by training carefully selected young apprentices in the difficult skill of toolmaking. At the same time, however, a new trend towards mechanization began, and in the last forty years the most difficult manual operations of the old toolmakers' craft have steadily been superseded by faster and more foolproof operations based on the use of machine tools [21].

Making holes at accurately defined positions in the various plates that make up a press-tool punch and die set was one of the principal operations in which the master toolmaker excelled. For this purpose he not only had a set of slip gauges — in those days they would be his own property — but also jigs and fixtures and working methods which he often regarded as his "trade secret".

The jig borer, which was pioneered by Moore in the United States and by the Société Genevoise des Instruments de Physique (SIP) in Europe, began in the thirties to take the place of traditional manual methods for the positioning of holes (*fig. 11*). The young electronics industry, with its great demand for tools, was in the forefront of this development.

The surface grinder was the second machine that very gradually superseded a manual operation. The magnetic chuck, with its reference planes finish-ground extremely accurately on the machine itself, was ideally suited to manufacturing profile punches with the same accuracy as those made by hand or even greater. Numerous aids were developed for the surface grinder such as "sine" tables, radius-dressing and pantographic equipment, which greatly extended the capabilities of this machine (*fig. 12*).

If the same form-grinding method is to be used for making apertures in the die plate, the parts of the die plate must be split up into sections to allow the grinding wheel access to the internal surfaces of the apertures to be ground. The fifties in particular saw a marked development of this method for the manufacture of hard-metal punch and die sets. These tools are very costly, but they can make tens or even hundreds of millions of products before the tool finally has to be discarded.

For steel die sets the meticulous craftsmanship of the old-style toolmaker has long remained the most efficient method of making shaped holes in the various plates of a punch and die set. As a result of rising wages and the growing scarcity of highly skilled toolmakers, it looks as if these old craft methods are also going to be superseded by a mechanical technique, spark machining.

Apart from the manufacture of punches for metal

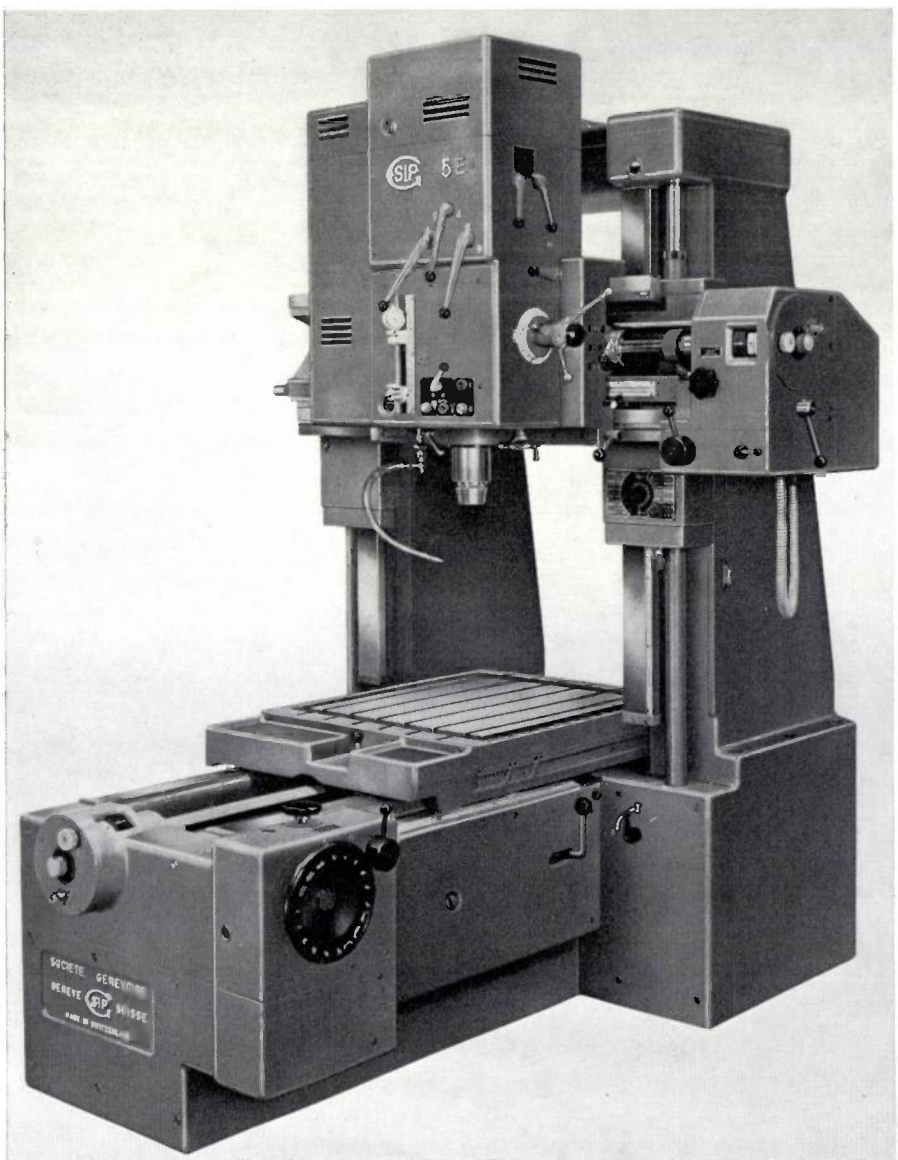


Fig. 11. Jig borer made by Société Genevoise des Instruments de Physique (SIP).

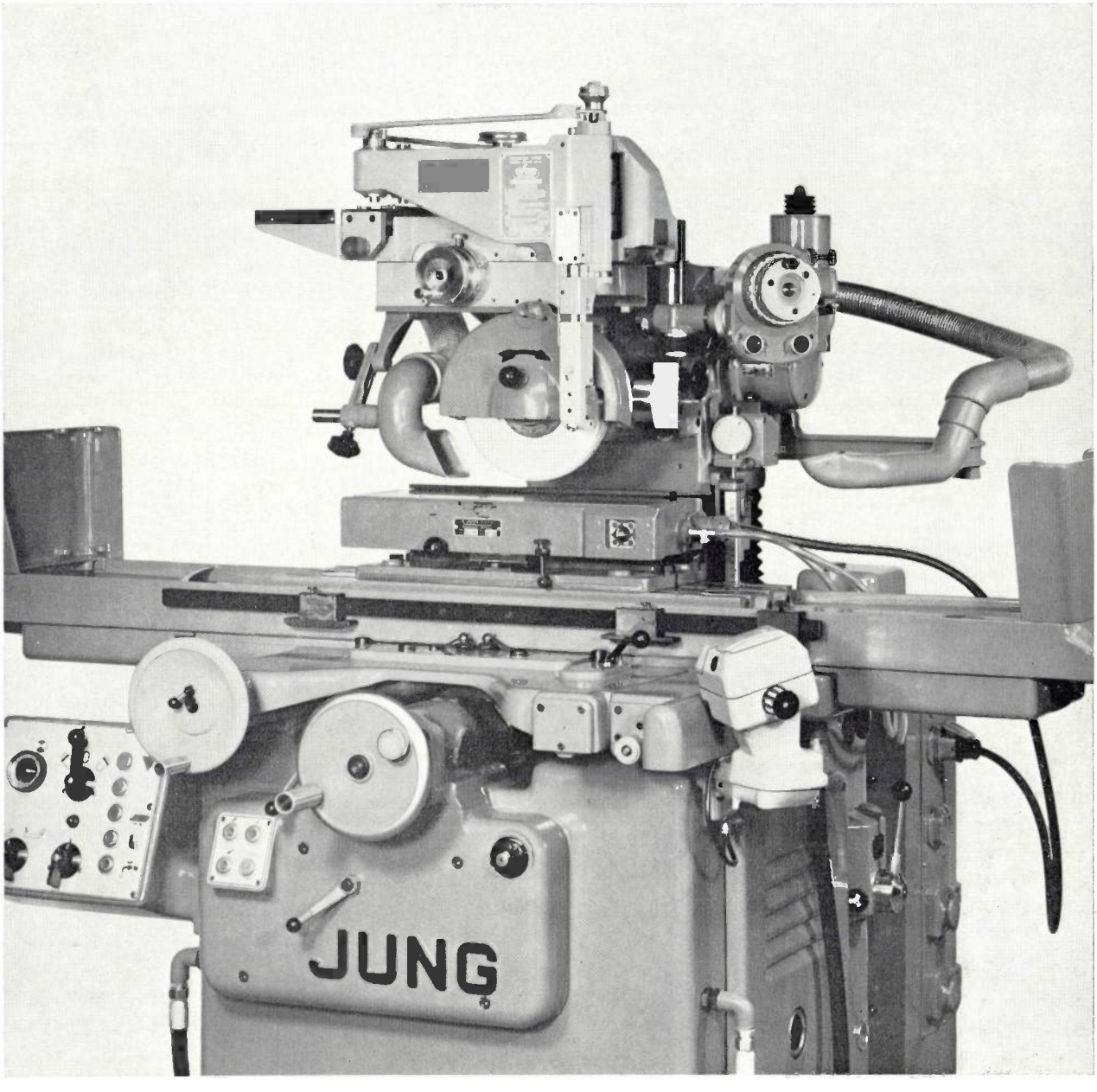


Fig. 12. Jung surface grinder with Diaform profiling system. The grinding disc is ground to shape with a diamond, whose movement is controlled by a jig via a pantograph.

parts, the radio industry also played a major part in promoting the use of plastic products manufactured in moulds. In the thirties the thermosetting plastic Bakelite, invented by L. H. Baekeland in 1909, and various associated condensation polymers were the synthetic materials most widely used. In the years after the war there was a gradual increase in the use of thermoplastic materials, and many new types were introduced. The traditional methods of compression moulding, like those used with plastics between the two World Wars, were largely superseded by the injection moulding of thermoplastics, but the thermosetting materials themselves still had much to offer. New thermosetting plastics with interesting properties such as polyesters and epoxy resins were developed, but a more important development was the adaptation of the properties of thermosetting plastics to the injection-moulding technique. The transfer mould was also further developed, so that there was no reason for the forming of thermosetting materials to lag behind that of the thermoplastics.

Injection moulding itself made considerable advances in the years after the war. Improvements in the injection-moulding machine, such as the introduction of screw plasticizing, programmed control, etc., made the

[21] S. Wiegersma, De ontwikkelingsgang van de vervaardigingsmethoden van snij-, buig- en trekgereedschappen, I. II, III, Metaalbewerking **22**, 276-281, 295-300, 319-321, 1956/57.

greatest contributions, but advances were also made in the design of the mould. The introduction of "automatic" three- and two-pot moulds which eject products and waste separately, and of "hot-runner" moulds which produce no wastage at all, are the most important recent developments.

The making of a mould cavity by a machining operation almost invariably involves difficult problems of access. The main difficulties do not arise in the metal-cutting process itself, but in the subsequent polishing required to remove the tool marks left after the previous operation. Visually, an object looks "right" with its projecting parts better finished than less accessible and consequently less visible parts. But since the toolmaker has to finish the product "in the negative", he naturally achieves the reverse; the protruding parts of the product are of course the ones that lie deep in the die or mould and are therefore difficult or impossible to finish completely. For the manufacture of dies and moulds a number of interesting "reversal methods" have therefore been developed, which can be used to form the appropriate cavity from a positive model of the product.

Casting, the most obvious reversal method, is used surprisingly little in the toolmaking shop. Unless the whole casting process can take place at room temperature, it is simply not accurate enough for the purposes of toolmaking. Moreover, cast metals are as a rule too impure and too inhomogeneous to be used as materials for making dies or moulds. Press moulds for glass are an exception to this, but even in this case a special casting technique is used to ensure that the layer of metal at the wall of the cavity is of high purity. Casting with the aid of non-shrinking materials on a ceramic or plastic base is a familiar method of making copying models. However, since these materials are not suitable for making the die itself, they do not solve the central problem of finishing a shape "in the negative".

Hobbing is a reversal technique which is particularly suitable for making multiple cavities in pressing and injection moulds. In the thirties this technique could only be carried out with pure iron. With the means then available this ductile and not particularly strong material was the only one in which a hardened steel punch could be sunk cold to the required depth. A drawback with this method was that the resultant cavity had to be hardened in water after carbonizing.

After the war news came from America about the development of a new type of steel that could be hobbed and subsequently hardened in air. Systematic investigations into the influence of alloying elements on the strength and "hardenability" (critical cooling speed) of this new material had shown the strongly ferrite-stabilizing elements chromium and molybdenum to be the best alloying elements and made it clear that the austenite-stabilizing elements, such as carbon, nitrogen and manganese, had to be kept as low as possible.

Meanwhile in Germany new, highly stable hobbing presses had been brought out that permitted the use of harder hobs. Towards the end of the fifties this hobbing technique had reached its peak. In the sixties a new development in the technology of spark-machining was to supersede hobbing to a great extent.

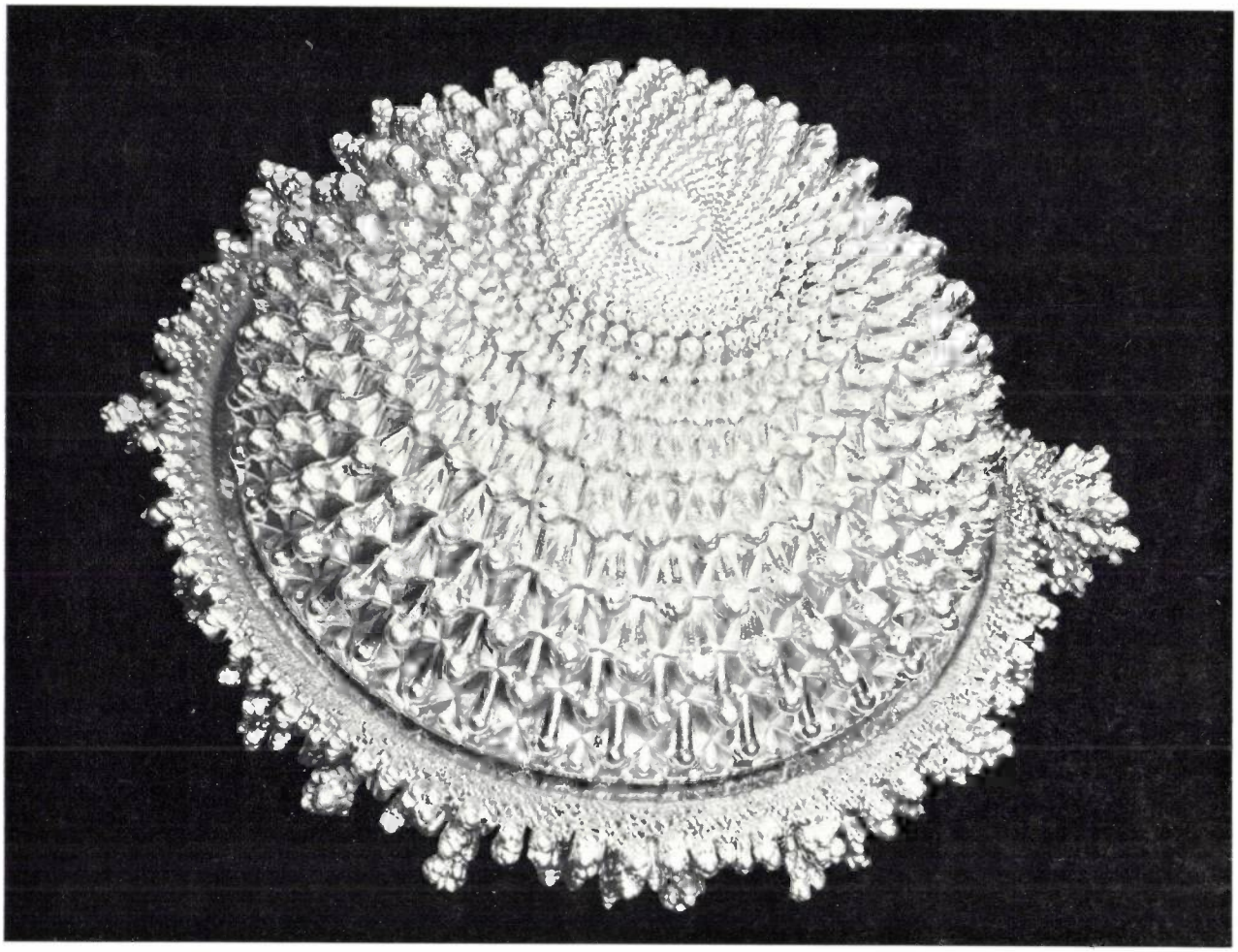
A very interesting reversal technique dating from before the war is electroforming. By the electrical deposition of a thick non-adhering layer of metal, usually hard nickel, on a positive metal mould, a shell-shaped negative is made that can be built into a die as a moulding cavity. The use of electroforming for the manufacture of gramophone-record moulds and surface-roughness samples indicates the great value of this method, in which the extremely pure and homogeneous electrolytic metal deposit enables the finest details of the model to be faithfully reproduced (*fig. 13*). Although in recent years spark-machining or erosion has seemed to be in the process of superseding all other reversal techniques in mould and die manufacture, it will never be able to take the place of electroforming in reproducing fine detail.

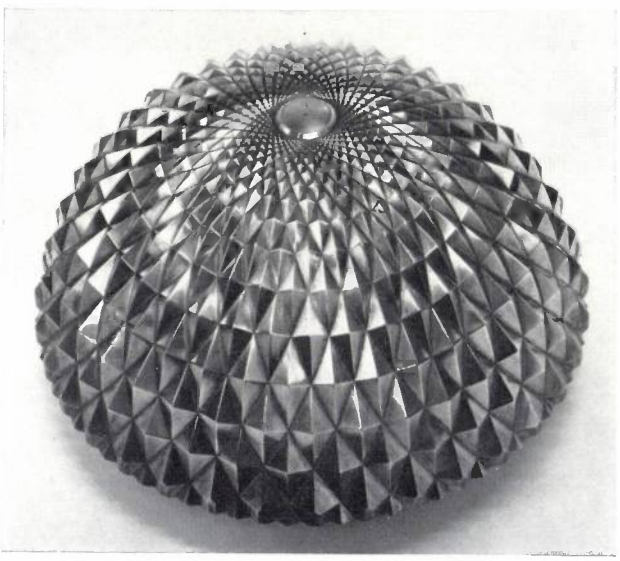
Spark machining is the most significant development in toolmaking since the war. The various methods of spark machining are well suited to reversal methods for toolmaking.

The first reversal method for which spark machining was used as an alternative was the finishing of profiled holes. In this operation alone the severe electrode wear — which was originally unavoidable — was overcome by the fact that the outer surface of the electrode could be made many times longer than the thickness of the die in which the hole had to be cut. Even the extremely unfavourable ratio between the eroded volumes of workpiece and electrode, as in the spark erosion of cemented carbide, could be overcome by using a sufficiently long electrode. Spark erosion was applied at a very early stage for making hard-metal die sets, and has contributed considerably to the wider use of these tools.

In the early days of spark machining — the early fifties — spark machining of blind holes was only accurate enough for making the shallow holes for forging dies, which did not require high accuracy or much detail.

In the middle of the sixties the picture was completely changed through the solution of the problem of electrode wear, which was reduced to about 0.1 % of the eroded volume of the workpiece. The method at once became a possibility for making deep, accurate cavities like the ones required in moulds for plastics. The range





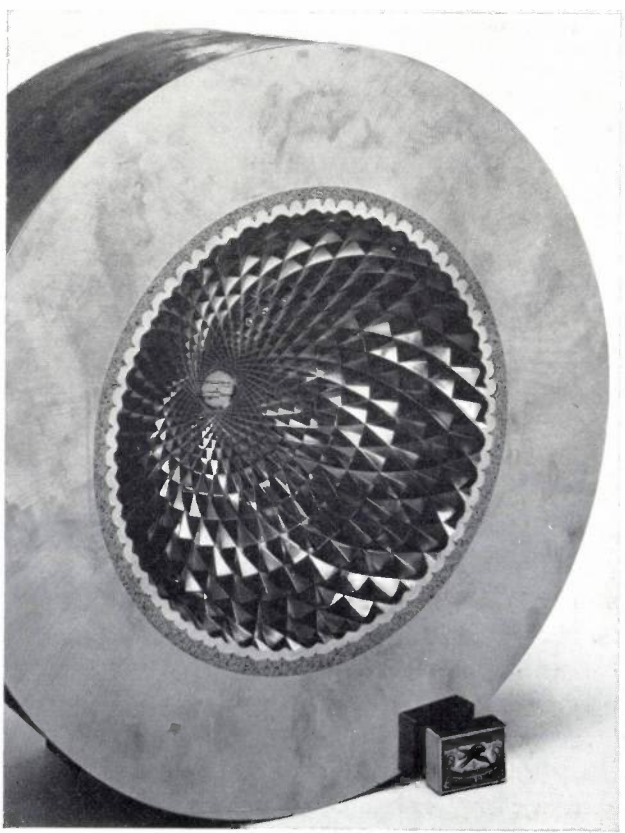


Fig. 13. Jig (above) with a metal shell deposited on it by electroforming (top picture) and die (right) in which the shell acts as mould cavity.

of applications of spark machining doubled in a few years as the older reversal techniques such as hobbing were superseded.

The revolutionary change in the technology of spark machining was attributable to a new spark generator equipped with power transistors. The old generator that is inversely proportional to the power. The latest type of spark has made it possible to erode dissimilar materials selectively, and with the correct choice of materials the erosion of the workpiece material can be a thousand times higher than that of the electrode (*fig. 14*).

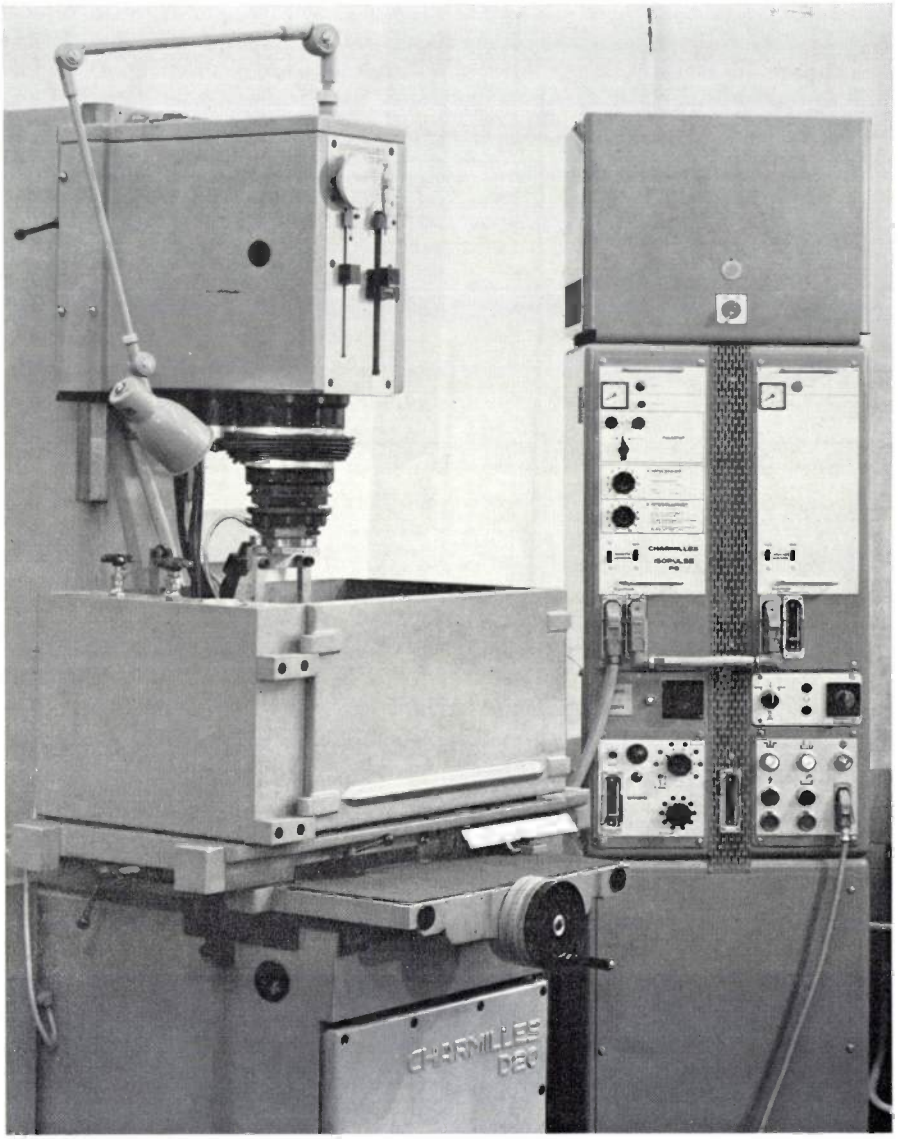


Fig. 14. Spark-erosion machine made by Charmilles Eleroda with Isopulse generator, in use at the Philips Centre for Manufacturing Techniques.

produced the necessary separate spark discharges by the discharge of a battery of capacitors, whereas the new generator simply opens and closes the connection to a d.c. source. The capacitor-discharge generator produces sparks of very high power and very short duration, whereas the sparks from a modern generator give the same energy for a much lower power and a duration

Although the reversal of a positive shape, enabling it to be transferred directly into the hard material, is by far the most widespread application of spark machining there are some other applications that are worth discussing.

In one form of spark machining the electrode consists of a taut wire which cuts contours out by erosion, in

much the same way as a fret-saw. Owing to the absence of machining forces the wire is not deflected and the accuracy can be high. The tensioned wire runs over rollers, so that the part of it subject to erosion is continuously renewed. This contour cutting can effectively be controlled by means of a numerical-control system.

specialized machine and generator needed for it were developed at Philips Research Laboratories [22].

With this high-precision spark-machining technique contour cutting of the same type as wire cutting can be carried out, but now using a finite wire electrode. In conjunction with an appropriate control system, this

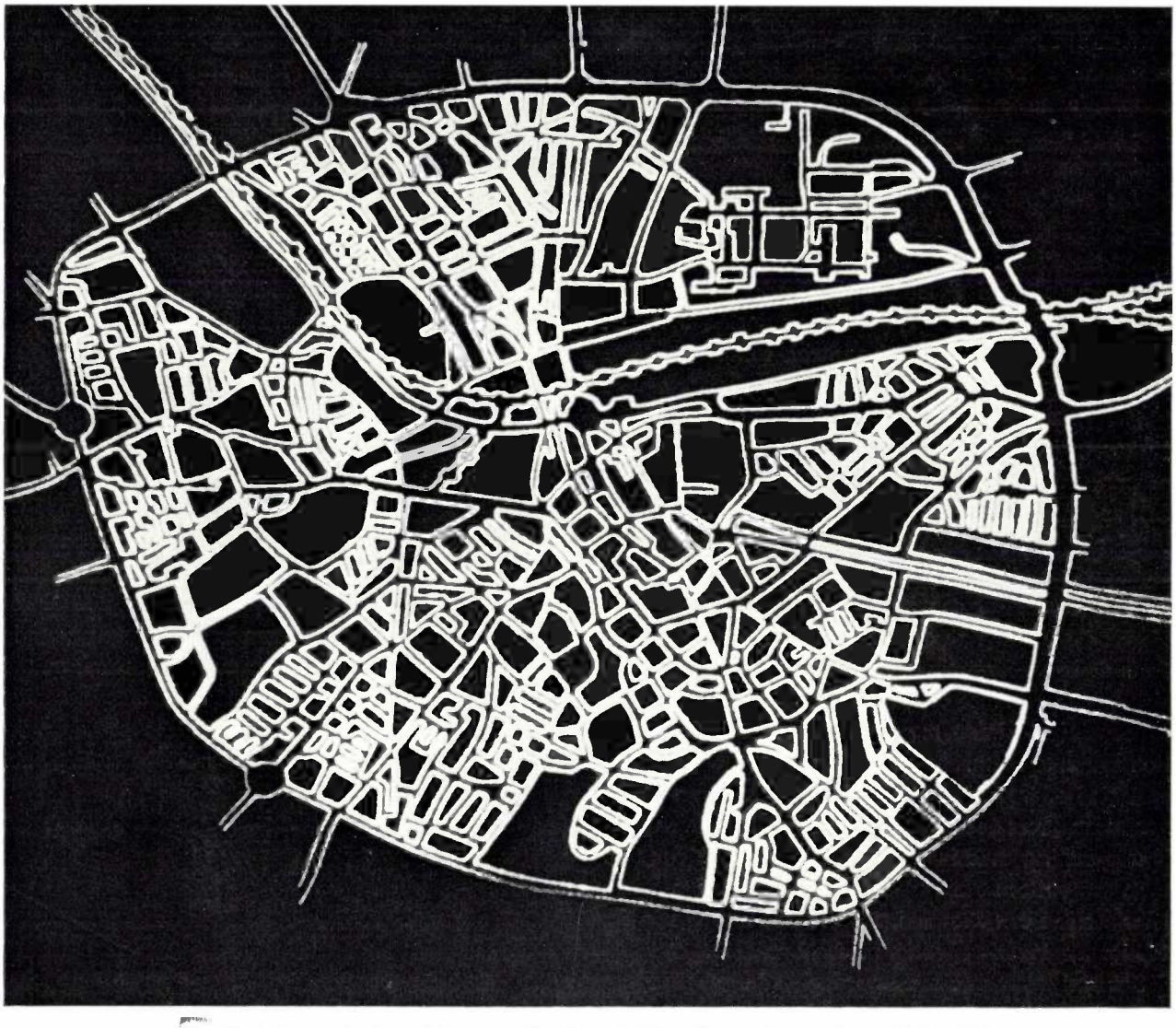


Fig. 15. A demonstration of the exceptional capabilities of precision spark machining. This map of the centre of Eindhoven was cut by spark machining in a metal film evaporated on to a glass plate, and its true size is only $2 \times 2\frac{1}{2}$ mm. A very special method was used to control the size of the spark gap and a photoelectric line-detection system was used to transfer the lines of the original drawing, much reduced, to the workpiece [22].

Spark machining can be adapted for drilling very small holes (up to a diameter of 5 μm). Conventional drilling is ruled out for holes of this size, since the hair-fine drills it would require would completely lack stiffness. The absence of machining forces also makes spark machining very suitable for this application. The method of carrying out the operation and the highly technique makes it possible to cut very accurate contours of complicated shape, as in the masks used in integrated-circuit work. *Fig. 15* gives an impression of what can be achieved with this technique.

[22] C. van Osenbruggen, High-precision spark machining, Philips tech. Rev. **30**, 195-208, 1969 (No. 6/7).

Numerical control

The advent of the digital computer led many people to predict a new "industrial revolution". Similar predictions were made about a revolution in the workshop when numerical control was introduced. In the first industrial revolution the steam power that drove the machines forced engineers to find new cutting materials that would enable this power to be used to the full. It was obviously to be expected that external changes, in this case development of data-handling techniques, would have far-reaching consequences for the workshop. We have already mentioned the information-intensive nature of the work at the interface of design and production.

The development of numerical control started in about 1950 in the United States of America. The pioneering work done at the Massachusetts Institute of Technology, with the financial support of the U.S. Government, related right from the outset to the most complex form of numerical control, contour milling. The motivation for the governmental support of this project was the recognition that the making of the numerous jigs and fixtures needed in the manufacture of intricate aircraft was an exceptionally lengthy process and thus affected the tooling-up time for war production of aircraft. The first development project aimed right away at the technically most significant application of numerical control, relating as it did to the primary manufacture of complex workpieces straight from design information. Moreover, these workpieces, since they contained contours not accessible to simple kinematic generation, required a great deal of information to make them.

A kind of workpiece that is simpler, but nevertheless contains a great deal of information, is the cam. In 1955 a cam-milling machine [23] of high accuracy for that time was built at Philips Research Laboratories, and a later version of this machine was taken into use in the Philips Engineering Works as early as 1961.

Strangely enough, the approach of the American machine-tool industry — and even more so that of the European industry — in the period that followed was almost diametrically opposed to this first large project in the field of numerical control. Two-axis positioning tables with a very simple positioning system sprang up everywhere. These simple jig tables had the advantage of being within the financial means of the ordinary firms, but in most cases they did not offer very much economic advantage. As *fig. 16* shows, numerical control becomes more attractive as the amount of geometrical information in the design increases: this information can be measured by the number of dimensions on the drawing that have to be achieved by metal-cutting operations on a machine tool [24].

An important step forward, in about 1956, was the idea of the "machining centre". One of the earliest and most successful versions of this idea was the Kearney and Trecker "MilwaukeeMatic". The machining centre utilizes the potentialities of numerical control by performing a large number of operations in one or more settings on a single machine. This machine is usually a horizontal-boring machine that can perform milling, drilling, boring, tapping and facing operations; it is the most versatile machine in the workshop, and performs all these operations on the frequently intricate types of workpiece that serve as machine frames or housings. Because of the large amount of information involved in their manufacture, these workpieces were made in the past by dividing the production process into a large number of simple sub-operations, and these components therefore determined the delivery time of the product in which they were used. *Fig. 17* shows a Scharmann machining centre with Philips S-NOR numerical control.

With the development of the equipment the problem of generating the information was recognized, as early as 1952, at M.I.T. This resulted in a programming language, known as APT, which is now used internationally for contour and positioning problems [25].

At first sight it would not appear particularly attractive to apply numerical control to lathe work. The drawings for lathe work contain on an average far fewer dimensions than those for frames and housings made on "machining centres". In many cases, however, the workpieces have to be turned in a large number of cuts. This amounts to producing a large number of workpieces each of which is made from the preceding one.

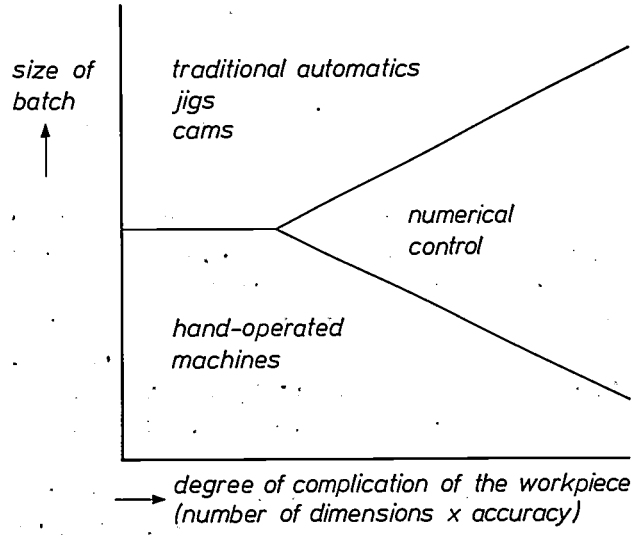


Fig. 16. The choice of the type of machine depends on the size of the batch to be manufactured and on the complexity of the workpiece. For very complicated workpieces numerical control is a paying proposition even for relatively small production-runs.

A drawing of the complete series of workpieces would in fact contain a very great deal of information.

Only the dimensions of the last workpiece are functional and therefore established by the designer. All dimensions of "intermediate workpieces" are chosen to suit the machining operations. These dimensions ought therefore to be established during tooling up, in the optimization of the working method. Numerically controlled turning will only give its maximum benefit when optimization of the working method by means of a computer program provides direct information suitable for controlling the machine. The "Miturn" program developed by the Metalworking Centre of the Netherlands Organization for Applied Scientific Research (TNO) is an important contribution in this respect.

Numerical control can most properly be referred to as a "breakthrough" where it concerns true contouring control. All the earlier existing machine tools could only generate simple kinematic shapes. The most usual shapes were cylinders and prisms, and the most intricate ones were screw threads and involute gearwheels. Copy-milling is only apparently an exception to this rule, since this operation requires an existing model. Making this master model is the crucial problem, not making the copies of it.

The amount of information needed for exactly defining a given contour is in theory infinitely large. In practice, of course, the quantity of information is not infinite, though it is always very large, and the more accurately the contour has to be defined the more information must be provided. The problems involved in producing such a contour are twofold: firstly a large amount of data relating to the shape must be generated in the design drawing office, and secondly this informa-

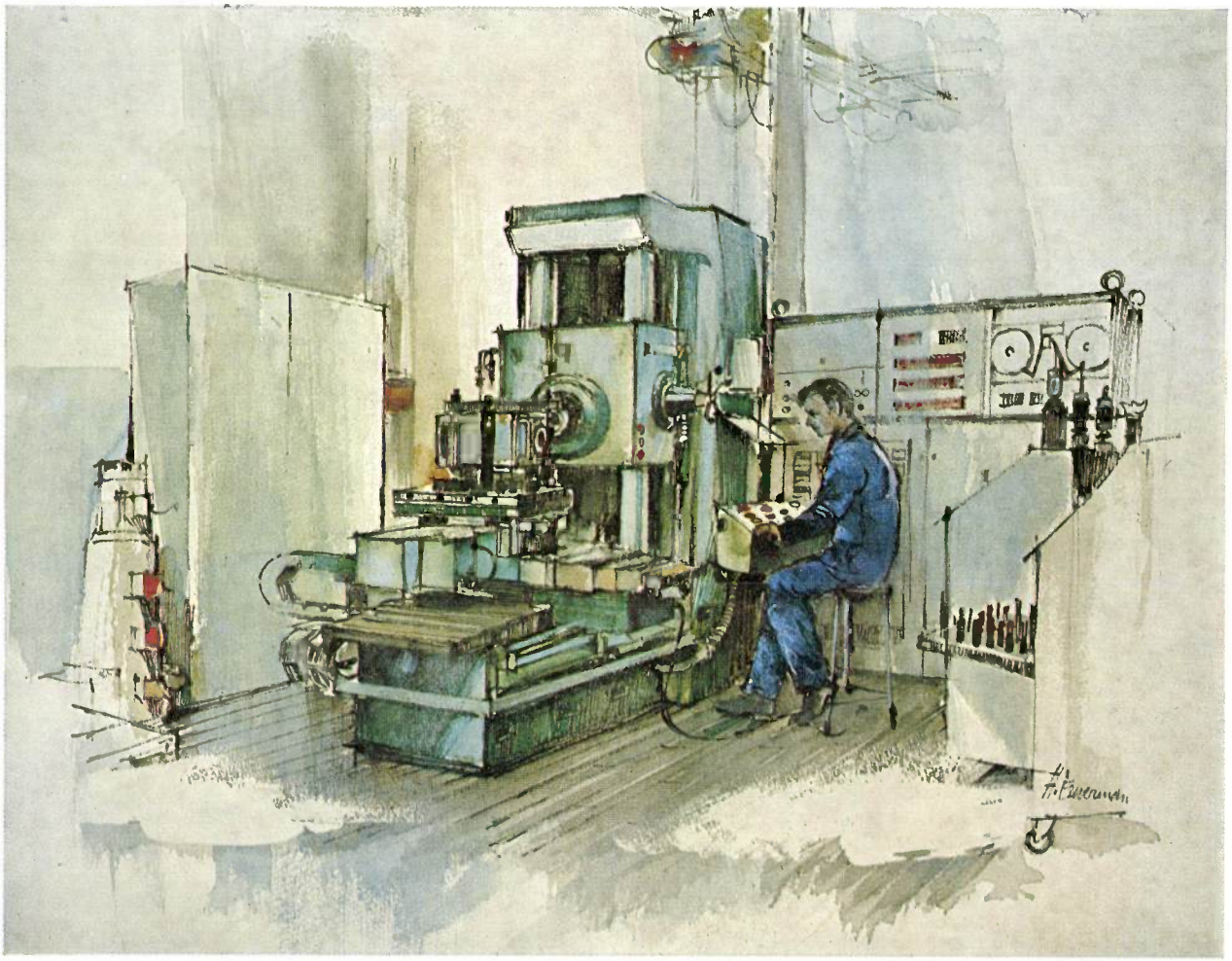


Fig. 17. Scharmann machining centre, equipped with Philips S-NOR numerical-control system. (Drawing by H. Euverman.)

[23] J. A. Haringx, R. C. van Ommering, G. C. M. Schoenaker and T. J. Viersma, A numerically controlled contour milling machine, Philips tech. Rev. **24**, 299-331, 1962/63.

[24] H. Huizing, Over de toepassing van numerieke besturing, Metaalbewerking **31**, 97-102, 1965/66.

[25] J. Vlietstra, The APT programming language for the numerical control of machine tools, Philips tech. Rev. **28**, 329-335, 1967.

tion must be processed while the machine is in use — and in such a way as to fit in with the timing of operations of, say, a milling machine.

Modern electronic equipment for digital data processing provides the answer for both sides of the problem: the computer can generate in an acceptable time enough information to define the shape with great accuracy, and the numerically controlled machine tool, controlled with this information, can make the required workpiece at a hitherto unprecedented rate. The transfer of the information generated is a problem here; it becomes more and more necessary to integrate the generation and the processing of the material.

An integration of this kind has been achieved at Philips Research Laboratories, for the manufacture of cams. With a cam we are concerned with two dimensions whose kinematic relationship is dependent on the application of the cam and also on the kind of cam-follower mechanism. The equations in which this relationship is expressed are very suitable for generating the shape information from simple design data. The desired contour is produced quickly and accurately on a milling machine of special construction. The machine is controlled by the same minicomputer that has just previously calculated the control information (*fig. 18*).

Shapes that are designed from aesthetic considerations and are not susceptible to mathematical definition derived from their function can be produced in a "conversation" between a "mathematically" programmed computer and an "aesthetically" evaluating designer. The computer can ensure that the boundary conditions incorporated in the program are continuously fulfilled during the process.

The amount of design information produced in this way can be made as large as is needed for the most accurate definition of the shape. This information is moreover available in numerical form, and is thus an ideal input for the numerical-control unit of a machine.

Integrated systems for computer-aided design, followed by the production of designs using numerically controlled milling machines, are now employed on a very wide scale for die manufacture in the automobile industry. The exact definition of shape and the exact observance of specific boundary conditions are particularly important in mass production. Parts of the bodywork of any desired shape can now be manufactured so as to be fully interchangeable and to fit each other just as reliably as pistons and cylinder blocks in the days of Henry Ford's "Model T".

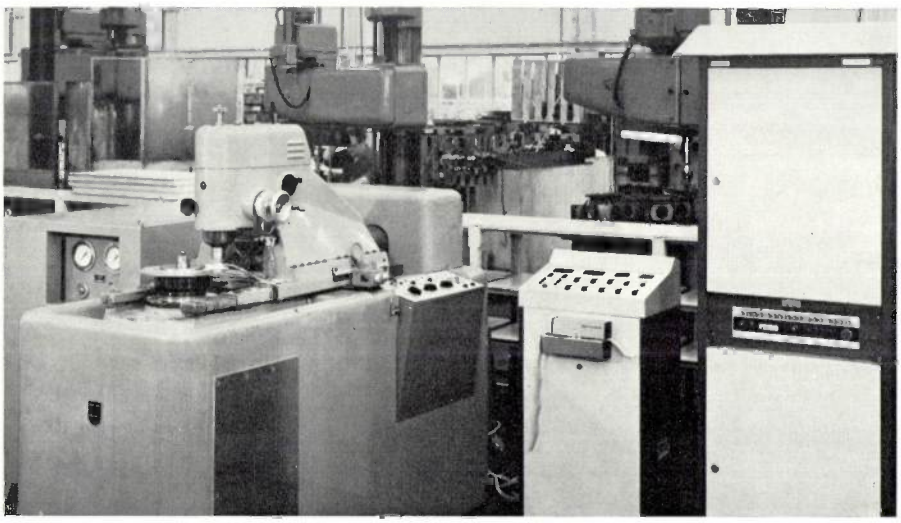


Fig. 18. Cam-milling machine, directly controlled by a Philips P 9201 computer. The control system and the software were designed by Philips Research Laboratories and the machine was designed by the Philips Engineering Workshops.

The numerically controlled three-, four- or five-axis milling machine is becoming a worthy and even more successful successor of the copying lathe in the tool-making shop and in the aircraft industry (*fig. 19*). The copying lathe, useful though it was, contributed nothing to the making of the master model, let alone to the generation of the enormous amount of design information needed to define a master model of arbitrary shape in three dimensions.

In contour milling the computer and the numerical-control system undertake work which in the past was performed by human hand and brain only when there was no other way. The aircraft and marine design engineer filled pages with calculated coordinate tables, and the man on the shop floor had an equally difficult job in interpreting this information to set up jigs, mark out plates or centre them, or drill and file templates.

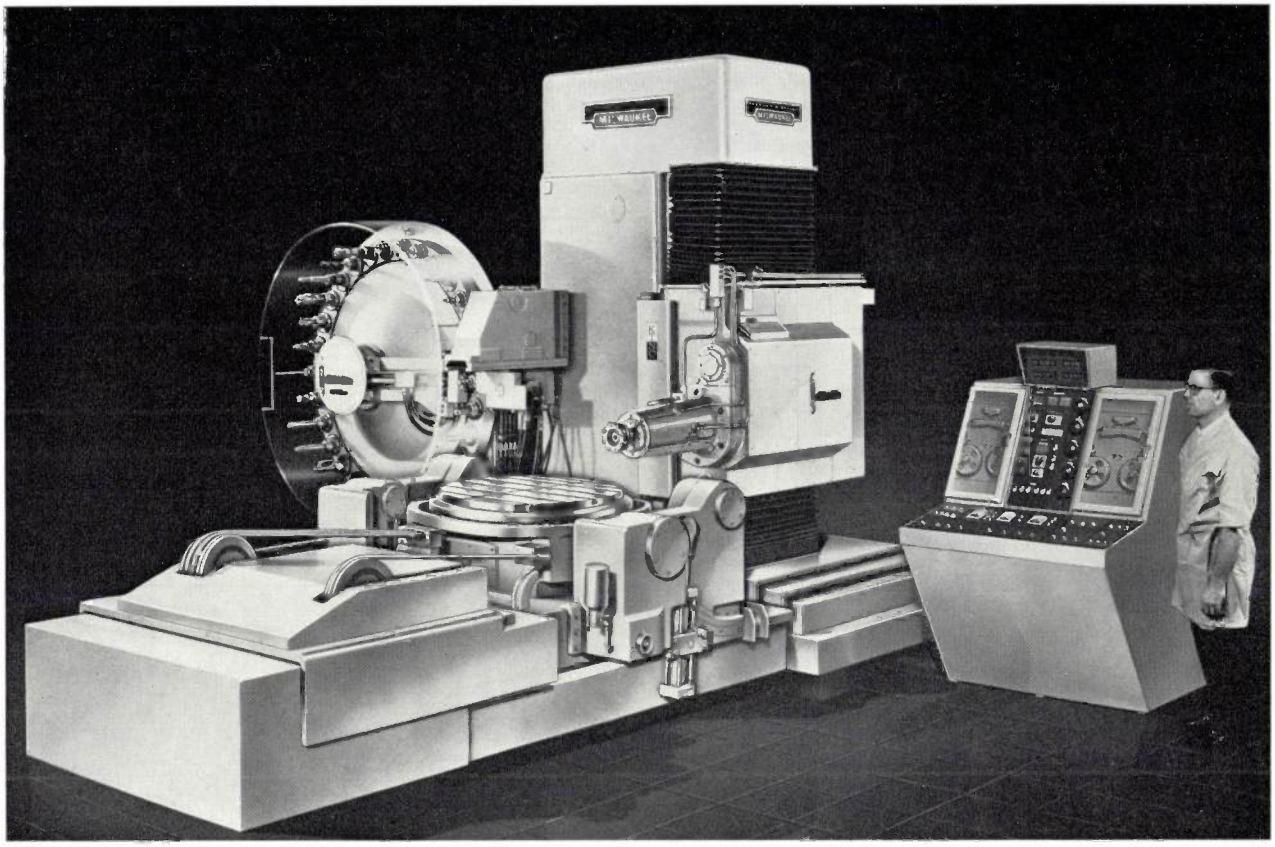


Fig. 19. Five-axis numerically controlled milling machine (Kearney and Trecker, Milwaukee-Matic model III, 5-axis). This machine will be set up in 1971 in the new workshop at the Philips Engineering Workshops to manufacture pressed-glass moulds for television tubes.

The less advanced forms of numerical control take over the more ordinary tasks of the craftsman. The great question is whether numerical control is going to turn the skilled craftsman into a machine-minder, as happened to the unskilled worker before him as a result of the earlier mechanization.

Like the skilled work of the craftsman, numerical control lies at the interface between design and production and is potentially capable of taking over here much of the work of the skilled craftsman. No one can predict at the present time, however, how far the evolution will go or how fast it will proceed. Complex machines and ingenious techniques have never yet made human work redundant, and, however much may change in the workshop, this is just as unlikely to be the case as it was when the manual skills of the old-style toolmaker were taken over by machines.

This "universal" laboratory building (WY) on the Waalre complex of Philips Research Laboratories has been in use since 1968. On its roof, on the right-hand side, there is an aerial for the experimental colour-television station. The following article gives an account of the buildings and also of the planning and development of the laboratory complex.

The Waalre complex of Philips Research Laboratories

L. A. de Haas and S. S. Wadman

When the Waalre complex of buildings for Philips Research Laboratories was officially inaugurated in 1963 it consisted principally of the low building WA and the multi-storey block WB with associated workshops [1]. Since then various new buildings have been built and put into use, including the multi-storey laboratory block WY.

Scientific research in industrial laboratories nowadays calls for great flexibility, not only in the overall layout of a complex of this type, but also in the individual buildings and the technical facilities. The basic plan devised at the outset for the Waalre laboratory complex has been found more than sufficiently adaptable to allow the continuously changing requirements to be met.

The emphasis has, of course, shifted here and there. For example, the idea of dividing the complex into separate sectors has rather tended to recede into the background. Personal contact between staff engaged in different branches of research has proved in practice to be even more important than was originally assumed, and the sector idea has obviously not been adhered to for very costly and specialized technical equipment and facilities that could not have been made economic in in a single sector. Nevertheless, the idea that a sector should consist of a universal laboratory, a number of specialized laboratories and a workshop has proved to be right, and it remains — albeit in a modified form — a guiding principle for the further extension of the complex. It is true to say, however, that the need for specialized laboratory space has grown much more in recent years than that for universal laboratory space; because of this there has also been an increase in the floor area needed per member of staff.

Another essential requirement of the basic plan was that the laboratories should be subject to the minimum of disturbance from outside. A separate area was therefore reserved in the basic plan to accommodate small buildings for research work in which there might be risk of fire or explosion, and also buildings such as the boiler house, stores for acids and containers, the telephone exchange, the water supply and drainage facilities and the maintenance workshop. This area was to be separated from the laboratory area by parking sites. In the expansion of the complex there has been no departure whatever from this idea, and indeed it will receive even more attention in future planning.

The consistent adherence to the third guiding principle of the basic plan — perhaps the most important one for many people inside and outside the laboratory — which was that the laboratory complex should have an open character and detract as little as possible from the natural landscape between Eindhoven and Waalre, appears from the photographs in this article and on the facing page. The original plans for landscape design [2] have been modified where necessary, but on the whole remained unchanged. It cannot be stressed enough that the harmonious blending of greenery and buildings in a complex of this size is essential to the creation of a good working and living environment.

Before taking a closer look at the buildings at present in use or under construction, we ought to mention an important part of the basic plan that still remains on paper. This is the administration building. In view of the more urgent need for laboratory space, this building will not be started until later. The plans have also been slightly modified so that the administration building will not be the only centrally situated building on the complex. The Patents Department has also been in great need of extra space, and because this Department plays such an important part in the day-to-day work of the laboratories a joint decision was taken to rehouse the Patents Department at Waalre, where it will continue its activities as an independent unit. The new accommodation for the Patents Department is already under construction and will be linked with the administration building of the laboratories.

Ir. L. A. de Haas is a Senior Architect with the Architectural and Civil-Engineering Department of the Philips Plant-Engineering Division; Ir. S. S. Wadman is in charge of the General Technical Services of Philips Research Laboratories.

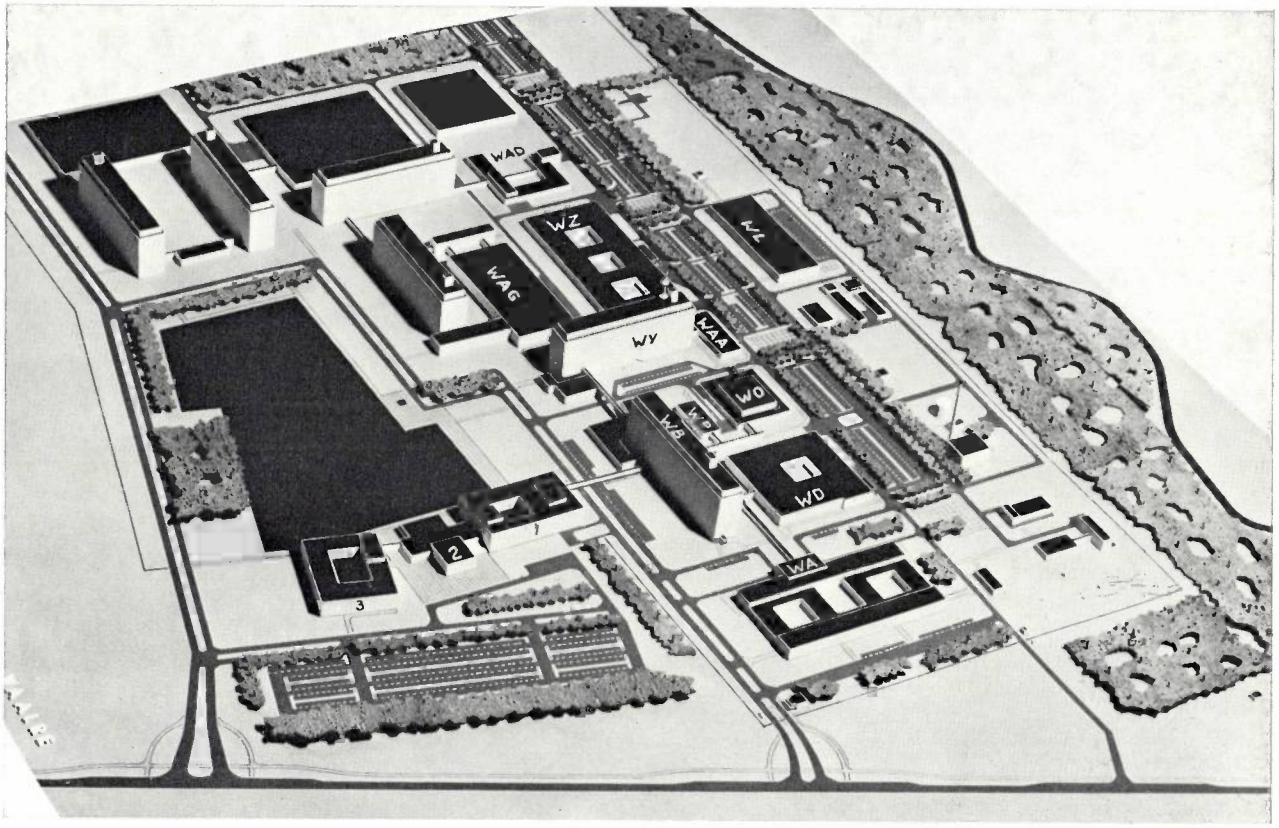
[1] See M. J. Jansen Gration, The planning of the new complex of buildings for Philips Research Laboratories in the Netherlands, Philips tech. Rev. 24, 385-395, 1962/63. An explanation of the letter coding of the buildings will be found in the photograph of the model shown on page 155.

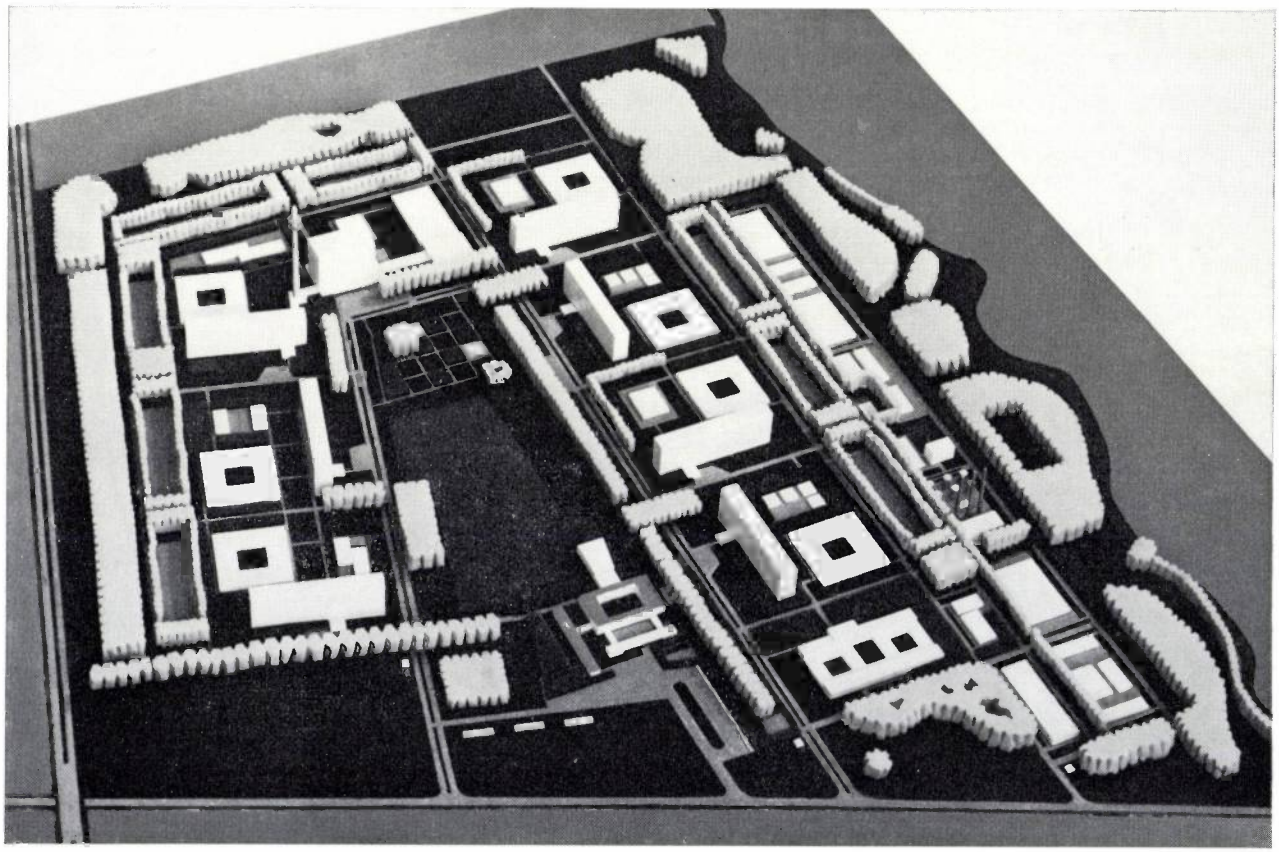
[2] Designed by Ir. J. Vallen, landscape architect, Roermond, the Netherlands.

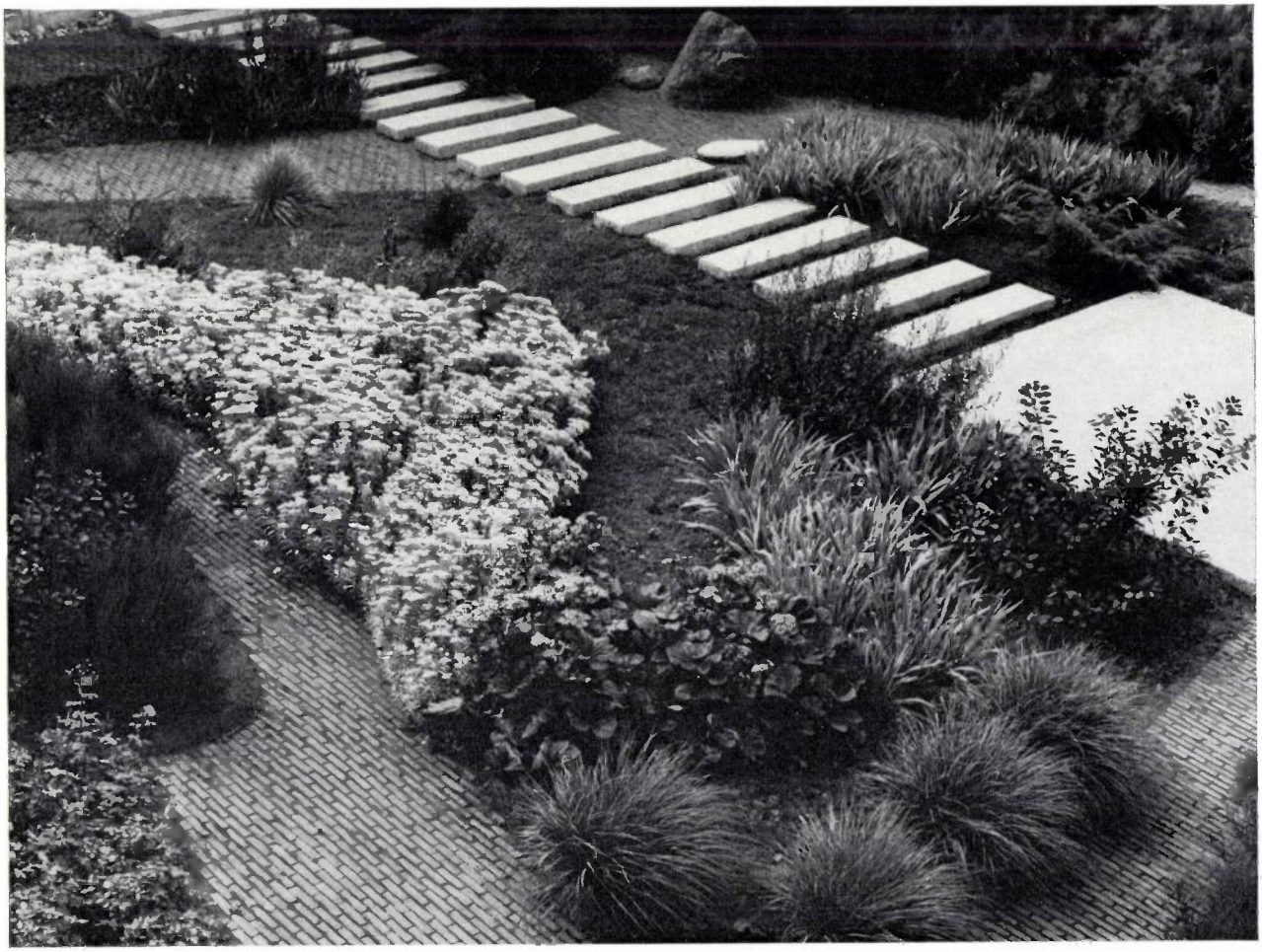
The necessary contacts between the main office block and the buildings in the individual sectors made their siting a problem that called for very careful study. The advantages of the layout envisaged in the basic plan will not be fully appreciated until the whole complex of buildings at Waalre has been completed.

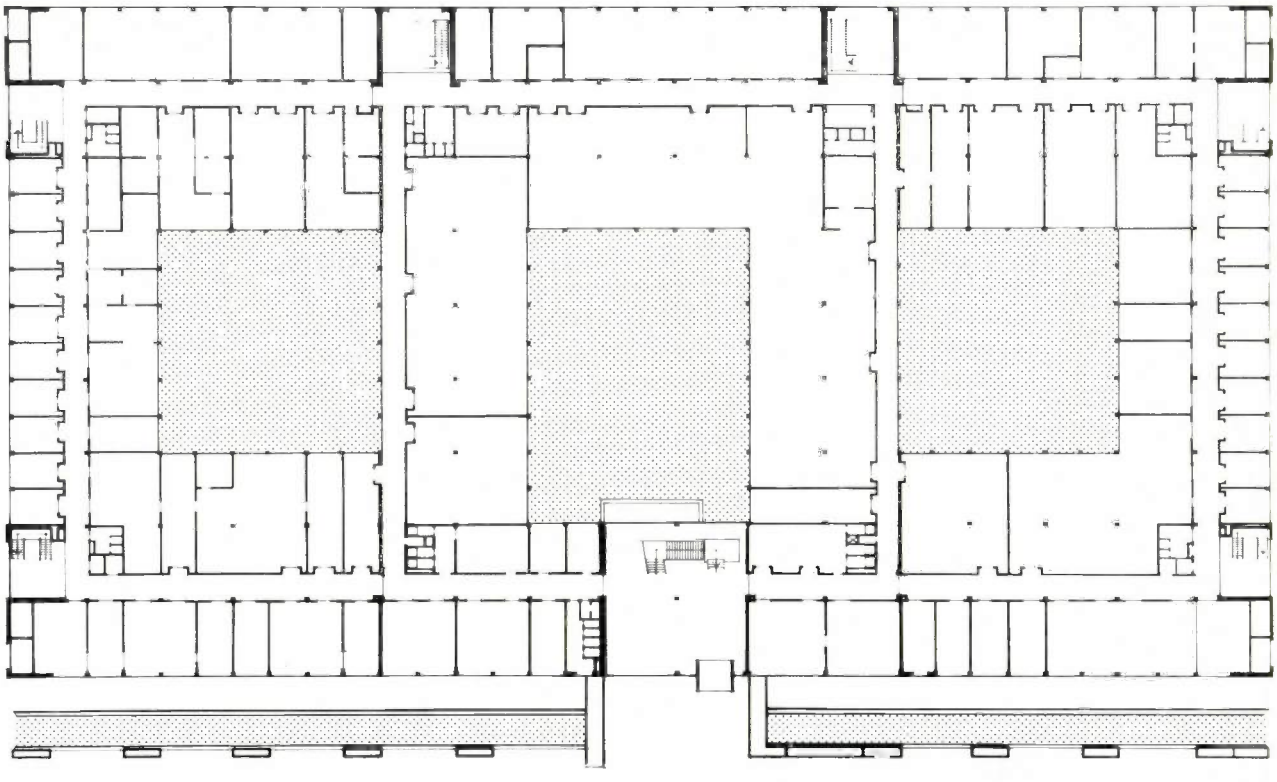
The first laboratory building to be erected on the Waalre site — building WA built in 1958 — was designed for technological research and may be regarded as a specialized laboratory. Because of the particular requirements they have to meet, the rooms in this building differ to some extent from the pattern of a universal laboratory. The building was also a pilot project to provide experience of equipment, ventilation and the supply system for water, gases etc. The setting of this first building, with its carefully planted garden courtyards and the surrounding ornamental gardens, bears witness to the not inconsiderable importance that we attach to the guiding principle relating to the landscaping of the laboratory site.

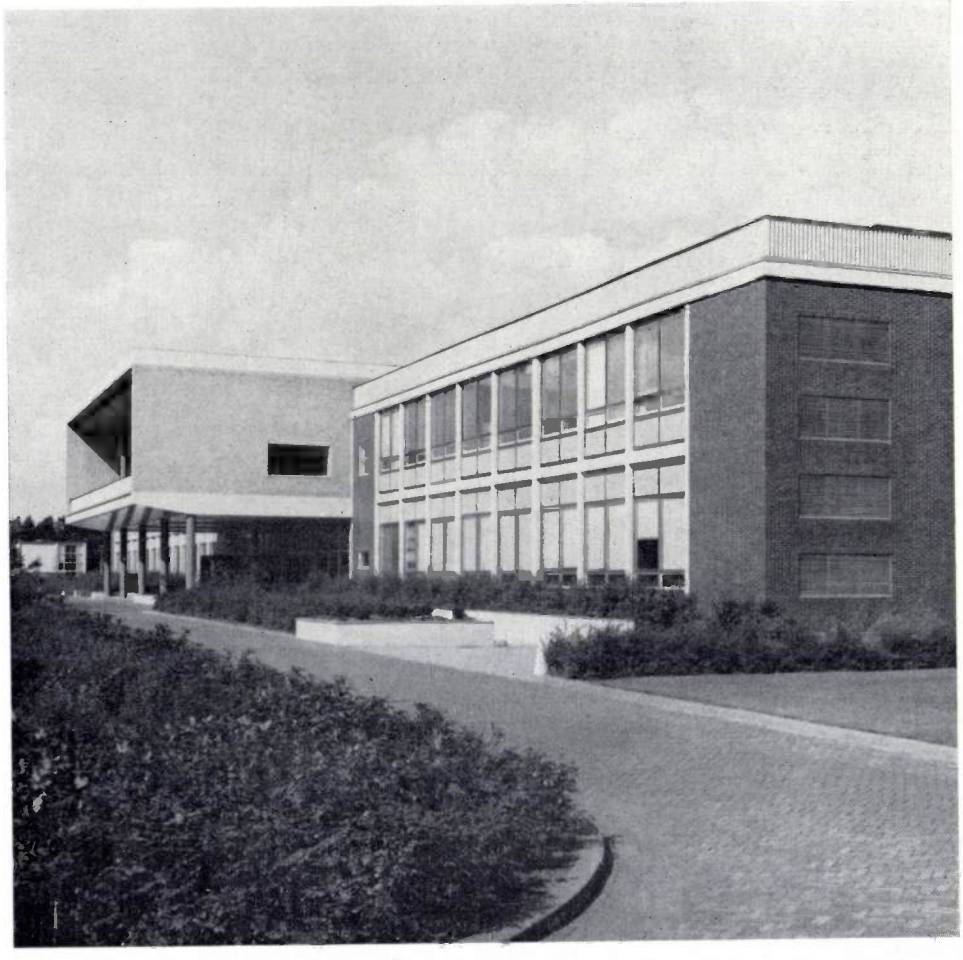
Above: an aerial photo of the Waalre complex taken in mid 1968. Upper right: a photograph of the model illustrating current plans; the model in the lower photograph shows the plans as they stood in 1958. The northern boundary to the site is formed by the river Dommel (on the right in the photograph). In the model in the upper photograph the most important buildings now in use or being built are marked by their code letters. The buildings on the nearer side of the lake are: *1* the administration building, *2* the large lecture theatre and *3* the Patents building.


WAD
WZ
WAG
WL
WY
WAA
WO
WB
WD
WA
2
3



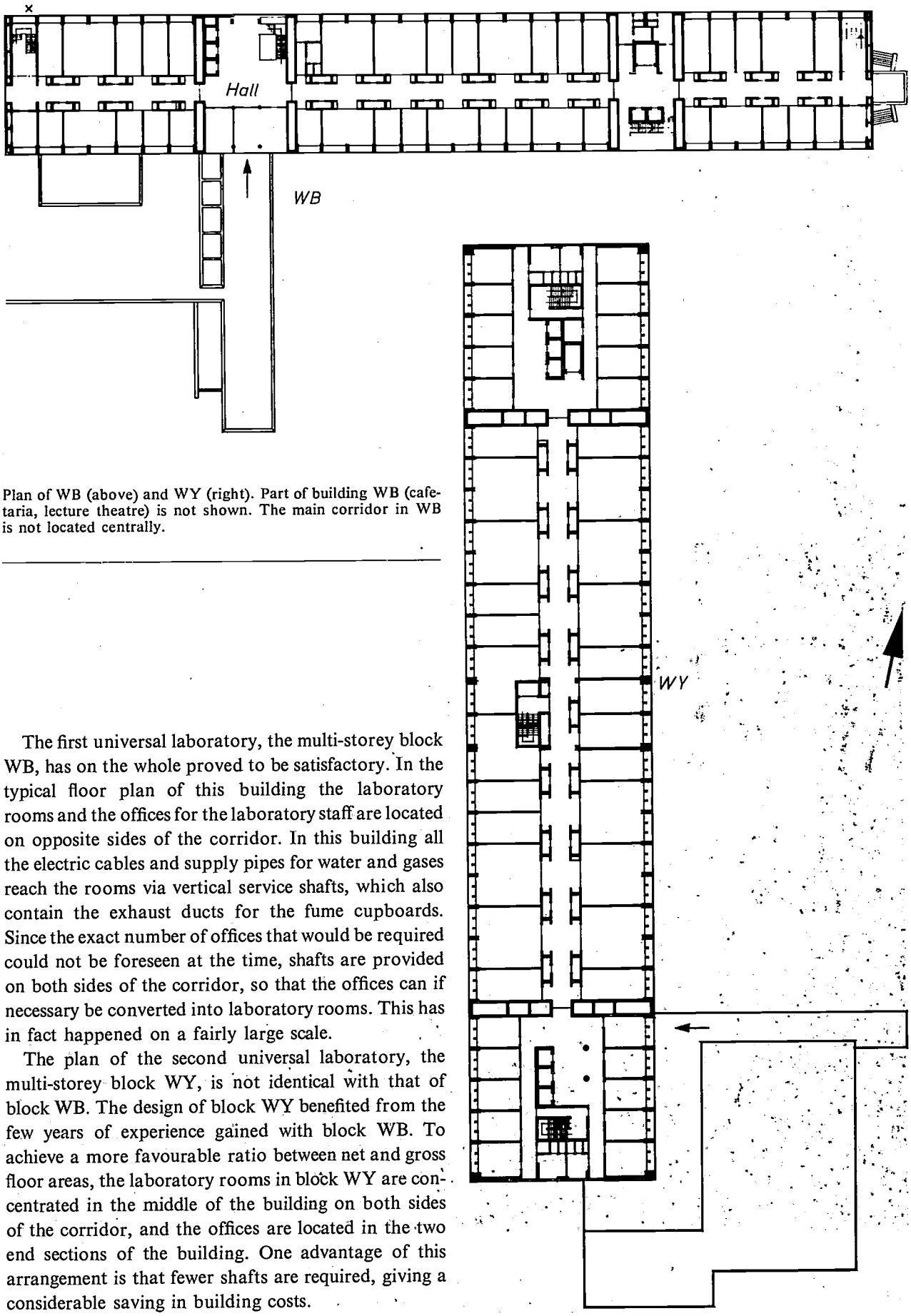






Building WA. Above: a plan of the ground floor; the three garden courtyards and the service roads to the basement are shown shaded. Right: the west front of the building, with the cafetaria projecting over the main entrance. The photographs opposite show the hall and one of the garden courtyards.

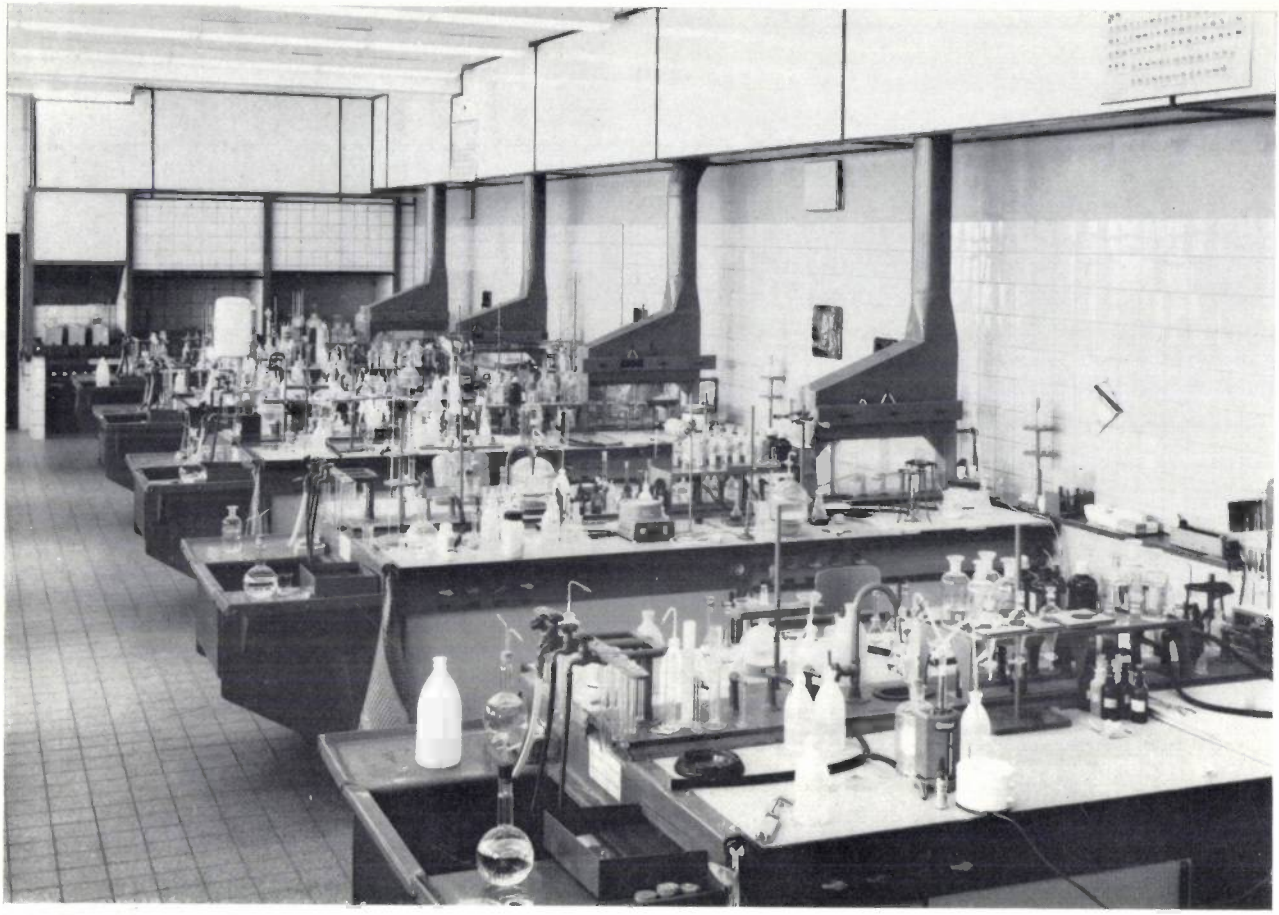
Building WY at night, seen from across the lake. When this photograph was taken building WAG had not yet been started.



Plan of WB (above) and WY (right). Part of building WB (cafetaria, lecture theatre) is not shown. The main corridor in WB is not located centrally.

The first universal laboratory, the multi-storey block WB, has on the whole proved to be satisfactory. In the typical floor plan of this building the laboratory rooms and the offices for the laboratory staff are located on opposite sides of the corridor. In this building all the electric cables and supply pipes for water and gases reach the rooms via vertical service shafts, which also contain the exhaust ducts for the fume cupboards. Since the exact number of offices that would be required could not be foreseen at the time, shafts are provided on both sides of the corridor, so that the offices can if necessary be converted into laboratory rooms. This has in fact happened on a fairly large scale.

The plan of the second universal laboratory, the multi-storey block WY, is not identical with that of block WB. The design of block WY benefited from the few years of experience gained with block WB. To achieve a more favourable ratio between net and gross floor areas, the laboratory rooms in block WY are concentrated in the middle of the building on both sides of the corridor, and the offices are located in the two end sections of the building. One advantage of this arrangement is that fewer shafts are required, giving a considerable saving in building costs.

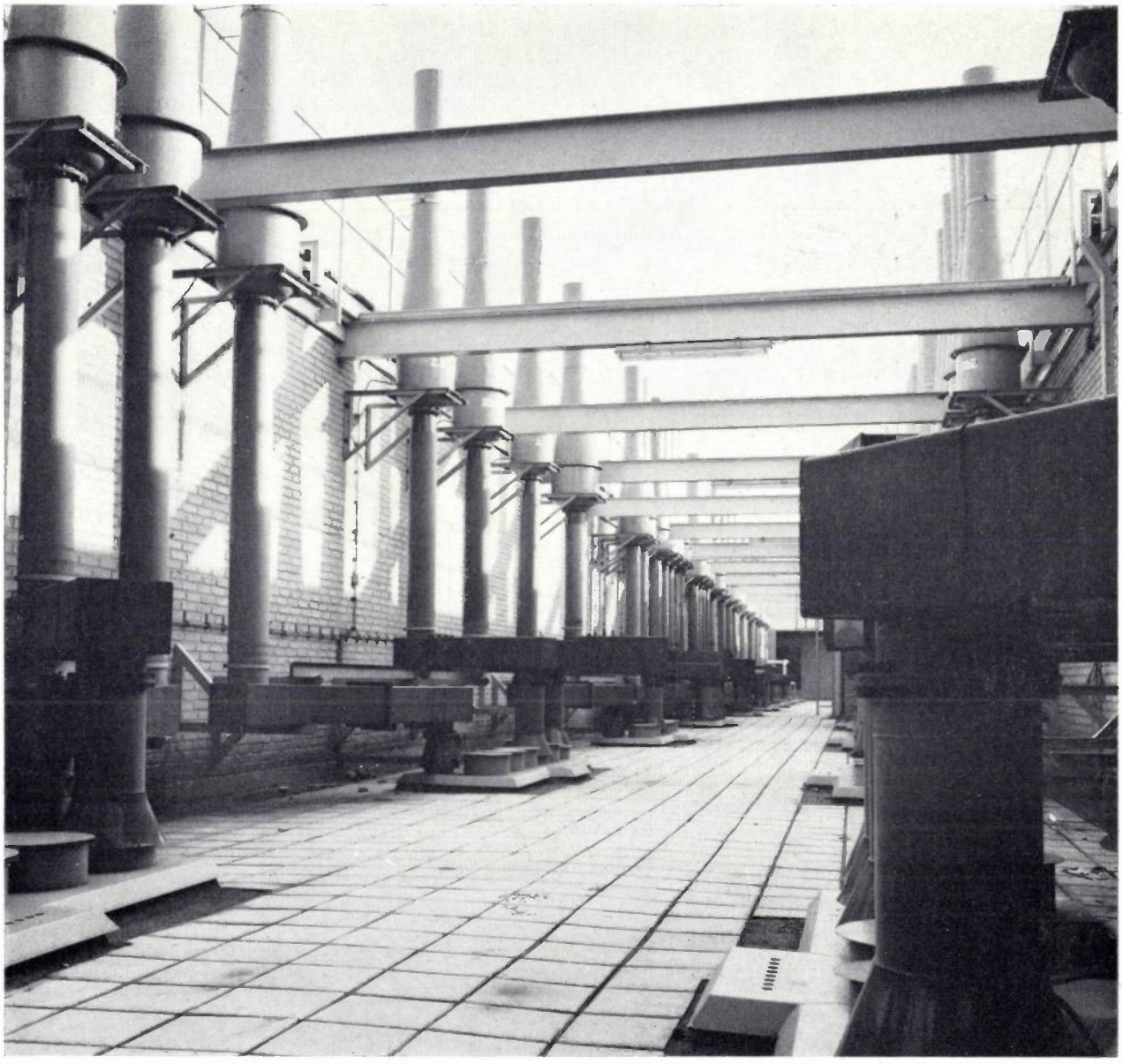


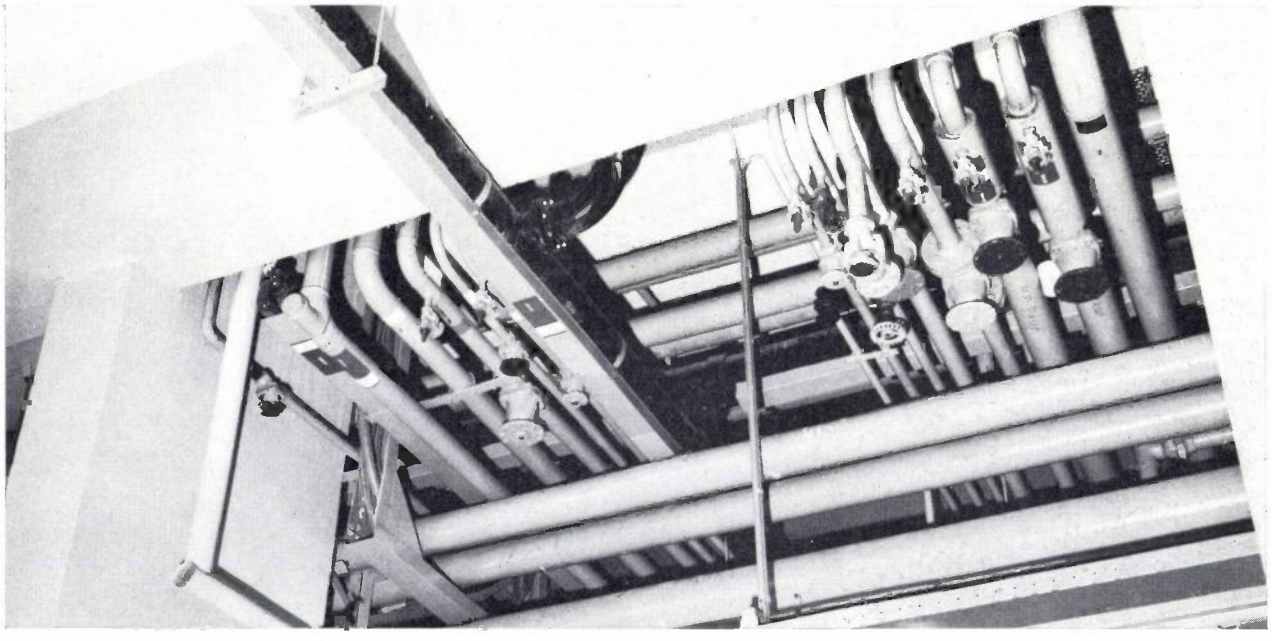
Above: one of the chemical laboratories in WB. Left: a service shaft in WB; doors on every floor give access to the service shafts. Upper right: on the roof of WY. The tops of the shafts can be seen; each shaft can carry up to eight ventilator outlets. Lower right: the "input" end of a service shaft in the basement of WY.

The vertical electrical distribution system used in block WB was changed to a horizontal system, giving a better distribution of the load. The ventilation system used in WB [1] was also changed in a number of respects, making more room available in the service shafts for supply lines and for special exhaust ducts.

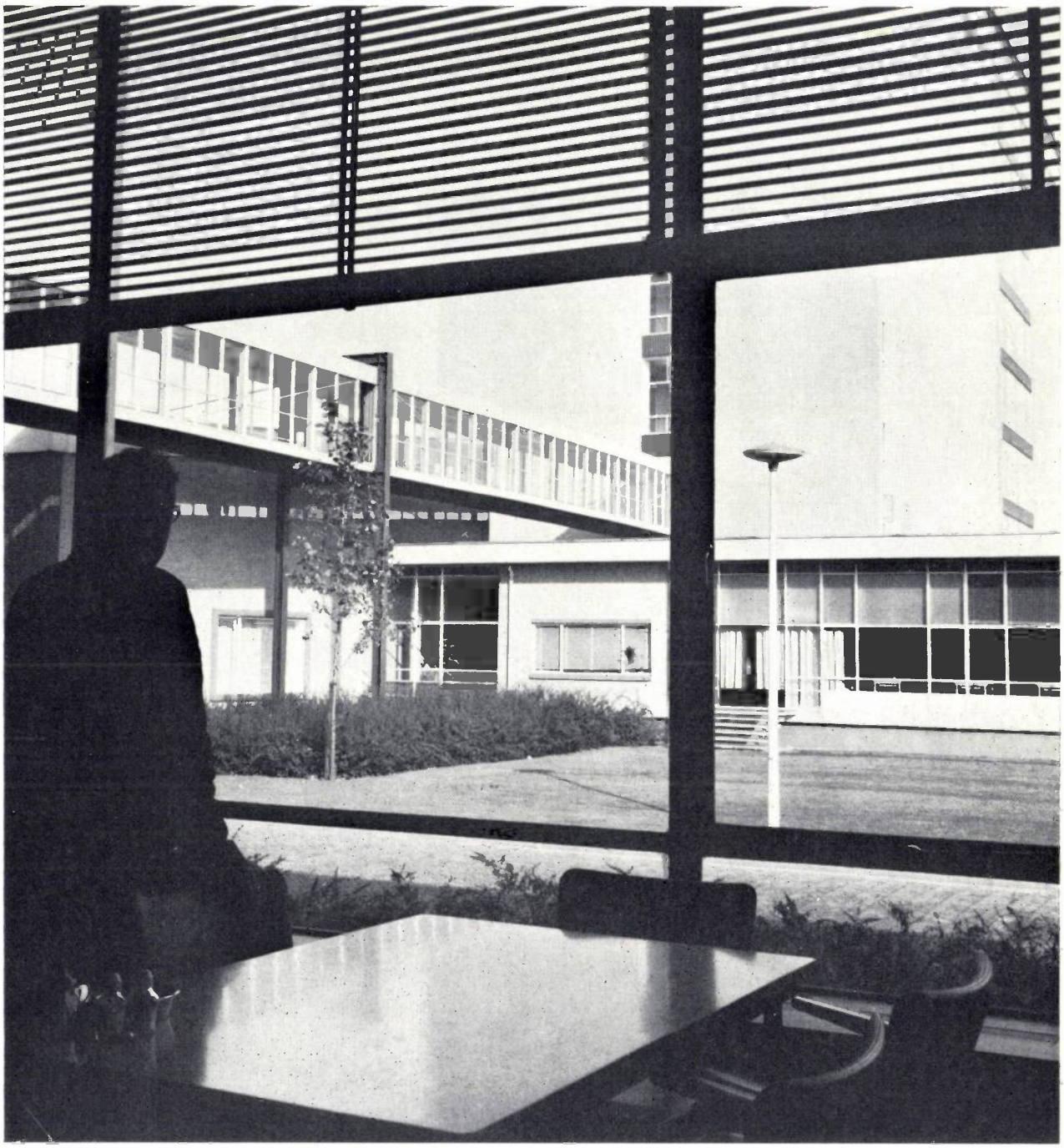
To help to achieve the best possible control of temperature and humidity in the rooms, the WY block, with its long sides facing East and West, has a more strongly profiled façade and smaller areas of glass than block WB.

In block WY the sound-proofing is applied direct to the underside of the floors, a measure that allowed a reduction in the height between storeys. As a result, although the building has roughly the same volume as block WB — which is fitted with lowered acoustic ceilings — it was possible to give it an extra storey.

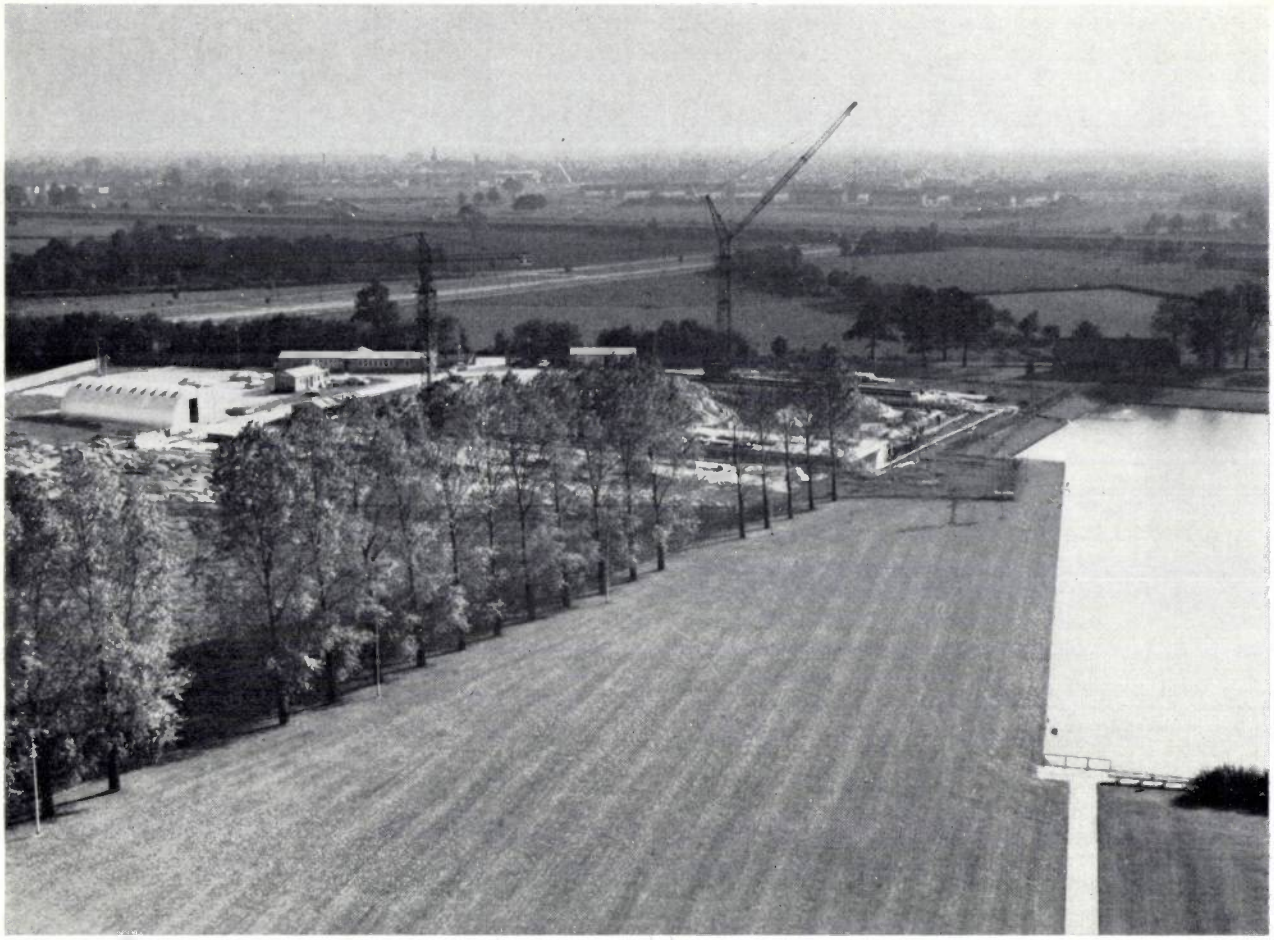




The lounge near the entrance to the lecture theatre, located in the annexe of building WB. The metal sculpture, by Toon Kelder (The Hague) was presented to the Laboratories in September 1969 by Prof. H. B. G. Casimir on the occasion of his 60th birthday.



A view of building WB from the cafetaria of WY. The covered bridge linking the two buildings can be seen at the upper left



There is also a distinct difference in the design of the workshops associated with the two universal laboratories. Workshop WD, belonging to the laboratory block WB, was designed as a simple workshop for glass and metal, with a basement to accommodate the supply lines and for storing materials. Workshop WZ, belonging to the laboratory block WY, has only an outward resemblance to building WD. Rapid advances in specialized workshop techniques made it necessary to design a workshop that would be capable of meeting specific requirements now and in the future, concerned particularly with temperature and humidity control and freedom from vibrations. The basement of this building is not exclusively intended for stores and pipelines, and a large part of it is laid out to accommodate special technical departments.

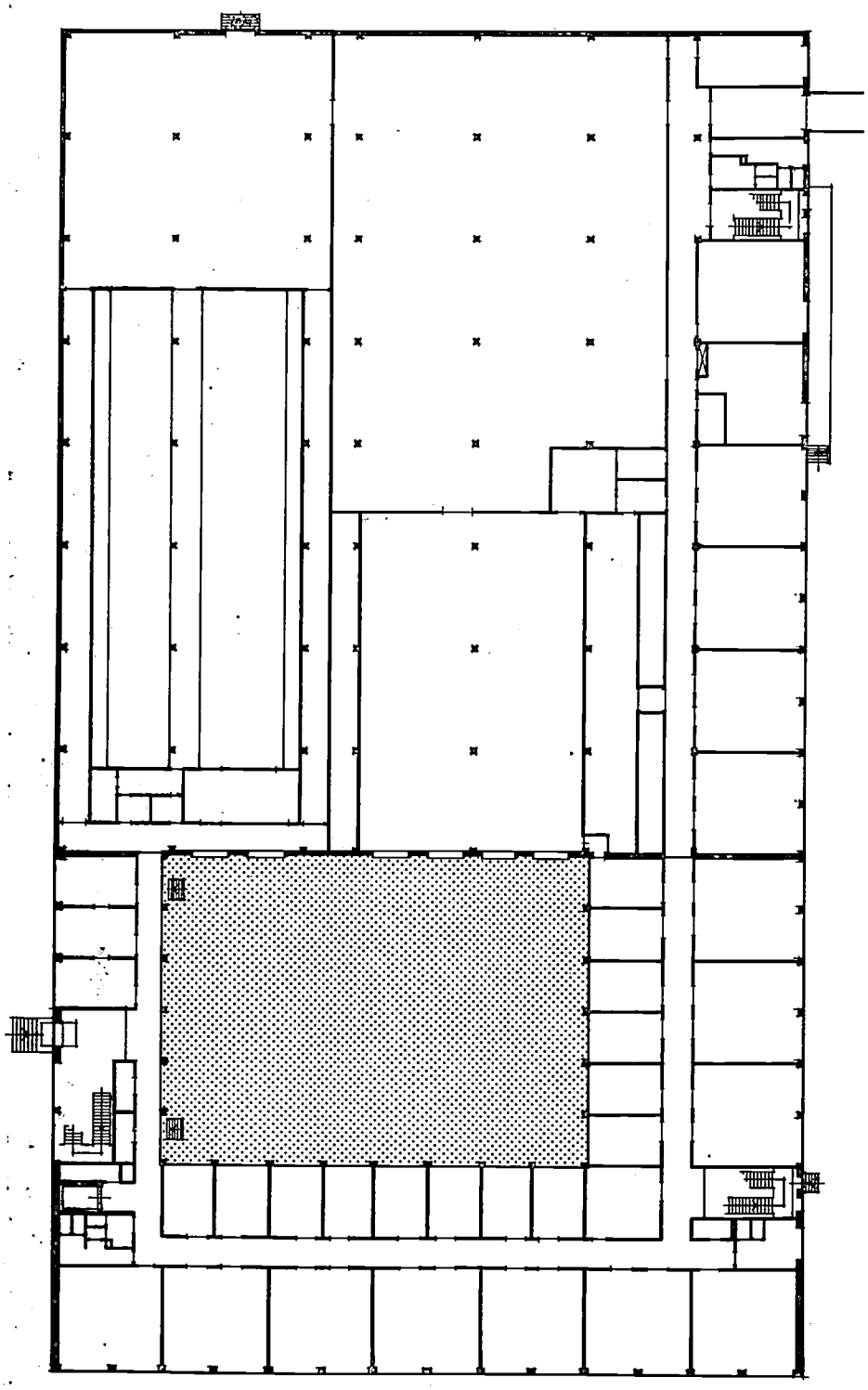
In its present phase of construction the Waalre complex contains, in addition to the pilot building WA, five other specialized laboratories, two of which are already in use and two under construction. In the first sector there is a building (WP) which houses a cryogenic

Above: the new Patents building in construction, seen from WB; part of the lake can be seen on the right of the photograph. Right: WAG in construction, seen from the building WZ under construction (see the picture of the model, p. 155); in the background the building WY can be seen.

laboratory, and a second building (WO) which accommodates a television and sound studio with associated laboratory rooms. The two buildings are connected with each other and with block WB by a covered passageway. The specialized laboratory WAA, in the second sector, is equipped as a scientific computing centre. Work has started on the erection of a laboratory for radiochemistry and a laboratory for special dust-free techniques, as required for the fabrication of certain semiconductor devices and integrated circuits.

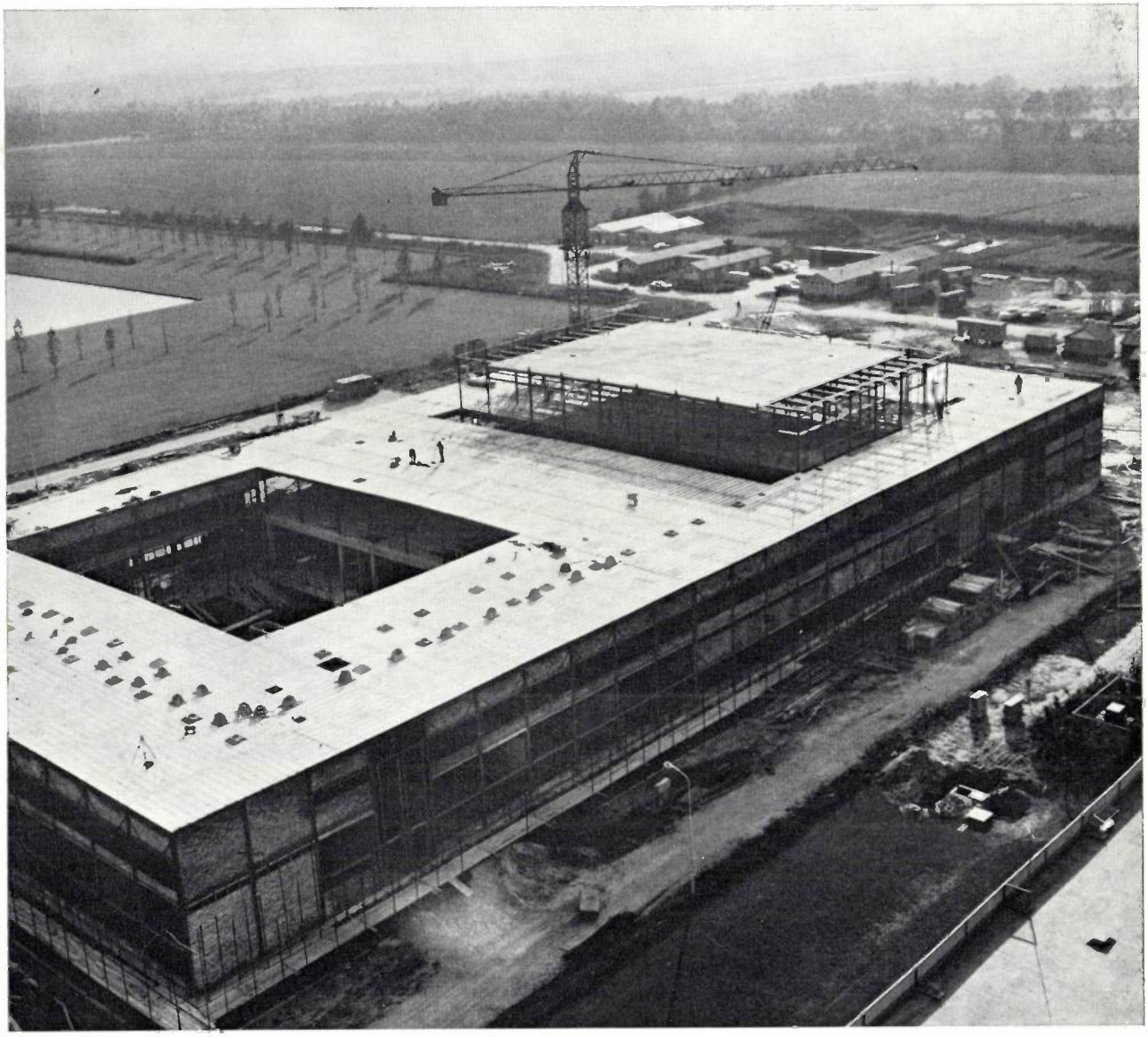
Since the start of the Waalre laboratory complex in 1958, which marked the beginning of a period in which the staff of the Research Laboratories increased from 1400 to 2300, the net area of the laboratories and workshops has grown to about 50 000 m^2. About 8000 m^2 of research accommodation will be added to this in the middle of 1971.

For the more distant future plans are being made to build a third universal laboratory and a specialized laboratory for nuclear-physics techniques.



Plan of WAG. The interior courtyard (shaded) is surrounded on three sides by offices and laboratories. On the fourth side (the upper one in the photograph) there are two dust-free rooms. The ventilation system for these rooms is located on a special floor, one storey above. The equipment for temperature and humidity control is located in the "superstructure" above the main roof; see the photograph on the right.

Another day's research comes to an end ....



A small generator-set with a Stirling engine driven by the heat from a brazier. The Stirling engine is connected to the heat source by a heat pipe. This method of heat transmission is of considerable interest for vehicle engines.

Prospects of the Stirling engine for vehicular propulsion

R. J. Meijer

Introduction

In our changing society increasing attention is being paid to the general climate of life on Earth. Indeed, it is becoming an ever greater problem to maintain the habitability of the Earth and to safeguard it for the future. Posterity may justly accuse us of squandering the common wealth of raw materials, which we consume on a vast scale by combustion or by dissipation as rubbish over the Earth.

One of the many facets of this problem is the sheer volume of the rubbish and pollution which we produce in our modern society. The volume of this rubbish is growing virtually exponentially, partly because of the growth of population and partly because consumption is the basis of our twentieth-century prosperity. At the same time the biosphere — the thin shell on the Earth in which we live — grows no larger and natural biological regeneration is in fact deteriorating [1].

The terms "rubbish" and "pollution" should be interpreted in a sense wider than the purely material: they may be held to refer to any worsening of our environment. Apart from air pollution and water pollution, our environment is also threatened by excessive noise, vibration and the dissipation of heat. In the literature the terms noise pollution and thermal pollution have been introduced as counterparts to air pollution and water pollution.

In discussing such matters we should try to avoid a too emotional approach; we should rely instead on the evidence of sober figures. Unfortunately there is far too little factual material to permit of making absolute statements, but present indications are disquieting enough. On the basis of available knowledge, extrapolation of our present rate of pollution of the biosphere implies that sooner or later an end must come to the society of men on Earth. Clearly it is our responsibility to consider what action should be taken to prevent such an eventuality. If nothing is done, the end may be sooner rather than later.

To obtain some idea of how far things have gone let us take for our example the expectation of the world's future consumption of energy. It is common knowledge that the generation or, rather, conversion, of energy is almost invariably accompanied by some pollution or other. Thus an atomic-energy power station, even if no accidents occur, still gives rise to thermal pollution, a coal-fired power station leads to thermal and air pollution, a petrol or diesel engine produces thermal and air pollution and noise. The total consumption of energy in the world depends on the world population and on the amount of energy consumed on average per person. *Fig. 1* shows the growth of the world's population and the reduction in available living space per person as a function of time, up to the year 2000. *Fig. 2* indicates the expected annual consumption of energy in the world.

Thus we see that if no special measures are taken in the coming decades there will be an enormous increase in the consumption of energy. One of the great users of energy, and therefore one of the agents giving rise to pollution of the environment, is the motor-car, which in the United States, for example, accounts for 20%

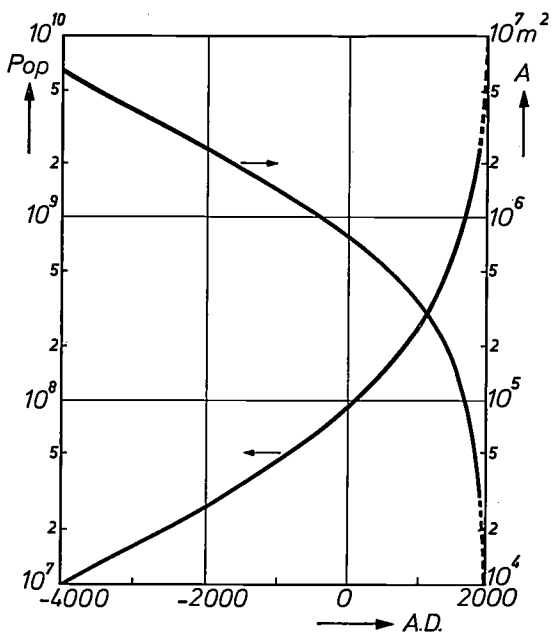


Fig. 1. The world's population (*Pop*) and the area (*A*) available on Earth per person, plotted against time [2]. If the present population explosion continues, the Earth will be inhabited by about 7 thousand million people in the year 2000, and the space available per person will have fallen to about 1 hectare (2½ acres).

Dr. Ir. R. J. Meijer is a Scientific Adviser with Philips Research Laboratories.

[1] See, for example, G. E. Hutchinson, The biosphere, Sci. Amer. 223, No. 3, 44-53, Sept. 1970 (issue on the grand-scale cyclic mechanisms of life on the Earth), and Committee on Resources and Man, Resources and man, Freeman, San Francisco 1969.

[2] Taken from P. S. Myers, Automobile emissions — a study in environmental benefits versus technological costs, paper 700182 of Society of Automotive Engineers, Inc.

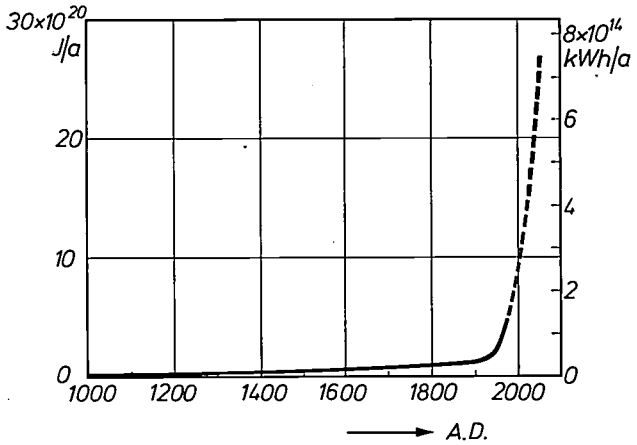


Fig. 2. The annual consumption of energy, plotted against time [3]. If the present increase continues, the consumption in the year 2000 will be about 9×10^{20} joule or 2.5×10^{14} kWh.

of the overall consumption of energy and for as much as 60% of the total air pollution produced there.

The air pollution caused by motor-cars poses a very complex problem. The main constituents of the exhaust gases considered to be noxious are: carbon monoxide (CO), oxides of nitrogen (NO and NO_2, generally indicated by NO_x), unburned hydrocarbons (C_xH_y), sulphur dioxide (SO_2), and in a certain sense also carbon dioxide (CO_2); also tiny solid constituents such as soot, and compounds of lead.

The abatement of air pollution caused by the engines of cars is being taken in hand by governments by the issuing of statutory regulations which lay down maximum amounts of the harmful substances that may be produced by a vehicle.

Though there is but little scientific knowledge about air pollution as such and about its consequences, it seems that care must be taken in the laying down of the maximum amounts, quite apart from economic reasons, because different pollutants can interact with each other. Thus the notorious smog is the result of a chemical reaction in which unburned hydrocarbons (C_xH_y), oxides of nitrogen (NO_x) and sunlight play an important role. Here circumstances can arise in which a reduction in the amount of NO_x causes the amount of smog to increase. This means that the specification of maximum permissible amounts is a matter of considerable subtlety.

Another example of a topic still somewhat controversial in scientific circles is the effect of carbon dioxide and of solid particles floating in the atmosphere on the average temperature of the Earth. The increase of CO_2, inherent in the combustion of fossil fuels, must on account of the "hot-house effect" cause the mean temperature of the Earth to rise. Yet this temperature has been falling during the last few decades; this fall is attributed to the greatly increased amount of dust in the atmosphere, which intercepts the solar radiation which would otherwise reach the Earth's surface and this effect prevails over the hot-house effect of CO_2.

It is against the background of the above problems and the statutory regulations that have been or will be taken that the properties of the Stirling engine will be described. As a prime mover for vehicles, the Stirling engine can make a contribution to the abatement of air pollution and noise; furthermore the Stirling engine can run on sources of heat other than those utilizing fossil fuels.

In the United States of America the first statutory measures were taken in the state of California, and these were later taken over by the federal government. For the 1970s a programme of specifications of gradually rising severity has been drawn up with which car engines must comply. A trend is furthermore discernible in the legislature to have these statutory requirements introduced at a faster pace. Though in Europe, too, some measures have already been taken, let us for greater clarity restrict ourselves to those of the United States.

In the drawing-up of statutory measures the problem was to lay down certain requirements quite unambiguously. For this purpose test procedures (cycles) representative of city traffic have now been established. *Fig. 3* represents the test prescribed by the U.S. Department of Health, Education, and Welfare (HEW) which has been taken over practically unchanged from the California test and which is applicable to the end of 1971. It is for the present restricted to unburned hydrocarbons (C_xH_y) and carbon monoxide (CO); in the future, oxides of nitrogen (NO_x) and solid particles will follow. The mean specific amounts produced by the car engine in the California test cycle — four times with a cold start and three times with warm starting — are expressed in grammes per mile.

In *Table I* we see in the first column the mean emis-

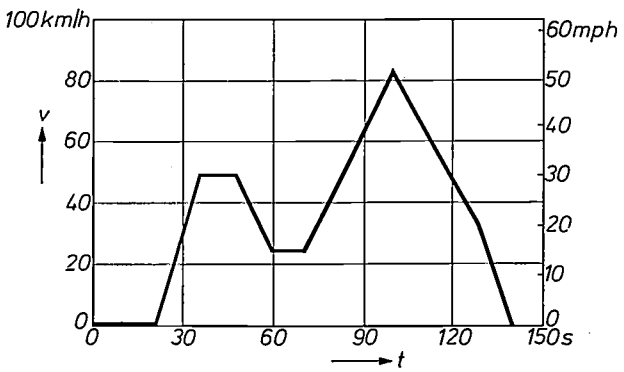


Fig. 3. In tests for the amounts of C_xH_y and CO in the exhaust gases of cars in accordance with the specifications now imposed by the U.S. Department of Health, Education, and Welfare (HEW) the velocity v of the car must be varied as a function of time in the manner here shown.

Table I. Exhaust gases of cars with petrol engines and a maximum weight of 2700 kg. The first column shows the situation in the United States of America in 1963. The other columns represent the specifications given in the years indicated, namely those of the Federal Government and those of the State of California [4]. The specifications of the Federal Government refer, in the years 1966-1971, to the test cycle of fig. 3, and those from 1972 onwards to the cycle of fig. 4. The California requirements all refer to the former cycle.

	1963		1966	1968	1969	1970	1971	1972	1973	1974	1975	1980	1980
hydrocarbons C_xH_y (g/mi)	5.7	Fed.	—	3.3	3.3	2.2	2.2	2.9+	2.9	2.9	0.5	0.25*	0.14**
		Calif.	3.4	3.4	2.2	2.2	2.2	1.5	1.5	1.5	0.5		
carbon monoxide CO (g/mi)	87.2	Fed.	—	34.0	34.0	23.0	23.0	37.0+	37.0	37.0	11.0	4.7*	6.2**
		Calif.	34.0	34.0	23.0	23.0	23.0	23.0	23.0	23.0	12.0		
oxides of nitrogen NO_x (g/mi)	5.8	Fed.	—	—	—	—	—	—	3.0*	3.0*	0.9*	0.4*	0.4**
		Calif.	—	—	—	—	4.0	3.0	3.0	1.3	1.0*		
particles (including lead) (g/mi)	0.3	Fed.	—	—	—	—	—	—	—	—	0.1*	0.03*	0.03**
		Calif.	—	—	—	—	—	—	—	—	—		
		Fed.	cycle in accordance with fig. 3					cycle in accordance with fig. 4					
		Calif.	cycle in accordance with fig. 3					cycle in accordance with fig. 3					

* Proposed.
** Research goals of National Air Pollution Control Administration (NAPCA) of U.S. Dept. of Health, Education, and Welfare.
\+ According to the Federal Register of 10 Nov. 1970 these values have recently been slightly increased (2.9 → 3.4; 37.0 → 39.0).

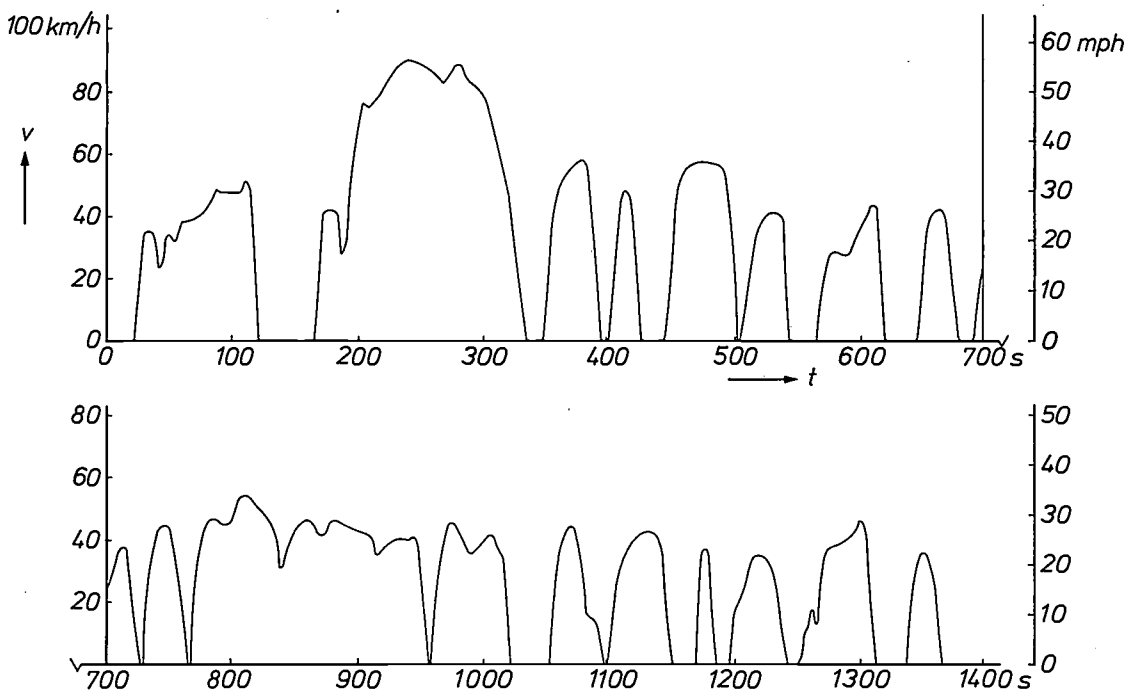


Fig. 4. As fig. 3, in accordance with the specifications imposed with effect from the beginning of 1972 by the U.S. Dept. of Health, Education, and Welfare (HEW). Later the amounts of NO_x and solid particles will also be measured.

sion prior to any measures against air pollution. In the column for 1973 we see for the first time a proposed specification for NO_x. In 1972 a new Federal test will apply to the whole of the United States, with much accelerating, braking, and idling over a total distance of 7.5 miles [4]. The entire test, including the cold start, lasts for 1370 seconds (*fig. 4*).

The reduction of the NO_x content to the values indicated in the proposal for 1980 will be the biggest obstacle for internal-combustion engines, and even for engines with external combustion it is not free from problems. In the case of the petrol engine the specifications of 1970 and 1971, in the matter of C_xH_y and CO, have been met mainly by tuning the engine to accept less rich mixtures; the temperature during combustion then becomes higher, so that less C_xH_y and CO, but more NO_x are formed.

[3] Taken from R. P. Hammond, Low cost energy: a new dimension, Science Journal **5**, No. 1, 34-42, 44, Jan. 1969.
[4] Control of air pollution from new motor vehicles and new motor vehicle engines, U.S. Dept. of Health, Education, and Welfare (HEW), Washington D.C., Federal Register **35**, 11334-11359, 1970 (No. 136, part II).

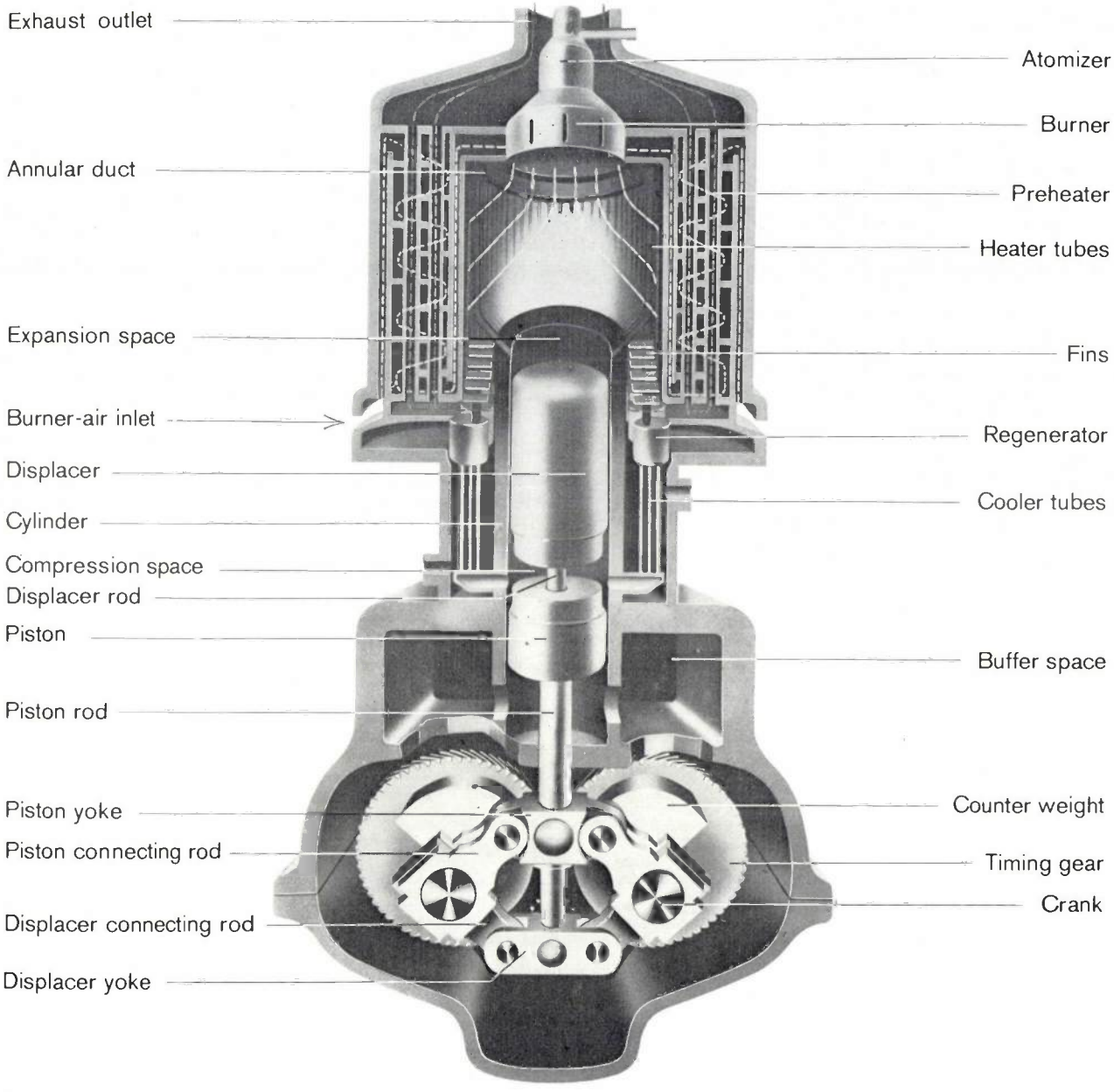


Fig. 5. Cross-section of a Philips Stirling engine. In the centre are the cylinder, piston and displacer. Below is the crankcase with the rhombic drive which governs the movement of the piston and displacer with respect to each other. The part above the cylinder is the heater. The air required by the burner enters at the place marked by the arrow and passes through a channel heated by the exhaust gases before reaching the flame. The exhaust gases escape at the top. It is equally possible to perform the heating by other methods.

The development of potentially cleaner engines for vehicles is being greatly stimulated by the U.S. Department of Health, Education, and Welfare. Thus a programme has been drawn up in which the motor industry is being activated to develop engines capable of complying with the severest emission tests, and in which attention is being paid also to mass production, price, safety, reliability, and fuel consumption [5]. By 1975 it is hoped to have made the selection from the alternatives so that industry can proceed with development, with or without government backing.

The exhaust gases of the Stirling engine

With the Stirling engine (*fig. 5*) the exhaust gases are relatively very clean even if fossil fuels are used in a normal — that is to say adiabatic — burner. Because of the continuous combustion of the fuel in a space surrounded by hot walls, the latitude in the choice of the air-to-fuel ratio and the considerable freedom in the design of the burner, the amount of unburned hydrocarbons becomes virtually negligible and the amount of carbon monoxide is very low. Strong preheating of the combustion air does, it is true, lead to a high flame

temperature, which promotes the formation of oxides of nitrogen. However, the combination of relatively brief residence time, lower peak temperatures than in internal-combustion engines, and the continuous combustion does lead to quite a small value for NO_x. Nevertheless this would in the long run still be a handicap if nothing further were done. On account of the external heating of the Stirling engine it is possible to incorporate modifications of the heater system without affecting the Stirling system, i.e. without diminishing the efficiency and specific power. The present-day efficiency of the engine is obtained at a temperature of the heater wall of only about 700 °C, so that the high adiabatic flame temperatures of more than 2000 °C are not essential (in contrast to engines with internal combustion, in which the highest temperature attained in the cylinder is decisive for the efficiency and power).

It is rather difficult to measure directly the effect of the flame temperature on the amount of NO_x formed, but it can be done quite well indirectly. For this purpose an electrically operated preheater was constructed for a 90 hp single-cylinder engine equipped with a normal adiabatic burner [6]. The production of nitrogen oxides was measured as a function of the temperature of the combustion air. The curve of *fig. 6* shows that the nitrogen-oxide content drops considerably as the air temperature is decreased. Work is going on at present with burners having a lower flame temperature than adiabatic burners, the so-called suppressed-flame-temperature burners, which function in combination with heat pipes (see below) to make the amount of NO_x negligibly small. Another method readily applicable to the Stirling engine for reducing the nitrogen oxide content is to recirculate a fraction of the flue gases to the fresh combustion air (*fig. 7*). With good mixing this does not lead to an appreciable change in the amounts of CO and C_xH_y.

Table II shows the test results obtained with the above 90 hp Stirling engine. Tests recently performed with a 10 hp engine gave corresponding results. In addition, published values for motor-car gas turbines (which also give clean exhaust gases) are included in this table. Because of the much greater excess of air and the poorer efficiency of the gas turbine in comparison with the Stirling engine the amount of exhaust gas per hp of the gas turbine is about eight times that of the Stirling engine. Hence a better comparison is obtained by calculating how many milligrammes of a

[5] Federal Clean Car Incentive Program, Report HEW National Air Pollution Control Administration (NAPCA), Oct. 1970, Attachment D.

[6] R. J. Meijer, The Philips Stirling engine, Ingenieur **81**, W 69-79, W 81-93, 1969 (Nos. 18, 19).

[7] Taken from Study of unconventional thermal, mechanical, and nuclear low-pollution-potential power sources for urban vehicles, HEW/NAPCA, Raleigh N.C., Oct. 1969.

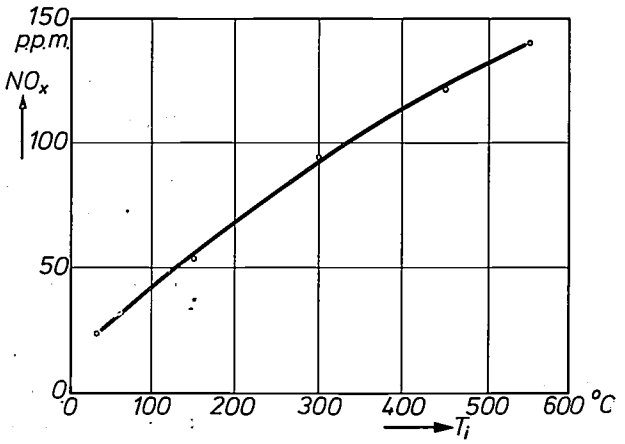


Fig. 6. The concentration of NO_x in the exhaust gases of a Stirling engine as a function of the temperature T_i of the combustion air.

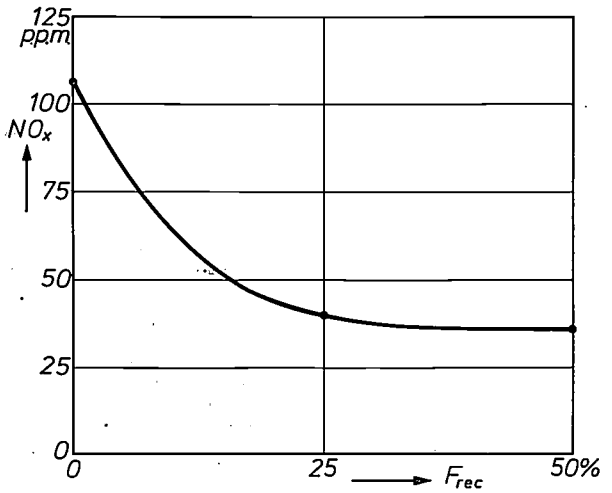


Fig. 7. The effect of recirculation on the amount of NO_x in the exhaust gases of a Stirling engine. The abscissa represents the fraction F_{rec} of the exhaust gases flowing back to the inlet. If F_{rec} is made 25% or more, the content of NO_x becomes three times smaller than that without recirculation.

Table II. The contents of C_xH_y, CO and NO_x in the exhaust gases of a Stirling engine. Excess air 40%. In one of the tests, 25% of the exhaust gases have been passed back and mixed with the fresh combustion air (recirculation; cf. fig. 6). By way of comparison, the values found in the case of a gas turbine for private cars [7] are also given.

	Stirling engine (ppm)	Gas turbine (ppm)
C_xH_y	1-2	1.5
CO	70-300	200-500
NO_x (adiabatic burner)	100-200	90-250
(with recirculation)	about 40	—

Table III. As Table II, but now calculated per horsepower and expressed in mg/s.

	Stirling engine	Gas turbine
C_xH_y	$3\text{-}6\times10^{-3}$	36×10^{-3}
CO	0.1-0.3	2.0-3.6
NO_x (adiabatic)	0.1-0.2	0.7-2.0
(25% recirc.)	0.04	

certain compound are emitted per second per hp. *Table III* shows the values found at full load, for which purpose the oxides of nitrogen and the unburned hydrocarbons are assumed to be respectively NO and C_6H_{14}.

It is not possible directly from these tests to express the emission of the Stirling engine in grammes per mile, as prescribed by the HEW test (which will apply up to the end of 1971). In the Federal test to come into force thereafter the emission must also be expressed in grammes per mile. For this purpose the engine must be installed in a vehicle or, if measurements are made with the engine in the test bed, actual conditions must be simulated as closely as possible.

However, certain estimates can be given: the first calculation, applicable to the present-day HEW test (fig. 3), can be performed because of the indication that the C_xH_y content, expressed in terms of C_6H_{14}, corresponds in the case of a large American car to 180 ppm [8]. Assuming that the excess air of a Stirling engine is 40% larger than in the petrol engine and that the efficiency is 1.4 times that of the petrol engine in this test, the following results are obtained:

C_xH_y	0.02 g/mile
CO	1.00 g/mile
$NO_{adiabatic}$	0.8 g/mile
$NO_{25\%\ recirculated}$	0.16 g/mile

The new HEW test programme (fig. 4) applicable from the beginning of 1972 is very complicated. To obtain an idea of what the emission of the Stirling engine would be if it were built into a car of 1800 kg and subjected exactly to this new test, calculations were made of the engine power required to overcome friction in the transmission, air and rolling resistance, and for accelerating and decelerating, allowance being made for the effect of power variations on the fuel consumption. *Fig. 8* shows the requisite crankshaft power generated in such a test, assuming a loss of 20% in the transmission. For calculating the emission per mile, the following further assumptions have been made: average excess air 60%; efficiency of Stirling engine (average over the entire test) 30%; C_xH_y 2 ppm; CO 300 ppm; NO_{ad} 200 ppm; $NO_{25\%\ rec}$ 40 ppm.

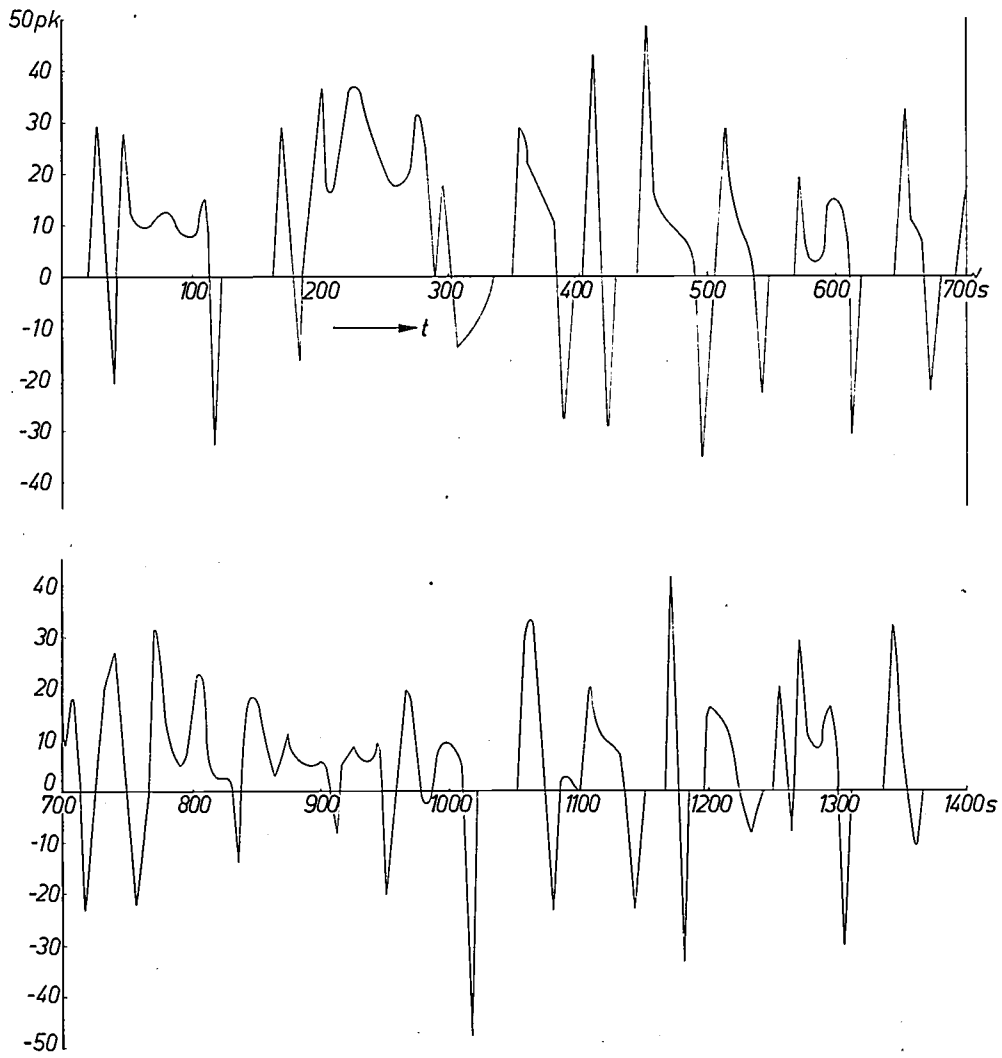


Fig. 8. Calculated variation, as a function of time, of the engine power (in hp) of a car weighing 1800 kg undergoing the test of fig. 4. The efficiency of the transmission is taken as 80%.

The results are shown in *Table IV*, which also gives the proposed Federal requirements for 1980 as well as the NAPCA research goals for 1980 (cf. Table I). A remarkable fact is that these two entirely different calculations, applied to entirely different tests, lead to practically the same results.

The numerical values in respect of the Stirling engine do not include cold starts. Measurements have shown that these would slightly increase the values of C_xH_y and CO but that the value for NO would undergo practically no change. The latter is not surprising, since, during warming up, the mean flame temperature is lower than when the air has been fully preheated (cf. fig. 6).

We have seen that with regard to its emission of exhaust gases the Stirling engine is very attractive. The question now arises as to how far the engine is suitable for vehicle propulsion in other respects.

Table IV. Exhaust gases, in grammes per mile, of a Stirling engine used for traction in a car of 1800 kg subjected to the test applicable in the United States of America at the present time and to the test that will apply from the beginning of 1972 (cf. fig. 3, fig. 4, and Table I). The third and fourth columns show respectively the requirements for 1980 and the NAPCA research goals.

	Stirling engine		1980 require-ments	NAPCA research goals 1980
	HEW 1970	HEW 1973		
C_xH_y	0.02	0.03	0.25	0.14
CO	1.00	1.4	4.7	6.2
NO_x (adiab.)	0.8	1.0	0.4	0.4
(recirc.)	0.16	0.20		

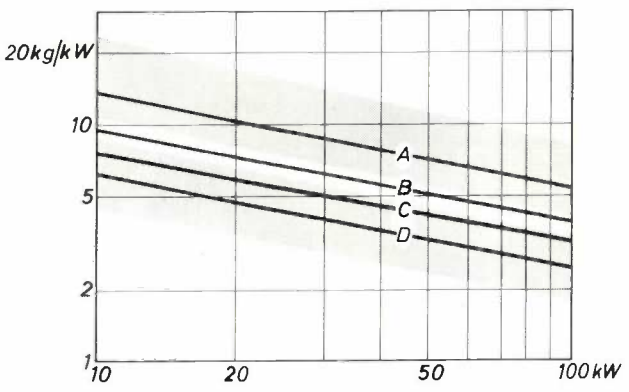


Fig. 9. The specific weight of the Stirling engine as a function of the shaft horsepower. *A* mean values of present-day laboratory models, with helium as working gas. *B*, as *A*, but with heaters permitting twice the working pressure. *C*, as *B*, if the design is based on the specific requirements of motor-cars. *D*, as *C*, but for an engine with hydrogen as the working gas. The curves have been calculated for a heater temperature of 750 °C and a cooling-water temperature of 55 °C. For petrol engines the corresponding curves lie in the lower grey band, for diesel engines they lie in the upper grey band.

The Stirling engine as a vehicle engine

Specific weight

One of the most important characteristics of a vehicle engine is the specific weight, i.e. the weight per kW or horsepower. In *fig. 9* the specific weight of American diesel and petrol engines is plotted against the shaft horsepower of the engine [9]. This figure shows also lines representing the Stirling engine. The upper line, *A*, represents the mean values of present-day laboratory models. The drives of these engines have been designed for twice as high a pressure and thus for twice as high a value of the shaft horsepower as that indicated. The heater material, however, does not yet have an adequate creep strength at such high pressures. When this material has been replaced by one of greater creep strength it will be possible to achieve this higher pressure in the engine for longer running periods. We then obtain line *B* for the Stirling engine. A reduction of 20% is expected if the engines are designed specifically for motor vehicles, so that provisionally the mean specific weight of the Stirling engine in a few years' time can be represented to a good degree of approximation by line *C*. (We shall see below that there are possibilities of reducing the specific weight even further without effect on the efficiency.)

Efficiency and other properties

The efficiency of the engine depends on a number of factors, such as specific weight, heater temperature, cooling-water temperature, and configuration. Line *C* of fig. 9 applies to engines with an efficiency of about 35% at a heater temperature of 750 °C and a temperature of the cooling water of 55 °C. This line will be displaced upward or downward, depending on whether high specific power or high efficiency is the primary aim.

Quite briefly, the following properties of the Stirling engine are also attractive in its application as a vehicle engine:

1. High efficiency at partial load.
2. Low noise production and low vibration (20 to 40 dB lower than with corresponding diesel engines).
3. Engine braking is possible (negative torque up to about 80% of full-load torque).
4. Wide speed range and very favourable torque characteristic and tangential-effort diagram.
5. No oil consumption and very infrequent oil changes.
6. Reliable starting, long service life.
7. Rapid power variation.
8. Very insusceptible to dust from the environment.
9. External heating (heating possible by means other than hydrocarbons).

[8] R. R. Allen and C. G. Gerhold, Catalytic converters for new and current (used) vehicles, paper read to the Fifth Technical Meeting, West Coast Section of the NAPCA, Oct. 1970.

[9] J. H. B. George, L. J. Stratton and R. G. Acton (Arthur D. Little, Inc., Cambridge, Mass.), Prospects for electric vehicles, paper prepared for HEW/NAPCA, May 1968.

In all these respects the Stirling engine is a very attractive proposition as a vehicle engine. It has only one, though surmountable, objection: the cooling water has to dissipate a rather large amount of heat at the lowest possible temperature. Optimization calculations of the entire system, i.e. engine with all auxiliaries and radiator, have shown that, in the case of an ambient temperature of 20 °C, the radiator must have a thermal dissipation power about 2½ times greater than that of a diesel engine. This is admittedly awkward, but not an insurmountable engineering problem.

Further research and development

In the enumeration of the properties of the Stirling engine nothing has been said about the price of the engine. The Stirling engine is certainly dearer than the petrol engine, but here it should be noted that the price is not the only factor determining the suitability of an engine; the diesel engine, too, is more expensive than the petrol engine. If it is furthermore remembered that in the United States the contribution of road vehicles to air pollution is about 60%, of which about 85% is attributable to the petrol engines of private cars, it is obvious that the Stirling engine in spite of its somewhat higher price can satisfy a demand as far as passenger vehicles are concerned.

What means are available in principle to make the engine cheaper? In the above considerations we have, of course, assumed that in developing our research models for large series or mass production the engine has been made cheaper: by using materials of high creep strength and other expensive materials only where strictly necessary, by shrewd design and fabrication methods. Essentially a lowering of the price, with retention of the favourable properties, including the high efficiency, implies in mass production a reduction of the specific weight or, which amounts to the same thing, an increase in the specific power.

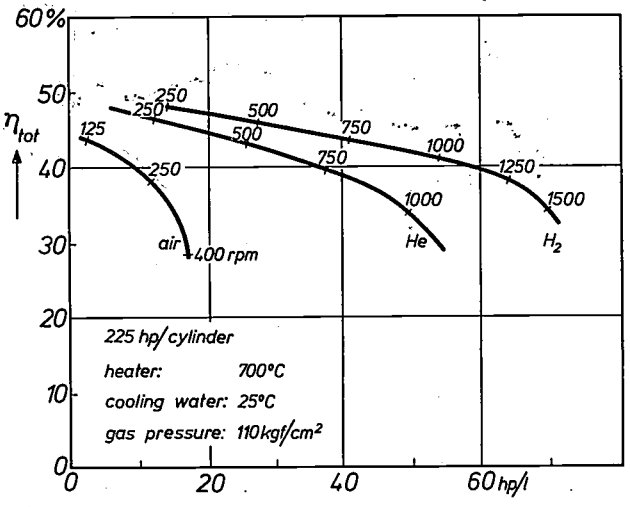
In the first few years after the invention of the rhombic drive all endeavours were directed to obtaining a reliable, robust engine of high efficiency. The application then primarily envisaged was the propulsion of boats, and the specific weight of about 6.8 kg/kW (5 kg/hp) achieved corresponded well to that of conventional boat engines. When in the sixties it became ever clearer that two features of the engine, namely clean exhaust gases and low noise, were becoming more important on account of increasing concern for our environment, more attention was given to the feasibility of the Stirling engine as a vehicle engine. Consequently our attention has become more and more directed to raising the specific power.

Quite generally there are three possibilities of raising the specific power: increasing the pressure of the engine, increasing the speed of the engine, and modifying the configuration.

The first point has already been discussed in the considerations concerning the specific power. The two other points will now be treated in more detail.

Towards higher speeds; the heat pipe

If a Stirling engine of a certain configuration is optimized for a relatively high speed, the efficiency will be less than if the engine is optimized for a lower speed. In *fig. 10* we see how the maximum attainable efficiency of a single-cylinder rhombic-drive engine with a nominal output of 225 hp depends on the specific power and the working medium. The manner in which the curves change shows that with the same working medium a rise in the specific power is accompanied by an increase in the speed (rpm) but also by a fall in the efficiency.



volume) in the case of three different working gases. A higher specific power requires a higher speed and leads to a fall in η.

The use of hydrogen instead of helium results in an appreciable increase in the specific power at the same efficiency. The snag is, however, that hydrogen slowly diffuses out through the hot parts of the heater, which means that in this respect the engine has a slow leak. There are applications where this is a serious objection, but in the case of vehicle engines this slight loss of hydrogen appears to be of little importance, so that, for traction purposes, hydrogen can certainly be used as the working medium. This will lead to a further 30% fall in the specific weight (fig. 9, line *D*).

If we now look at the Stirling system more closely we find that the limitation of the speed is due mainly to the form of the heater assembly. In spite of the large difference in temperature between the flame gases and

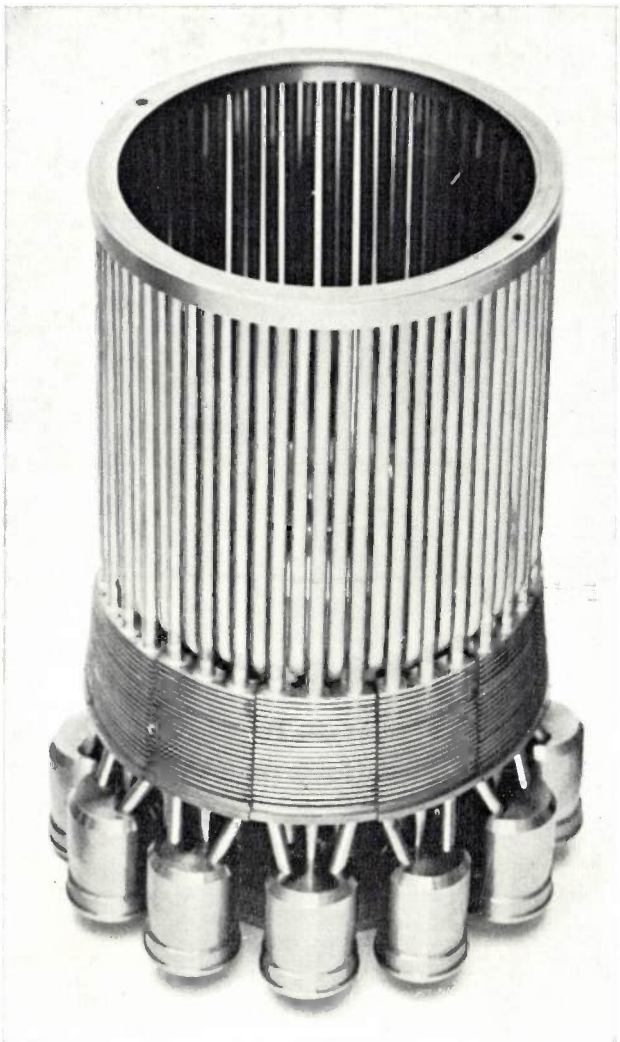


Fig. 11. Design of the heater of a Stirling engine heated by flame gases. The regenerator is distributed over the cylindrical canisters around the lower end. Three tubes leave each section and extend upwards to form the vertical tubes of the heater. Alternating with them are the down-going tubes coming from the annular duct at the top and discharging into the expansion space of the engine. At the lower end of the tubular wall thus formed fins are brazed to improve the heat transfer between the combustion gases and the tubes.

the heater tubes the heat transfer from the flame gases to the tube wall is relatively poor, so that a rather large tube surface is required. Large tubes (large surface area) giving optimum heat transfer on the outside do not, however, give an optimum heat transfer on the inside. The heater assembly is therefore a compromise, see *fig. 11*. In order to improve the heat transfer of the flame gases the heater is made in the form of a cage of tubes, to which fins are brazed at the lower ends. The more the specific power is raised, the worse the compromise becomes. It would be much better if the Stirling system could be optimized without any need to take the transfer of heat on the outside into account. This becomes possible if we make use of *indirect* heating. A very suitable system to realize this is the so-called heat pipe, by which large amounts of heat can be transferred from a large surface to a small surface with a very small difference in temperature.

In principle a heat pipe consists of a hermetically sealed chamber, the inside walls of which are provided with a lining of porous material in which a liquid is absorbed by means of capillary forces. The simplest form is a sealed pipe (*fig. 12*). For the temperatures of interest for the Stirling engine (700 to 800 °C) sodium is a suitable transport medium. On local heating of the wall, part of the sodium evaporates, absorbing heat. On account of the difference in pressure the vapour flows to the colder area and condenses on the unheated surfaces, giving off the amount of heat previously absorbed. The liquid thus formed flows back again, under the influence of capillary forces, to the area of evaporation. Thus a cycle is set up in which sodium goes successively through the vapour and the liquid phase and in which large amounts of heat can be transferred with very slight differences in temperature. In comparison with the thermal conduction of a copper rod of the same dimensions the flow of heat through a heat pipe can, with the same difference in temperature, be several thousand times greater. The heat-pipe system can in principle be regarded as a transformer of the heat-flux density. For example, a large area can be heated with a low heat-flux density, and the amount of heat taken up can then be released at a high heat-flux density, during condensation, to a small surface. This property is precisely what we need in the heating of the Stirling

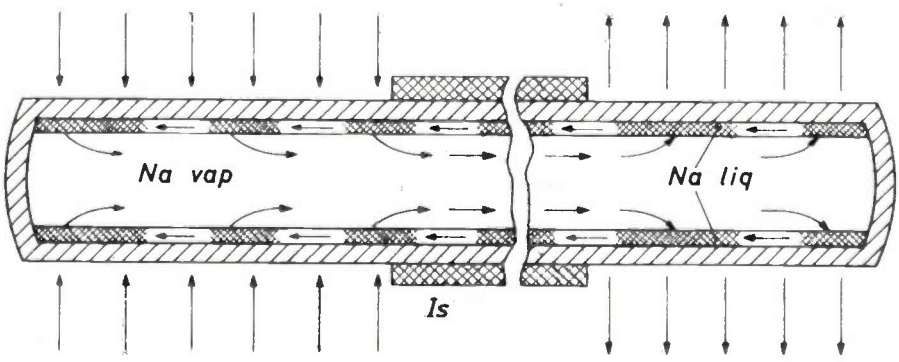


Fig. 12. Schematic cross-section of a heat pipe. The inner wall is lined with porous material. If one of the ends (here on the left) is heated, then the heat transfer medium (here sodium) melts and evaporates. The vapour (*Na vap*) flows to the cool end (right) and condenses there, during which process heat is given up. The condensed sodium (*Na liq*) flows back under the action of capillary forces in the porous lining to the warm end. Between the places where the pipe takes up and gives up the heat it is surrounded by a layer of insulating material *Is*.

engine, whether the heating is by flame gases or by other heat sources (see title photograph, p. 168). We shall return to this point later. Using such a heat pipe the heat source and the Stirling system can be separated, and four important advantages accrue:

1. Since the heat transfer by means of the condensation of sodium can be considered infinitely great in comparison with heat conduction through the wall of the tubes and the transfer of heat from the walls to the gas inside, the Stirling system can be optimized for a constant temperature of the outside wall of the heater tubes.
2. Since the heat pipe can act as a transformer of the heat-flux density we can make the heat-transfer surface subject to the flame gases large, so that a high burner efficiency can be obtained.
3. Since the heat-pipe system functions practically isothermally (absence of hot-spots) the average temperature of the pipe can be 50 to 75 °C higher for the same life; accordingly the power and the efficiency can again be considerably raised.
4. Because the heat-transfer surface at the flame-gas end can be made arbitrarily large it is possible to work with a smaller temperature difference between flame gases and wall while yet retaining a high burner efficiency; hence it is possible in principle to work with a lower flame temperature than in the "adiabatic" case — the so-called "suppressed-flame temperature" — allowing of drastic reduction of the NO_x content.

Fig. 13 shows a schematic diagram of a system of indirect heating by means of flame gases.

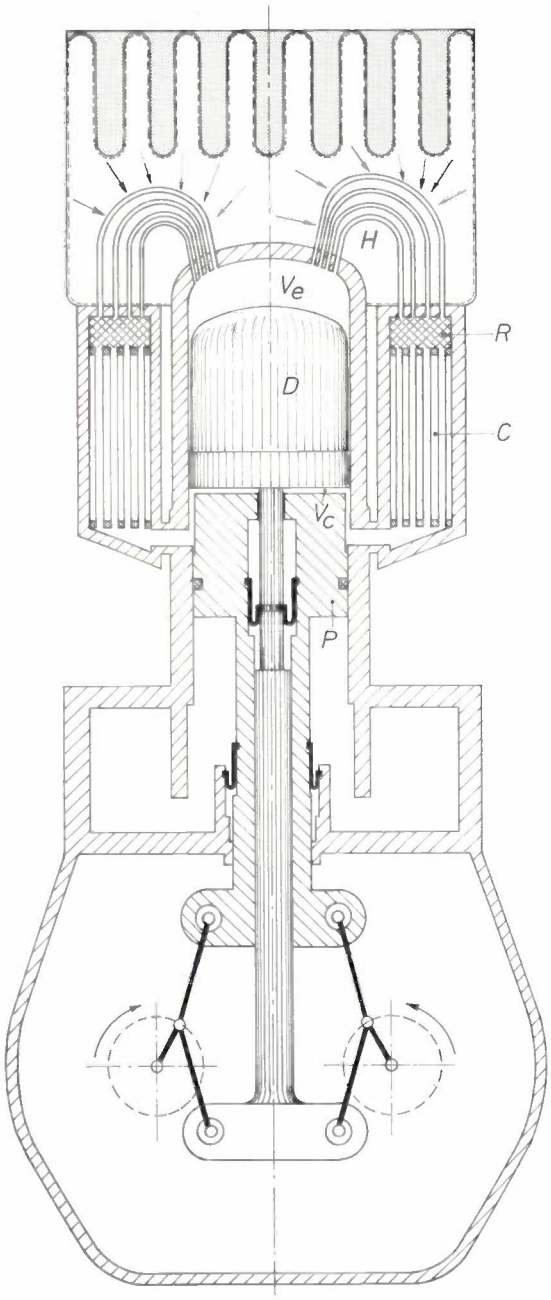


Fig. 13. Cross-section through a Stirling engine with a system of indirect heating by flame gases via a heat pipe (above). The broken line, drawn just inside the wall, represents the porous lining. The wall is heated by flame gases flowing through the shaded spaces. The sodium vapour condenses on the pipes *H*, which connect the expansion space V_e of the engine with the regenerator *R*. Below the regenerator is the cooler *C*, which is connected to the compression space V_c. *P* is the piston. *D* is the displacer. Below, the crankcase and the rhombic drive.

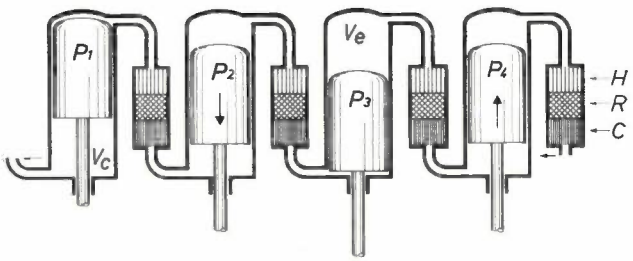


Fig. 14. Schematic representation of a double-acting four-cylinder Stirling engine. The symbols have the same meaning as in fig. 13. The open end, on the right below *C*, must be thought of as connected to the opening on the left of the left-hand cylinder. Here there are no separate pistons and displacers; the expansion space of one cylinder is connected via *H*, *R* and *C* to the compression space of the next one.

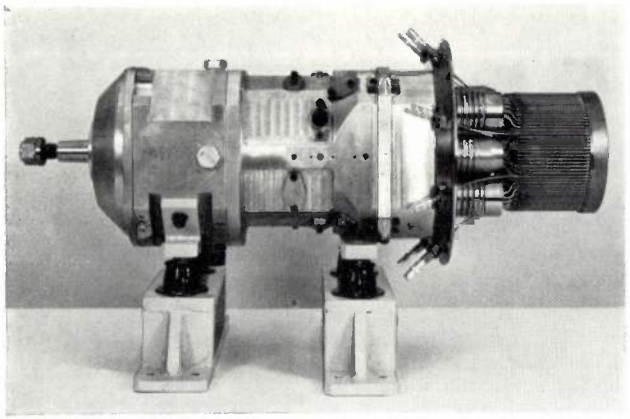


Fig. 15. A second-generation Stirling engine. Double-acting four-cylinder engine with direct heating; the burner cage is on the right.

Other engine configurations

After 1945 intensive work was done on the double-acting hot-air engine. In principle this engine is very simple, because various functions of the Stirling system can be combined (*fig. 14*). One of the main advantages at that time was that the crankcase did not have to be under pressure, so that this cleared the way to bigger engines. Owing to enormous difficulties, then quite insurmountable, this promising hot-air engine could not at the time be realized. When the rhombic drive was invented in 1953 it was possible to go back once more to the displacer principle, in which the functions

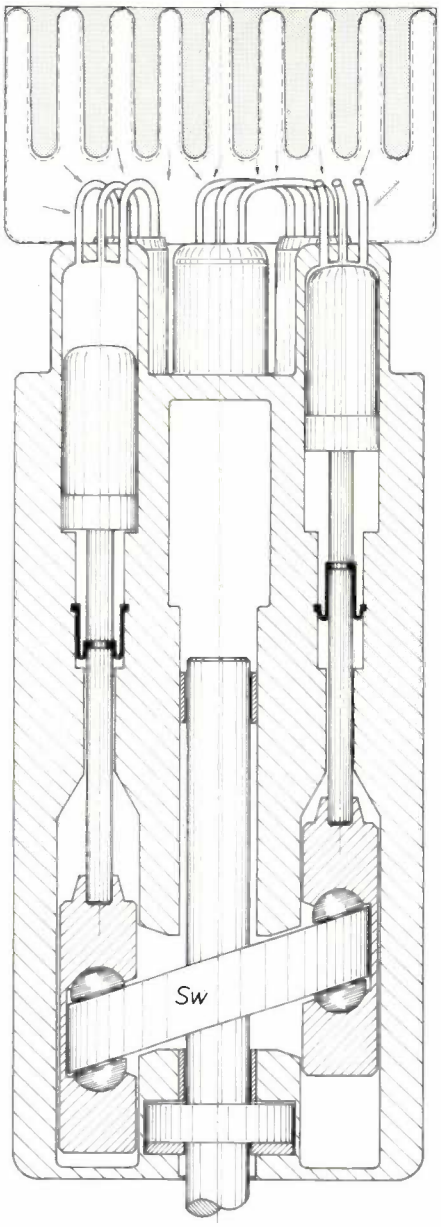


Fig. 16. Engine of the type of fig. 15, but now equipped with a system for indirect heating (cf. fig. 13). Two of the four cylinders are shown in cross-section, and in the centre the third cylinder can be seen. The cylindrical enclosures behind the first two cylinders contain the regenerators. In these engines the movement of the pistons is transmitted to the main shaft by a swash-plate *Sw*.

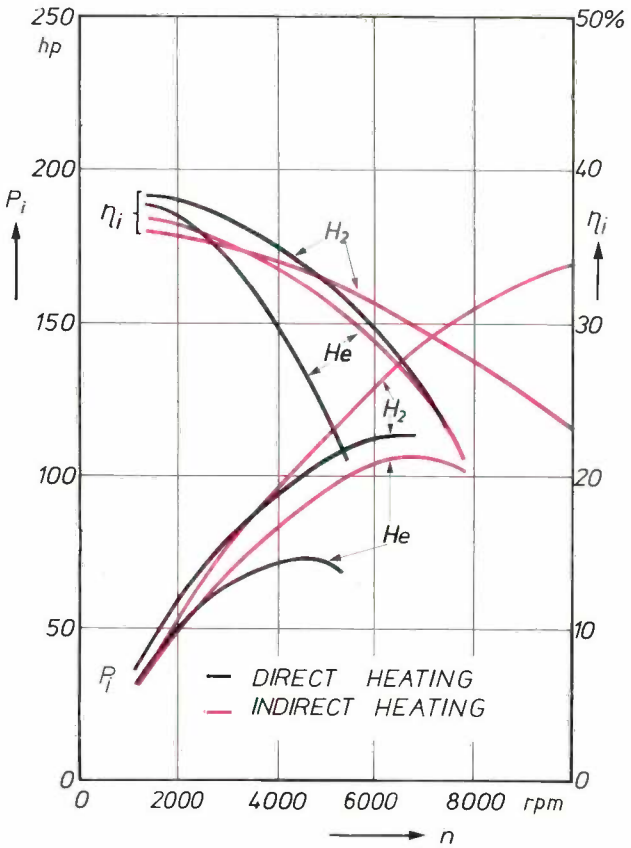


Fig. 17. The indicated power P_i and the indicated efficiency η_i, of the engine of fig. 15 as a function of the speed (rpm) n for two different working gases. The black lines refer to direct heating of the engine, the red lines to indirect heating (fig. 16). Indirect heating is much more advantageous because it permits of higher engine speeds.

are separated, while retaining such advantages as a pressureless crankcase and complete balancing, but at the cost of the compactness of the double-acting engine. But if we now strike a balance of the problems which at that time were insoluble we find that many of these problems have in fact been solved during the development of the engine with rhombic drive. This, in view of the desire to achieve a higher specific power, justified a cautious resumption of research into the double-acting Stirling engine. We call this type a "second-generation" Stirling engine (*fig. 15*). Research is in progress and initial successes are in sight.

In order to run this four-cylinder engine on one burner, heated directly by flame gases, the four heaters are combined into a single heater cage, but the systems are still completely separate (though this cannot be seen from the outside). It is interesting, on the basis of our calculations of this engine, which was designed specially for a high specific power, to demonstrate the effect of indirect heating (*fig. 16*). From *fig. 17*, in which the calculated indicated efficiency and indicated power are plotted, the difference can be seen between helium and hydrogen as the working medium of the engine and the

difference between direct and indirect heating. Indirect heating shifts the maximum power towards higher engine speeds and raises the power enormously. It should be noted that the volume of the engine, apart from the preheater, is the same for all the calculated curves. On the basis of tests and calculations it is to be expected that with this type of engine for this power range, a specific weight (including auxiliaries) of less than 1.5 kg/kW or 1.1 kg/hp will be attainable.

From the foregoing we may conclude that the Stirling engine is a technically very interesting engine for traction purposes and that it is potentially possible to lower the cost price by increasing the specific power. Yet it is easy to imagine that a company considering venturing into the field of the Stirling engine will want to know, before making large investments, what possibilities the Stirling engine has to offer if, for whatever reason, fossil fuels are no longer available.

In summarizing the various properties of the engine, mention was made of the feature of external heating, which makes various heating systems possible. We are therefore not restricted to the combustion of various types of liquid and gaseous hydrocarbons, as our demonstration model (*fig. 18*) may perhaps wrongly suggest: quite generally, the salient feature of the Stirling engine is that the type of heat source is irrelevant as long as the heat is supplied at a sufficiently high temperature. And with the intervention of the heat pipe, the applicability of the Stirling engine has become still more general.

Below we shall discuss two methods of heating which may become interesting in the future in view of air pollution or a possible scarcity of fossil fuels: heating by means of a "heat accumulator" [10], and heating by the combustion of hydrogen, stored in a hydrogen accumulator.

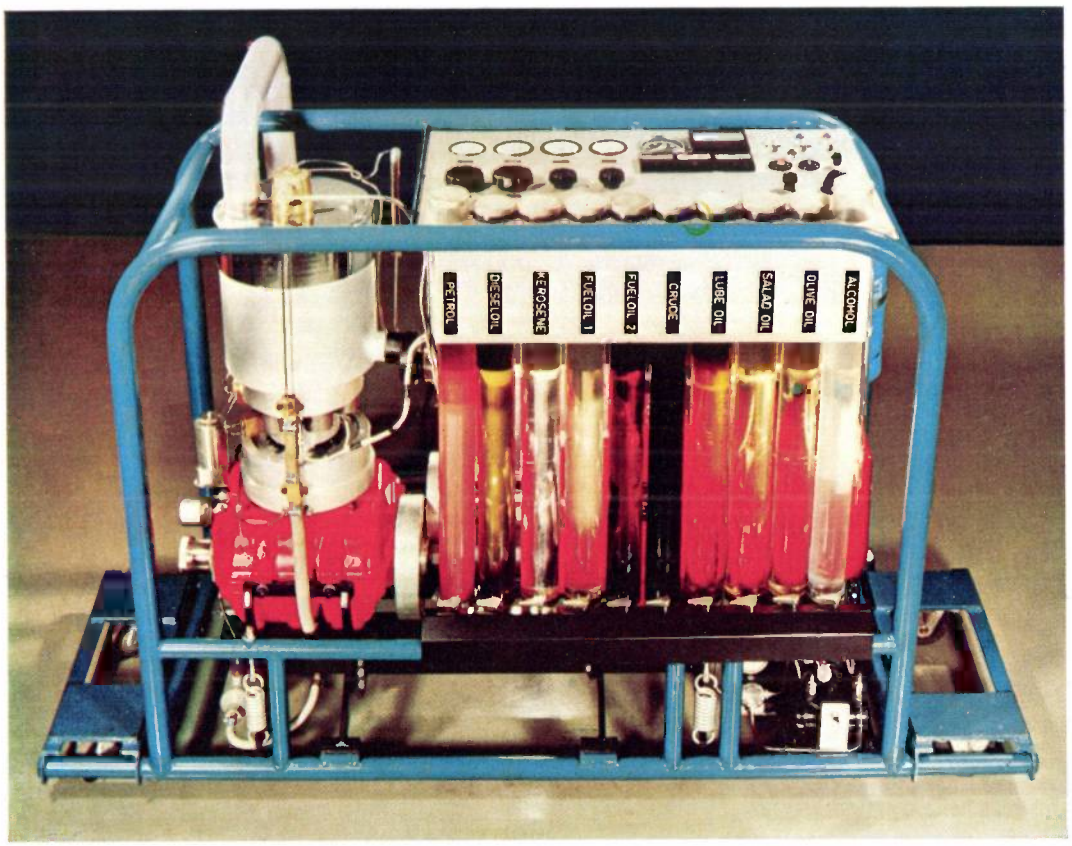


Fig. 18. Laboratory model for demonstrating that a Stirling engine will run on the heat obtained from all manner of fuels, here ten different liquids.

New sources of heat

The heat accumulator

Essentially a heat accumulator consists of a container which is filled with a material capable of absorbing and releasing large amounts of heat. This heat can be given up to the heater of the engine, for example via a system of heat pipes. We have done most of our experiments with lithium fluoride as the heat-accumulation material and what follows will be restricted to that material.

[10] See also R. J. Meijer, Mit Elektro-Wärmespeicher und Stirlingmotor — eine mechanische Antriebsalternative, Denkschrift 11/1969 Deutsche Forschungsgemeinschaft: Elektrospeicherfahrzeuge, pp. 143-164.

Lithium fluoride is a chemically very stable salt with a melting point of 848 °C, a heat of solidification of 250 kcal/kg, a mean specific heat of 0.56 kcal/kg°C between 550 °C and 848 °C, a density at 870 °C (liquid) of 1.79 g/cm³, and a density at 700 °C (solid) of 2.64 g/cm³. If the heat of solidification and the sensible heat obtained on cooling to 550 °C are passed to the engine, account being taken of the efficiencies corresponding to these temperatures, we find a specific *mechanical* energy of about 200 Wh per kilogramme of lithium fluoride (*fig. 19*). During the last ten years superinsulation materials for these temperatures have been developed in the United States to enable radio isotopes to be used in space, and these are just the thing for our purposes.

In order to transport the heat from the accumulator to the engine we make use, as already mentioned, of a system of heat pipes, and their property of also being a transformer for the heat-flux density is here very useful. The thermal conduction of solidified LiF is not very high, so that a relatively large wall surface with a small heat-flux density is necessary in order to avoid an excessive temperature gradient in the solidified salt.

Fig. 20 shows schematically an experimental set-up of heat accumulator, heat-pipe system, and Stirling engine. The heat accumulator actually consists of an enclosure containing small thin-walled sealed elements filled with lithium fluoride. In the liquid state the salt occupies practically the whole volume, a small volume above the molten salt being occupied by argon at a pressure such that, when the salt solidifies, there is always a slight overpressure inside the element to prevent it from collapsing. The broken lines in fig. 20 indicate the porous lining through which the sodium, the heat-transfer medium, flows back as liquid to the site of evaporation. The operation of the system is as follows: let there be a kind of valve at *A* which seals off the connection between the accumulator and the engine. If the bottom of the large container is electrically heated, sodium evaporates and condenses on the elements, where the heat of condensation is given up. The liquid sodium is led back to the bottom by the porous lining. This process can go on until all the salt in the elements has been melted, which is indicated by an increase in the rate at which the temperature rises. The heat accumulator is now "charged" and the supply of electrical energy can be shut off. If we wish to start the engine, we open the "valve" at *A*, so that the sodium

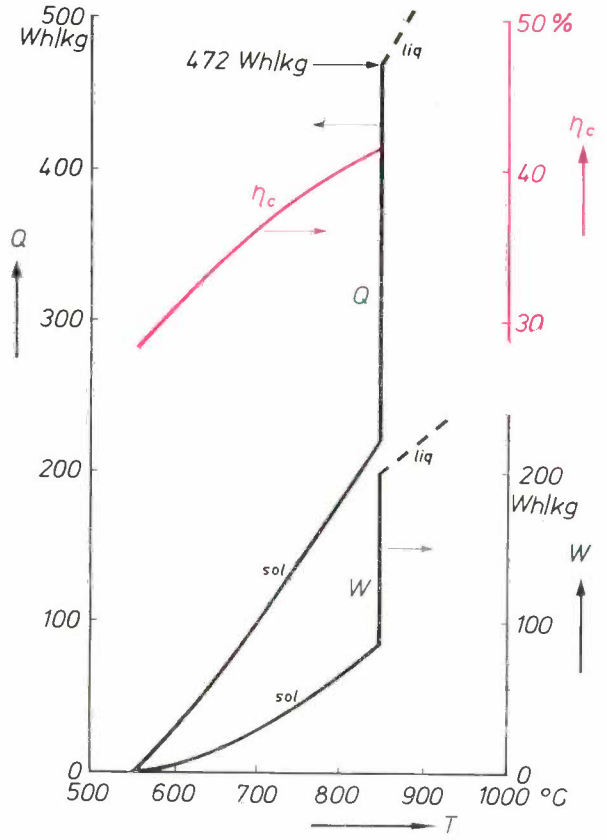


Fig. 19. The energy content Q per kilogramme of lithium fluoride as a function of the temperature T. At 848 °C the salt melts, and the (large) heat of fusion is taken up (vertical part of the curve). When all the lithium fluoride has melted, Q amounts to 472 Wh/kg. W is the mechanical energy obtainable from Q, η_c is the conversion efficiency.

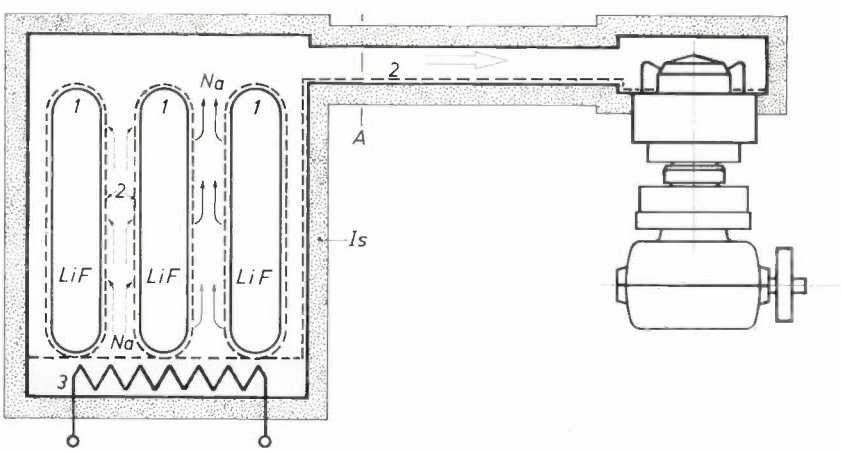


Fig. 20. Schematic representation of a Stirling engine (right) connected via a heat pipe to a heat accumulator (left). *1* closed cells (elements) filled with LiF. *2* porous lining. *3* electric heater. *Is* insulation. *A* isolating valve in heat pipe. During charging, heat is supplied until all the LiF has melted. The walls of the elements then act as the cold end of a heat pipe and the electric heater as the hot end. When *A* is open, the elements form the warm end of the heat pipe which connects the accumulator to the engine (arrows).

vapour condenses on the heater tubes and the liquid flows back again to the outer surface of the elements, which now constitute the evaporating part of this heat-pipe system. *Fig. 21* is a radiograph of a test element surrounded by a system of heat pipes. The salt is here in the solidified state. *Fig. 22* shows a test set-up of the entire system in accordance with fig. 20.

It is an interesting exercise, with our present data, to work out the size and weight of a motor vehicle equipped with the above system if it is specified that the accumulator is charged only once every twenty-four hours and that the car's radius of action is the same as that of a petrol-engined car.

For this calculation we use the data of *Table V*, taken from the previously-mentioned study by Arthur D. Little, Inc. for the U.S. Department of Health, Education, and Welfare in respect of six types of cars. This study includes, in addition to present-day vehicles, also so-called lightweight cars which, it is assumed, it will be possible to design in the future.

The principal assumptions which we have made for these calculations are: the specific power lies in the region between lines *C* and *D* of fig. 9; the loss of heat in twenty-four hours is 12% of the maximum amount of stored heat (half of this is conducted away by the insulation, and the remainder is lost via supporting and connecting pieces); the insulation thickness of the superinsulation material of medium quality is calculated for a hot-face temperature of 850 °C and an ambient temperature of 20 °C as being

$$\lambda = 1.4 \times 10^{-5} \text{W cm/cm}^2 \text{ °C};$$

finally the shape of the heat accumulator is cylindrical, its length equal to the diameter. The results are shown in *Table VI*. We see that the propulsion system with heat accumulator, for example for a large American passenger car, will become 275 kg too heavy if its range has to be 322 km. In the case of a lightweight car the difference is only 45 kg. For practically all other cars the specifications can be met with ease. Thus a commuter car, if it satisfies the prescribed weight, has almost twice the specified radius of action.

It is clear that, if we wish to continue in this direction, a great deal more development work will have to be done, though even in its present form the system shows great promise.

Besides the large radius of action and the complete absence of exhaust gases, the heat accumulator offers still more advantages. "Charging" can be relatively fast and can conveniently be done at night, taking advantage of cheaper tariffs for electric power at off-peak periods and evening out the load on the electricity grid. Furthermore, full engine power is available at all times, even if the heat accumulator is almost exhausted, because then the temperature of the heater head drops, so that the engine may be charged with a higher gas pressure. Last but not least the interior of the car can be warmed by heat taken either from the cooling water or from the heat accumulator. In the case of cars with electric propulsion this has always been a particularly difficult problem.

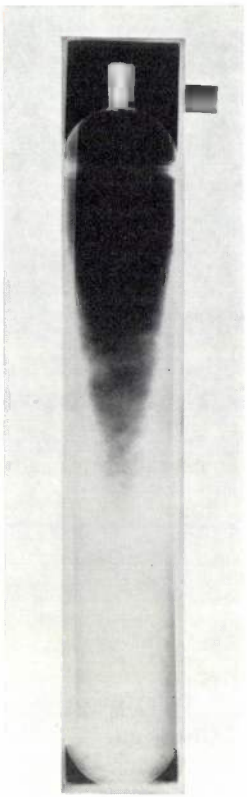


Fig. 21. A radiograph of one element of a heat accumulator, surrounded by a heat pipe. The salt (light region) is here in the solid state.

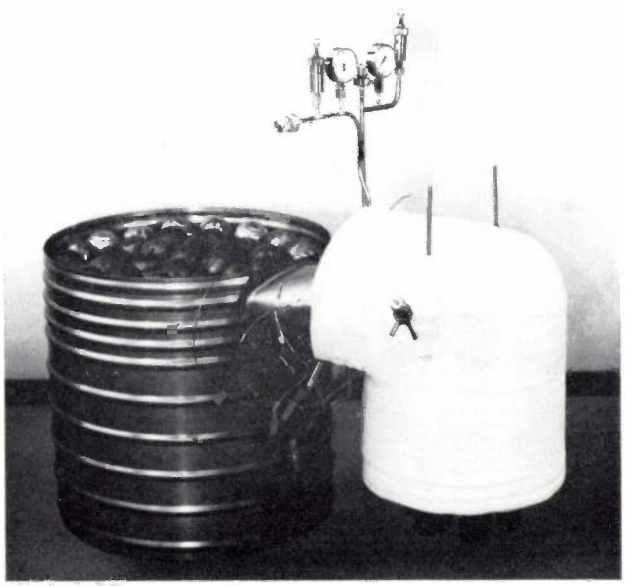


Fig. 22. Test set-up in accordance with fig. 20. On the left, with cover removed, the heat accumulator. The upper ends of the elements are just visible.

Table V. Some data concerning the principal types of cars at present in use [9].

		American family car	small European car (commuter car)	utility car	delivery van	city taxi	city bus
1. range of operation	km	322	161	80	97	241	193
2. maximum speed	km/h	161	129	105	90	124	88
3. acceleration to	km/h	97	97	48	64	64	48
in	s	15	30	10	20	15	15
4. maximum power output	kW	70	22	12	49	36	135
5. loaded weight	kg	1815	1135	770	3175	1815	13610
6. total weight assignable to new propulsion system							
a. conventional construction	kg	565	340	225	635	565	2270
b. lightweight construction	kg	795	475	320	910	795	3175
7. energy delivered	kWh	100	20	8	45	75	300

Table VI. Data calculated for the cars of Table V when equipped with a Stirling engine with LiF heat accumulator. Engine and accumulator are connected by a heat pipe. The accumulator is assumed to be cylindrical, with the height equal to the diameter.

		American family car	small European car	utility car	delivery van	city taxi	city bus
1. volume of heat accumulator tank	dm^3	385	77	30	174	289	1154
2. diameter of tank	cm	79	46	34	61	72	114
3. thickness of superinsulation material	cm	0.55	0.95	1.43	0.74	0.63	0.38
4. weight of engine + radiator	kg	216	82	49	162	124	379
5. weight of heat-pipe system	kg	32	12	7	24	19	57
6. weight of heat-accumulator material	kg	530	106	42	239	398	1590
7. weight of container + insulation	kg	62	21	12	37	52	130
8. total weight of propulsion system	kg	840	221	110	462	593	2156
9. weight assignable to propulsion system (Table V)							
a. conventional construction	kg	565	340	225	635	565	2270
b. lightweight construction	kg	795	475	320	910	795	3175
10. difference in weight (item 8 minus item 9)							
a. conventional construction	kg	+275	−119	−115	−173	+28	−114
b. lightweight construction	kg	+45	−254	−210	−448	−202	−1019
11. range of operation required	km	322	161	80	97	241	193
12. actual range of operation							
a. conventional construction	km	172	311	248	157	226	206
b. lightweight construction	km	298	480	387	252	350	308

Table VII. Data calculated for the cars of Table V when equipped with a Stirling engine heated by the combustion of hydrogen carried in a $LaNi_5$ hydrogen accumulator.

		American family car	small European car	utility car	delivery van	city taxi	city bus
1. volume of hydrogen accumulator tank	dm^3	132	26.5	10.5	60	99	396
2. diameter of tank	cm	44	25.5	19	33.5	40	63
3. length of tank	cm	88	51	38	67	80	126
4. weight of engine + radiator	kg	250	95	57	183	141	435
5. weight of $LaNi_5$	kg	288	98	39	219	366	1463
6. weight of tank	kg	26	5	2	12	20	78
7. weight of hydrogen	kg	7.9	1.6	0.6	3.6	5.9	23.7
8. total weight of propulsion system	kg	772	200	99	418	533	2000
9. weight assignable to propulsion system							
a. conventional construction	kg	565	340	225	635	565	2270
b. lightweight construction	kg	795	475	320	910	795	3175
10. difference in weight (item 8 minus item 9)							
a. conventional construction	kg	+207	−140	−126	−217	−32	−270
b. lightweight construction	kg	−23	−275	−221	−492	−262	−1165
11. range of operation required	km	322	161	80	97	241	193
12. actual range of operation							
a. conventional construction	km	194	376	322	187	261	225
b. lightweight construction	km	336	584	505	300	402	337

Hydrogen as fuel; the hydrogen accumulator

Though at first sight it seems strange to use hydrogen as a fuel for an engine, it is clear that at a given moment we may have to give thought to a synthetic fuel which, in production and use, does the least possible damage to the natural cycle on Earth [11]. Apart from thermal pollution, which is inherent in all thermal engines, hydrogen is a very "clean" fuel. On combustion there are no problems with carbon dioxide, carbon monoxide, and unburned hydrocarbons, and if the reaction of combustion is allowed to take place at not too high a temperature, the final product is just water. The use of hydrogen as a fuel holds no problems for the Stirling engine.

The applicability of hydrogen as a fuel depends mainly on the solution of two problems. In the first place a method must be found to produce hydrogen economically from *water*, and here the oxygen produced must be allowed to "escape" into the atmosphere so that subsequent combustion of the hydrogen restores the status quo. Naturally this is necessary only if hydrogen is used on a large scale as a fuel. Recently a very interesting chemical cycle has been described in which water is split into hydrogen and oxygen with a theoretical efficiency of 75%; the process requires the supply of heat at a temperature of at most 750 °C [12]. Once such a cycle is mastered, the heat liberated in nuclear reactions could be used directly to produce hydrogen on a mass scale. Still higher operating temperatures of nuclear reactors would, with suitable chemical cycles for those temperatures, yield still higher conversion efficiencies, and produce hydrogen still more economically.

The second problem that has to be solved is to find an easy way of storing large amounts of hydrogen, so that sufficient fuel can be carried in the vehicle. A recent discovery in our laboratory [13] may be a significant step in this direction. It has been found that some hexagonal intermetallic compounds of composition AB_5, in which A is a rare-earth metal and B is nickel or cobalt, readily absorb and desorb large amounts of hydrogen at a pressure of a few atmospheres. At 2.5 atm and room temperature the density of hydrogen in $LaNi_5$, for example, is almost twice as great as that of liquid hydrogen.

As in the case of the heat accumulator we have calculated for all six types of cars mentioned in Table V the dimensions of a tank filled with $LaNi_5$ capable of storing an amount of hydrogen at 20 °C sufficient for the previously mentioned specifications to be met. The calculation is based on the following data: 1) the amount of hydrogen that can be absorbed and desorbed is 180 cm^3(NTP)/g $LaNi_5$; 2) the plateau pressure is about 2.5 atm.; 3) the density of $LaNi_5$ is 8.5; 4) the heat of reaction is 0.059 kcal/g $LaNi_5$; 5) the net calorific value of hydrogen is 2570 kcal/m^3 (at NTP).

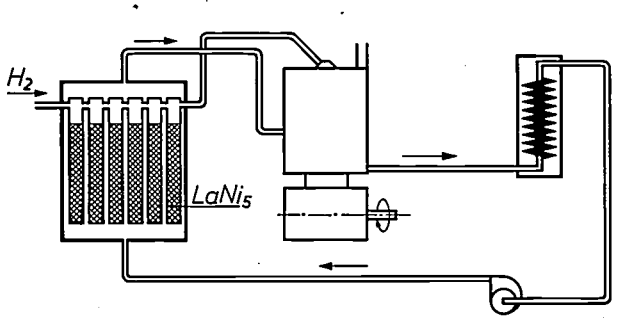


Fig. 23. Schematic representation of a Stirling engine (centre) with $LaNi_5$ hydrogen accumulator (left). Because heat is liberated during the charging of the accumulator and heat must be supplied during discharge, the accumulator is incorporated in the cooling-water system of the engine. The water is forced by a pump (bottom right) through the accumulator, the engine, and the radiator (top right).

When the tank is filled with hydrogen a great deal of heat is liberated (about 13% of the calorific value). We have assumed that the tank is built into the cooling-water system, as shown in *fig. 23*. When the engine is running, hydrogen is continuously withdrawn from the tank, which will thereby tend to cool. This heat is made good by the cooling water of the engine after it has passed through the radiator. This causes the cooling water to cool down further (about 10% of the amount of heat which has to be withdrawn from the cooling system of the engine can be used to prevent cooling of the $LaNi_5$). For the calculation of the $LaNi_5$ tank we have assumed furthermore that the volume of the tank is 2.3 times greater than the volume of the $LaNi_5$ because a) the $LaNi_5$ is present in powder form, b) the $LaNi_5$ powder expands during the uptake of hydrogen, and c) about 15% of the volume is required for the heat exchanger in contact with the cooling water. We have also assumed the tank to have a cylindrical shape with a length of twice the diameter. The efficiency of the Stirling engine is taken as 38%. With these data and assumptions the values shown in *Table VII* have been obtained.

We see from this table that the propulsion system with hydrogen accumulator, just like that with the heat accumulator, is too heavy for the large American motor-car by 207 kg if a radius of action of 322 km is specified. But it can easily satisfy the requirements of a lightweight car (—23 kg). For all other motor vehicles the specifications can be met with an ample margin. Thus a small European car with an all-up weight of 1135 kg (Table VII) has a range of 376 km, while the specifications call for a range of 161 km.

Although the above considerations on the application of heat accumulators and hydrogen accumulators in various types of motor-cars are purely theoretical, they do indicate the wide applicability of the Stirling engine.

Conclusions

There are signs which indicate that the biosphere is undergoing changes in the negative sense because of the activities of man. The interference with nature on the part of man in the course of his struggle for survival and the pursuit of pleasure has vast consequences both for himself and for other life on Earth. Above all the exponential increase in the consumption of raw materials and in the production of rubbish — which pollutes the Earth — gives rise to great concern. At present it is air and water pollution that give cause for particular anxiety. The air is being polluted not only by industry but also by the exhaust gases of motor-car engines. In the United States the contribution of vehicle engines to air pollution is more than half of the total, but locally it is much greater because of the heavy concentration of cars in and around cities (about 80% for Los Angeles). In addition the diesel engine in buses and lorries contributes to other sources of annoyance: noise, smell and soot.

It is against the background of these facts that a description has been given of the Stirling engine, which, from the technical point of view, could well replace conventional vehicle engines. With its many special features, the Stirling engine could make a great contribution to environmental hygiene, not just today and tomorrow, while fossil fuels are still in use, but also the day after tomorrow, when these are no longer available.

However, it must be admitted that the really *large-scale* introduction of the Stirling engine for vehicular propulsion will be extremely difficult for economic reasons. To a question put to me a few years ago in the United States after a lecture on the Stirling engine, namely "What is wrong with the engine?", the answer had to be "The existence of other engines". Indeed it is difficult to compete against engines whose price is little more than that of so many pounds of steel, cast iron, or aluminium. That is why, up to a few years ago, only those applications were envisaged in which, purely on economic grounds, the introduction of the Stirling engine would have a fair chance of success. These are the areas where the conventional internal-combustion engine cannot be used or where the abatement of the nuisance caused by these engines is expensive.

We may fairly ask whether the changed conditions regarding pollution will give the Stirling engine, or some other alternative to the petrol engine, a place in the near future as a prime mover for vehicles. Much will depend on regulations and other circumstances. Let us consider once more the United States, because that is where plans and regulations are furthest developed.

The first question is whether the legislator will maintain the strict specifications of the emission of exhaust gases proposed for the year 1980.

The second question is whether the present-day petrol engine can be improved so far as to meet *all* specifications (including NO_x) without giving rise to big price increases. Here of course it is not just a question of the purchase price but also the more expensive maintenance, stricter government surveillance, greater fuel consumption — particularly if lower compression ratios have to be used — and the higher price of petrols with a high octane value obtained by means other than lead dopes.

In the third place the new regulation which specifies the maximum amount of exhaust gases, irrespective of the size of the car; will certainly lead to the introduction of smaller, lighter cars of lower engine power. The question is how far one must and can go in this respect.

The fourth question is how far the real cost price of the Stirling engine can be lowered by the measures mentioned in this article, so that this engine can, within the framework of the regulations concerning exhaust-gas emission, compete on level terms with all the other alternative methods of propulsion.

The final question is what further measures the governments will take to combat environmental pollution: the introduction of a new type of engine — even the engine with the best qualifications — is in fact feasible only on the basis of general government policy [14].

Irrespective of whether an improved petrol engine or a new prime mover is going to propel the private vehicle of the future, it is obvious that limitation of noxious substances in exhaust gases is going to cost a great deal of money. In this connection there is a growing opinion in America that continual and frequent government supervision of cars with "cleaned-up" internal-combustion engines will turn out to be impracticable [15]. That is why thoughts are steadily turning to power sources which are "clean" by nature: prevention is better than cure.

---

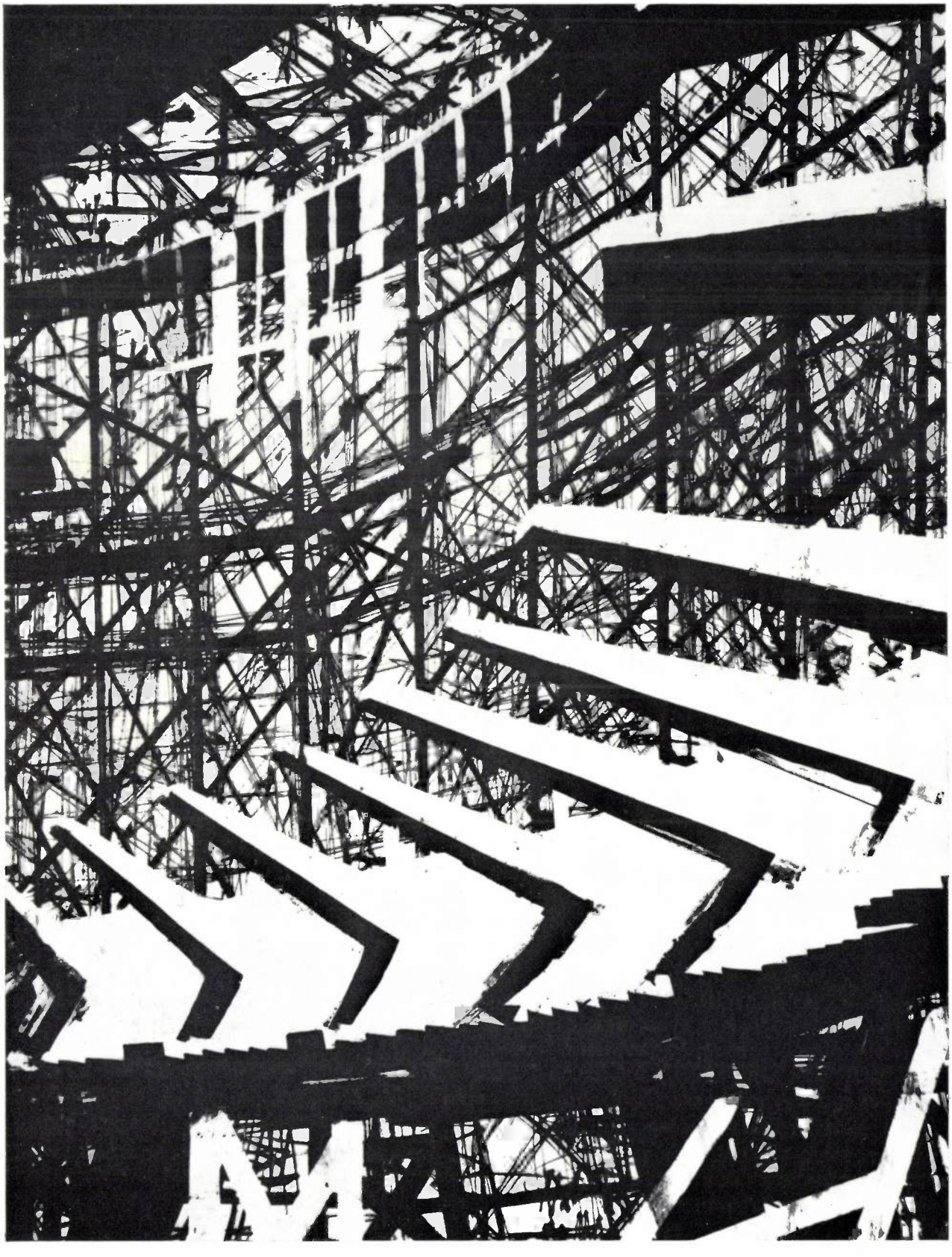
[11] L. Green, Jr., Energy needs versus environmental pollution: a reconciliation?, Science **156**, 1448-1450, 1967.
J. McHale, World energy resources in the future, Futures **1**, 4-13, 1968.
P. E. Glaser, Solar energy — an alternative source for power generation, Futures **1**, 304-313, 1969.

[12] G. de Beni and C. Marchetti, Hydrogen, key to the energy market, Euro-Spectra **9**, 46-50, 1970 (No. 2).

[13] J. H. N. van Vucht, F. A. Kuijpers and H. C. A. M. Bruning, Reversible room-temperature absorption of large quantities of hydrogen by intermetallic compounds, Philips Res. Repts. **25**, 133-140, 1970 (No. 2).

[14] D. D. Kummerfeld and G. Wilcox, Federal policy on auto air pollution control, research report of Center for Political Research, Washington D.C., April 1970.

[15] In the United States attempts are being made to encourage the design of clean engines: in order to bring mass production into the realm of practical politics, the "winner" of the Federal Clean Car Incentive Program [5] can, according to a certain time-table, replace all government vehicles (several hundred thousand) at a reasonable return (at most, twice the normal price) by cars equipped with the new power unit.



The construction of a very unusual building seen through the eyes of an artist. The following pages give an account of the birth of the idea behind it and of the equally unusual exhibition it houses.

The EVOLUON

A permanent Philips exhibition

J. F. Schouten

In 1961 Ir. F. J. Philips, thinking ahead to the forthcoming 75th anniversary of Philips in 1966, put the following question:

"*Is there any sense in continuing to take part in world exhibitions, as we did in Brussels and as we might do, for example, in New York and Montreal? We do this because other firms do it. The costs, however, are very high and the effects, though spectacular, are shortlived. Suppose that we stopped taking part, saved up the millions that we would have spent on these exhibitions, and then used that money, perhaps with a supplementary investment, to set up a permanent exhibition about our own company?*"

Not all of Ir. Philips's colleagues on the Board of Management were as enthusiastic about this idea as he was. The investments required would have to be made at the expense of industrial projects, and it was difficult to assess the indirect profit from such an undertaking.

Despite this the plan was finally accepted. Ir. L. C. Kalff, former arts director of Philips, was commissioned to design the building in collaboration with De Bever, a firm of architects. The services of Mr. James Gardner RDI were obtained for the design of the exhibition, and Mr. Jacobus Kleiboer was appointed adviser.

The search for a "distinctive form for a distinguished content" led to the concept of an enclosed shallow bowl (like a "flying saucer"), supported by twelve V-shaped pillars to give an illusion of weightlessness. Within the dome thus formed there were to be three galleries, the diameter increasing towards the top. Two platforms projected inwards from the third and uppermost gallery. Galleries 3 and 2 each comprised about a third of the total available floor area for the exhibition, gallery 1 a sixth, and the two platforms the remaining sixth.

We shall not deal here with the particularly interesting way of building the dome with prefabricated concrete elements — 288 flat sheets form the lower part and 822 hexagonal prisms the upper part. The building was started in September 1964.

Prof. Dr. J. F. Schouten is a Scientific Adviser with Philips Research Laboratories and a Professor Extraordinary at the Technical University of Eindhoven. Prof. Schouten is also Director of the Institute for Perception Research (IPO), Eindhoven, a member of the Advisory Council of the Evoluon and a member of the Board of Consultants of the Netherlands Young Scientists Association.

The birth of the central theme

From 1961 onwards all kinds of ideas were put forward about what Philips should exhibit in this building and about the way in which the exhibits should be linked together in a framework of topics and themes.

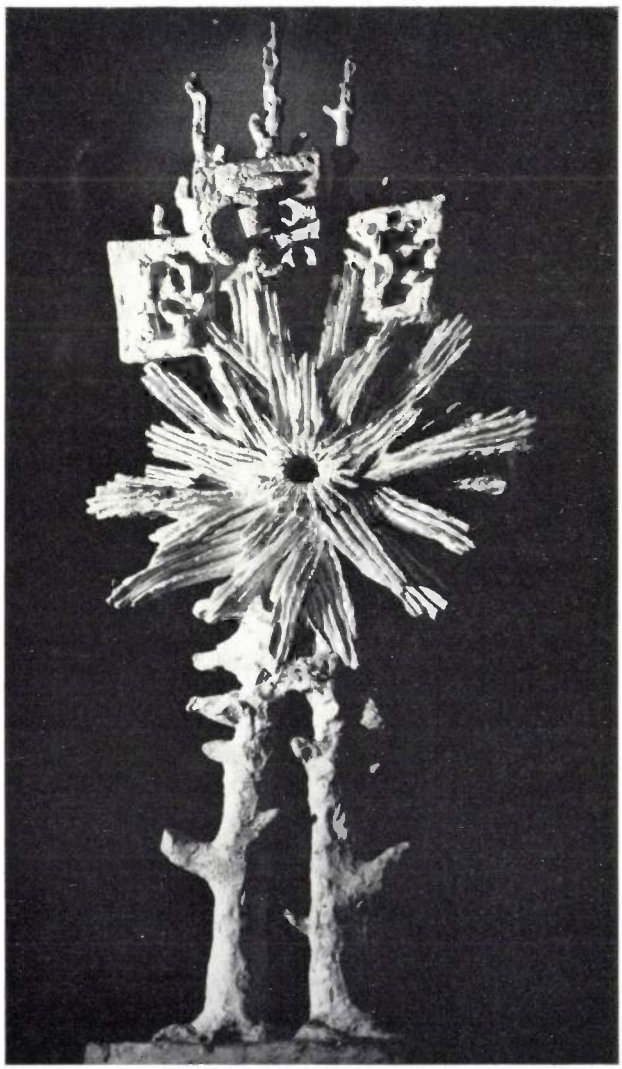
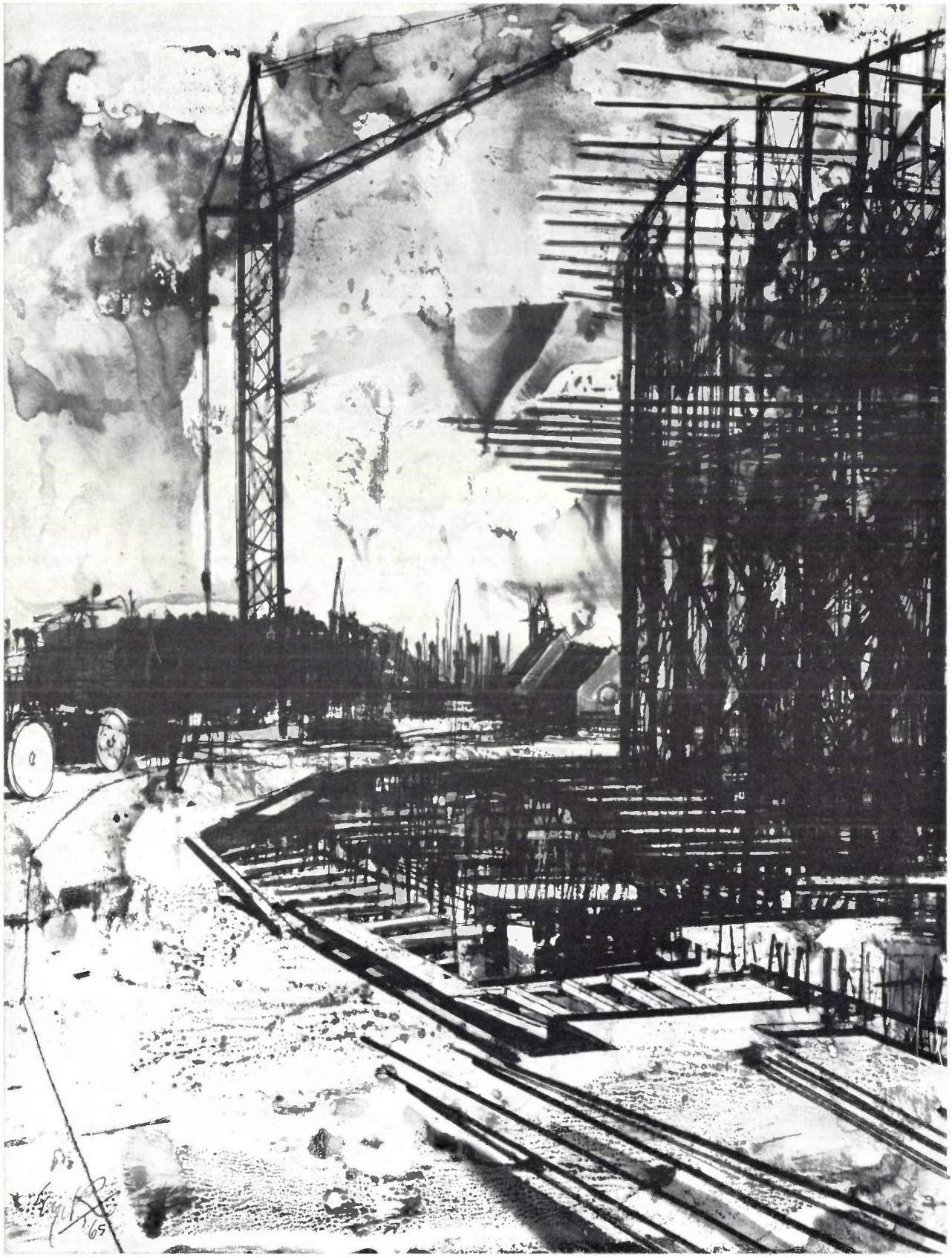
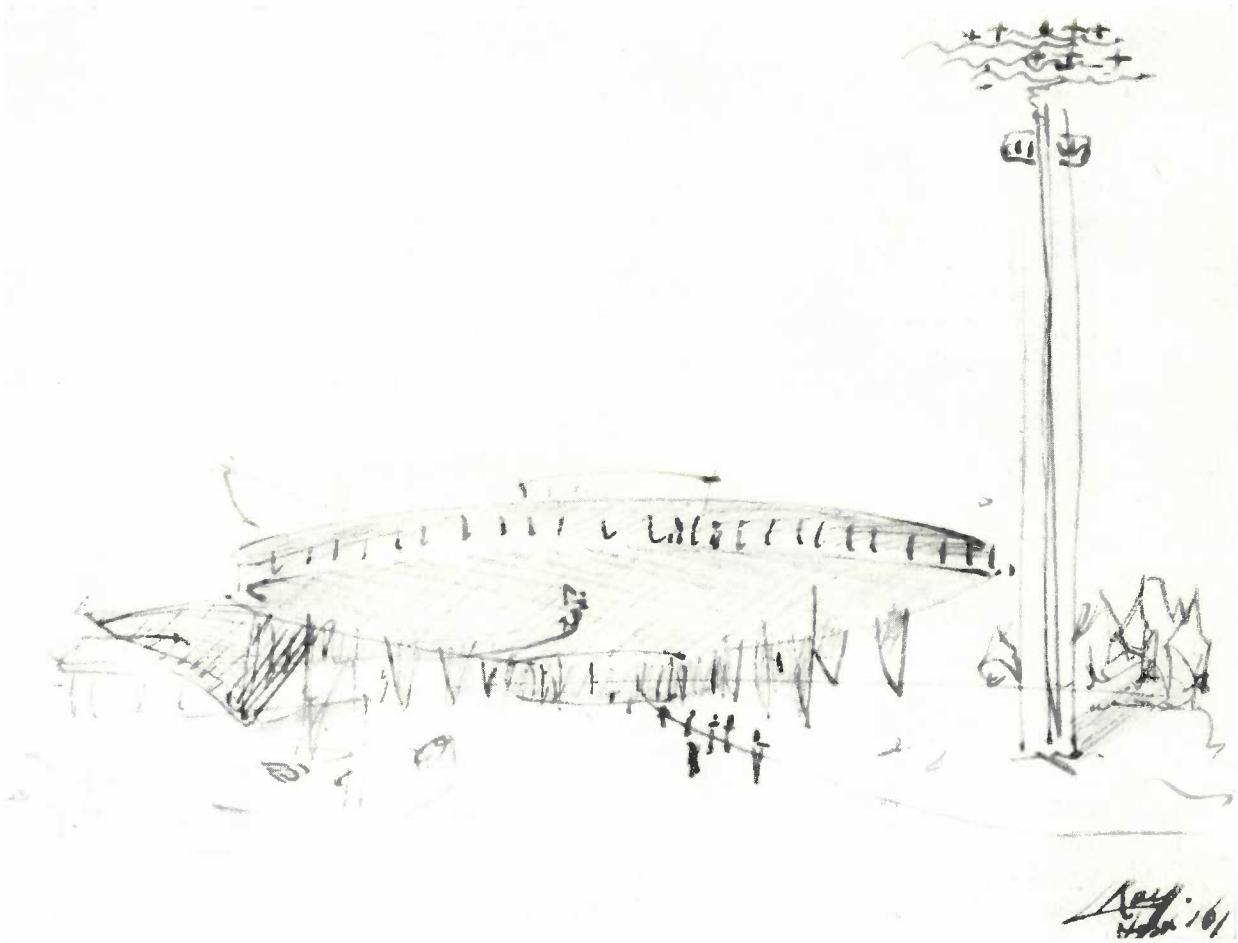


Photo Lucien Herné, Paris

Bronze by Ossip Zadkine "Eclatement de l'atome", which was presented to Ir. L. G. Kalff, the former arts director of Philips, by the society "Recherches et formes de demain", as a prize for the whole of his completed work, and is now on display at the Evoluon. Various other works of art inspired by natural events are also to be found in the building.



One of the impressions drawn by Charles Eyck of Maastricht during the building of the Evoluon; there is another on p. 186.



Above: the first sketch of the design by Ir. Kalff. On the following pages there are some photographs of the construction of the building.

As early as 1961 Mr. J. Kleiboer came up with the name EVOLUON or EVOLEION. It was a name, he said, that symbolized growth in time and growth in complexity, and would characterize an enterprise like Philips very well. Moreover, speaking from his experience as an exhibition designer, he thought it would be an enormous advantage to have a new name instead of one of the hackneyed variants on "expo" and "rama". At the time his proposal was not taken up because a term borrowed from biology did not appeal particularly to those who had a technological exhibition in mind. By the beginning of 1965 it was clear that this name would in fact be a particularly appropriate one for an exhibition whose dominant theme was to be industrial evolution.

This concept of industrial evolution was developed in a memorandum, published as a booklet in July 1965. The booklet served as a working document to indicate the general lines on which the exhibition was to be designed, and a translation of the key section of the booklet is given below.

WHAT THE EXHIBITION IS NOT TO BE

Not a showroom for the products of the Philips Product Divisions. This would attract only transient interest. Moreover, many Product Divisions already have better showrooms than they could hope to get from their share of the total floor area of about 5000 m².

Not a kind of Science Museum. Here again, the interest would be too limited. Moreover, it would not give a truly representative picture of the industrial group that is presenting the exhibition.

No surfeit of technology, or of mechanization and automation. This could make the visitor feel humbled and insignificant when he left the exhibition. It would quicken his fear of losing his job to a robot, the more so in a period of economic recession. Moreover, we consider that such an overemphasis would give an untrue and unbalanced picture of the essential significance of technology to man.

Finally, the thinking of the exhibition must be based not on ourselves — Philips — but on the visitor. Obviously, the exhibits relate to Philips, but the visitor's attention can only be caught and held by appealing to his interest and natural curiosity.

THE EVOLUTION IDEA

Man has always striven to acquire mastery over matter and to form stable societies with his fellows.

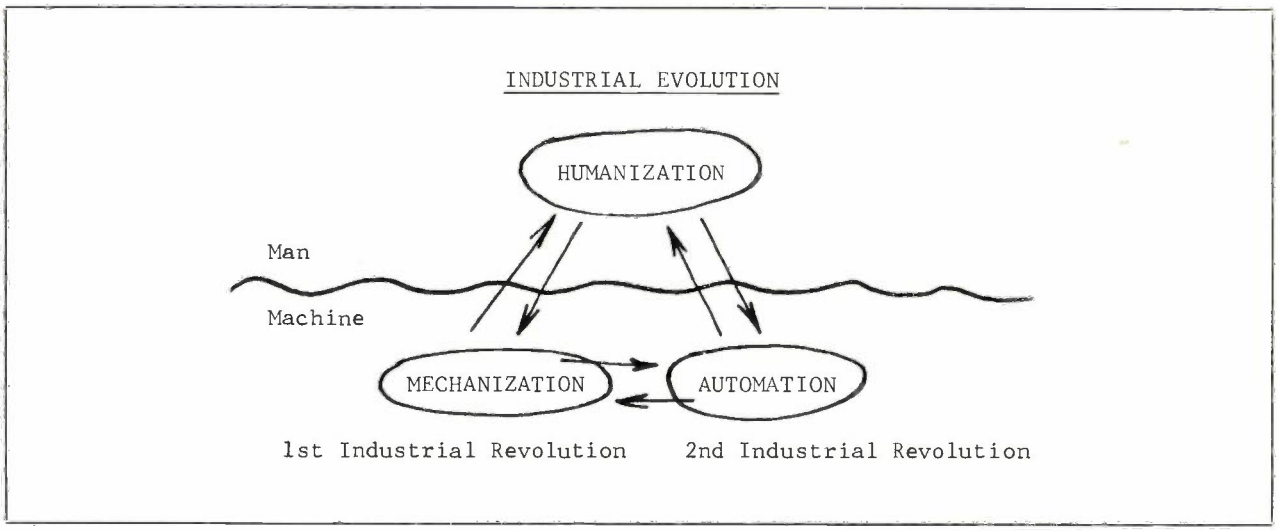
By the end of the 19th century, in a society already unmistakably industrial, many people had to work as *producers* in conditions that we should consider appalling. The hours were long, the work hard, and the conditions unhealthy. But their share as *consumers* (of food and clothing, housing, medical care, schooling, leisure, etc.) seems pitiably small to us today.

The key to the enormous improvement that has taken place in Western society during this century, both in working conditions and the goods available to the average consumer, is to be found in the enormous progress made in the industrial process. More rational production methods brought about by *mechanization* and *automation* have vastly increased output per worker, while at the same time making it economically possible to bring about the much desired *humanization* of working conditions (shorter working hours, better working conditions, social facilities, etc.). As for the consumer, what used to be the luxury for the few has become the everyday fare for the entire population.

Mechanization is sometimes referred to as the first industrial revolution, and automation as the second. However, these two trends, plus the humanization of working conditions — all developing in equilibrium — ought really to be regarded as an *industrial evolution.*

If we define industrial evolution in this way we are not indulging in an easy optimism, for too hasty an introduction of mechanization and automation by the employers, or too eager demands from the employees for better conditions, could upset the balance and bring about a degeneration from evolution to revolution.

We might also speak of the *evolution of man through technology evolved in a controlled way by man himself.*


INDUSTRIAL EVOLUTION
HUMANIZATION
Man
Machine
MECHANIZATION
AUTOMATION
1st Industrial Revolution
2nd Industrial Revolution

THE EVOLUON

In this evolutionary industrial development Philips clearly play an extremely important and fascinating role. It is the role of a growing enterprise which has made available to the whole world a vast array of important products and systems that contribute to the consumer needs of man and society; which is organizationally and financially sounder and more forward-looking than many state systems; which is a welcome partner, supplier and customer of other business enterprises; and which offers employment and the opportunity for personal development to some two hundred and fifty thousand people all over the world.

This is essentially a dynamic role. All that we have to offer is today's solution of yesterday's problem, and we have the keen determination and unbounded confidence that we shall solve today's problems by tomorrow.

This leads us to see the exhibition not as a technological exhibition but as an *industrial* one. This in its turn means that we should show that all that has been done in the last 75 years in organization, methods, products and the distribution of those products has been achieved by a *community of people*, who have time and again been faced with almost insoluble problems, who have struggled to overcome those problems and who have achieved no small success.

It should moreover be made clear that an industrial firm is a community which makes its products to sell them, and which must see every guilder it spends come back to the till.

An exhibition is of course a display of material objects, adorned with texts and art work. Everything turns, however, on the ideas which are communicated to the visitor through this display.

The main idea is this: that technology should be the servant of man. In view of his needs as a consumer, man cannot do without technology. He must therefore organize technology so effectively that he is able to meet his consumer needs in the fullest possible way by matching human labour to man's capacity and potential.

This exhibition will be opened on the occasion of a 75th anniversary, and is therefore bound to call up a perspective of those 75 years. This perspective, this dynamism, this continuous struggle and process of creation in the midst of a struggling society which is itself still in the making, should form, quite apart from the anniversary, the main and permanent idea behind the EVOLUON.

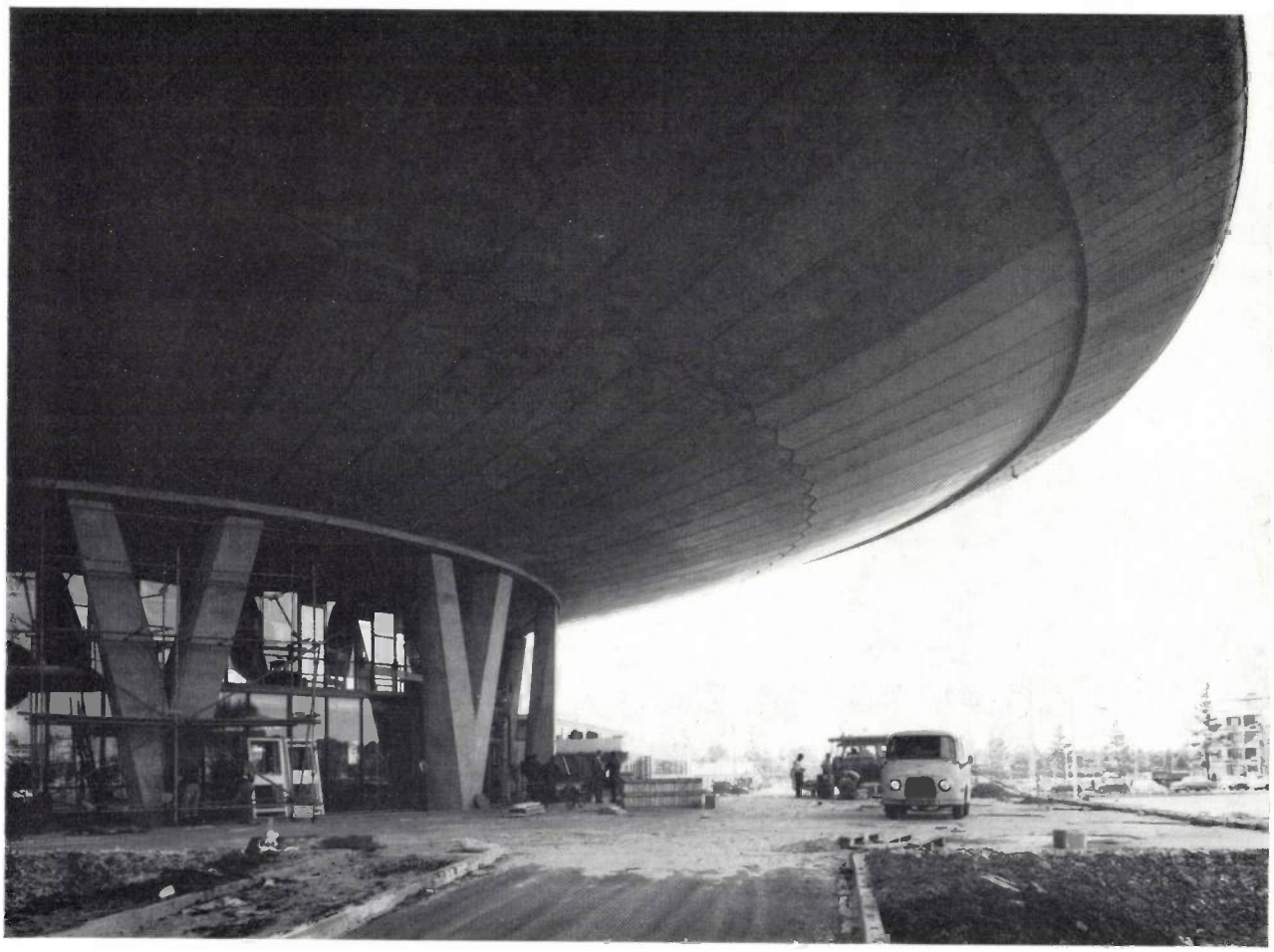
We shall then be presenting not a static display of what now exists, as in a zoo, but the evolutionary dynamism of the creative process and of mankind's struggle to overcome its problems.

Most industrial exhibitions lack this evolutionary theme. It will therefore give our exhibition a style and distinction of its own.

Another tremendous advantage of placing our activities in the context of an evolving society is that it will enable people to "understand" Philips. Every visitor, whether employee, housewife, child, scholar, layman, industrialist or statesman, will recognize some part of the display that bears on the life he or she knows.

This recognition should leave a more lasting impression, with the result that the visitors will continue to think about and discuss the problems (television, education, upbringing, recreation, industry, society, evolution, revolution, etc.), even after they have left the exhibition.

The EVOLUON, then, whose name embodies its theme of evolution, is a human story. A story of people banding together to create technology and distribute its benefits to those who have need of them.



The choice of the main and subsidiary themes

After these rather general digressions, the question was now to arrive at significant main themes on the basis of the principles outlined. The presentation of these themes would have to be adapted to the space available for them, and it would also be necessary to take into account the visitor's progress through the building. The idea was that he would first go by lift to the large platform, and then descend via galleries 3, 2 and 1 to the ground floor. Our original choice was the following:

On gallery 3: the impact of technology on society.
On the large platform: research in the natural sciences.
On the small platform: research in the life sciences.
On gallery 2: technology.
On gallery 1: Philips as an industrial enterprise.

If the social significance of technology is to be displayed, this main theme should be divided into subsidiary themes relating to social aspects. These were:

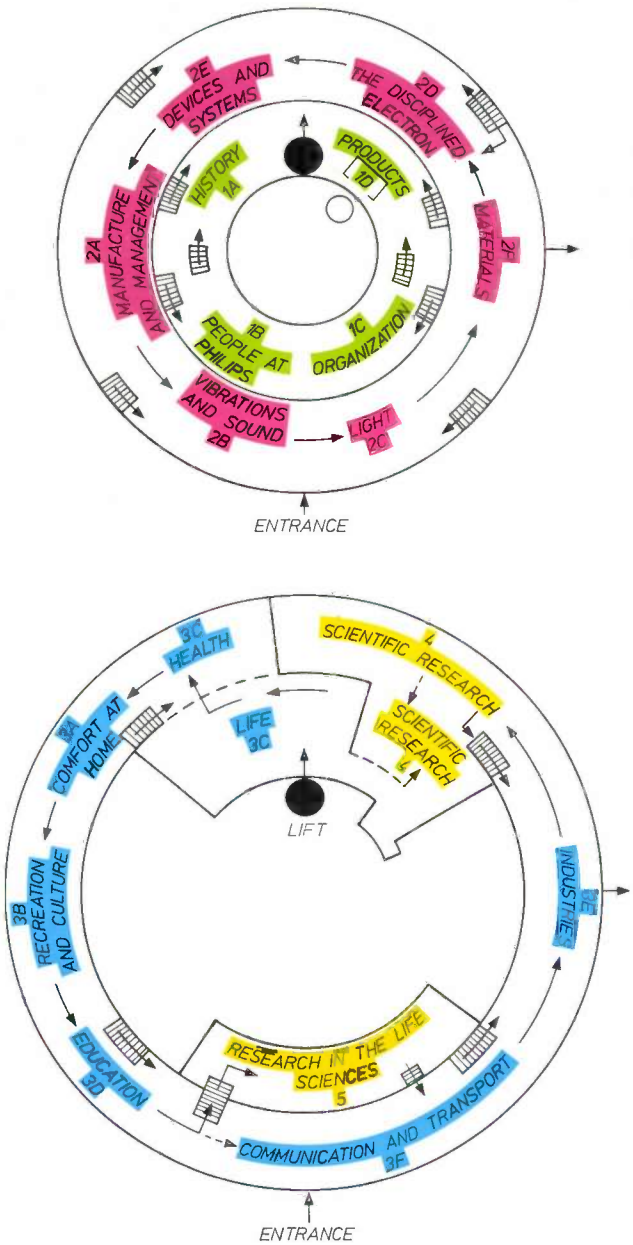
1. Life and health
2. Comfort in the home
3. Recreation
4. Education
5. Communication
6. Technology for other industrial firms.

In broad lines this scheme was adhered to in the realization of the exhibition, but in one point of detail a change was made at once. Mr. Gardner wanted the visitor, as soon as he left the lift at the large platform, to be confronted with a spectacular representation of the origin of the Earth and of life on Earth. This meant that research in the natural sciences could not take up the whole of the platform, and it was decided to split this subject up, situating some of the display on the right side of the platform and the rest on the adjoining area of gallery 3 linking up with it.

The line of thought behind the scheme outlined above was the following. On gallery 3 the visitor will be able to find an answer to the question, "What use is technology to me?" and thereupon start taking a closer interest in technology itself. This he will find on gallery 2, divided into six subsidiary themes:

1. Sound
2. Light
3. Matter, the key to all production
4. The disciplined electron
5. Equipment and systems
6. Manufacture and production control.

After completing gallery 2 the visitor has covered five-sixths of the exhibition but has as yet scarcely encountered the name Philips. This confrontation takes place on gallery 1, which is devoted entirely to Philips. Here the visitor sees the history of the company, the story of people at Philips, the organization of Philips as a multinational Group and a representative display of the products made by the fourteen Product Divisions.



One of the pages from the "work book", with the planned division of space; the present division differs somewhat from this.

Photo Jan Bijvank, Eindhoven

The interior, seen from gallery 3. The three galleries can clearly be distinguished. One of the features on the upper gallery is a 1/4 scale model of the ELDO telemetric aerial for tracking satellites. The molecular model hanging from the ceiling is a representation of polypropylene. The photograph was taken while a light-diffusion grid was being hoisted into place.

Photo Gunnar Wahlén, Stockholm

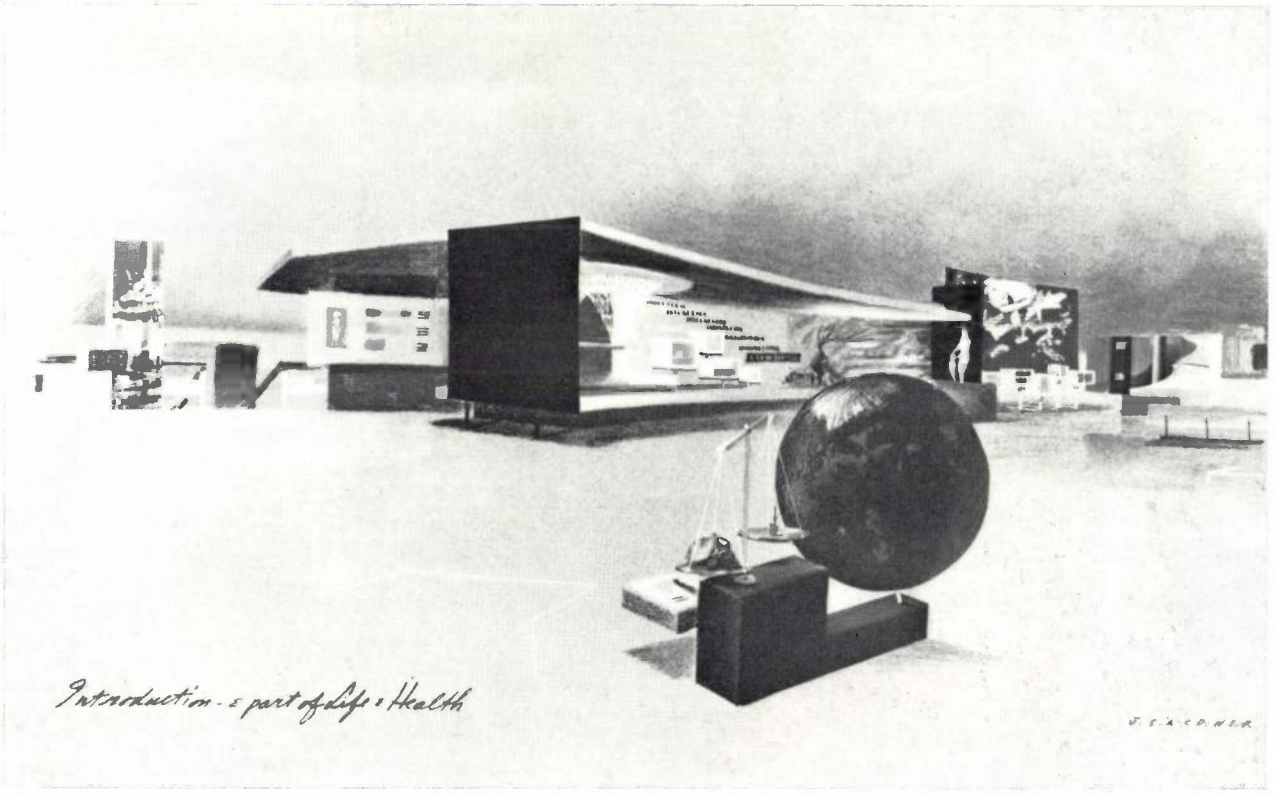
Realization

To carry out the ideas outlined above eighteen working groups were formed — one for each of the eighteen subsidiary themes — with a total of about 120 members. These working groups were to take as their starting point the premise that the value of an exhibition to the public need *not* depend solely on the intrinsic technological or scientific contents. It was necessary to think up a "story" which the industrial designer could get across to the public. Moreover it was impressed upon them that a rigid sequence of ideas or rigorous consistency soon becomes boring. We called this requirement the "alternation of aspects". The visitor was to be alternately confronted with logic and surprise, matter and man, seriousness and humour, "don't touch!" and do-it-yourself, the topical and the historical, today and the future, and so on. It was expressly stipulated that completeness in the treatment of a theme was never a desirable aim.

By far the most difficult task was to give expression to the human aspect in what was obviously a material exhibition. To this end eighty "pioneers" in the fields of the subsidiary themes were depicted. This was to serve several purposes. Not only would it bring out the necessary human aspect, but it would bring home to the visitor what science, technology and society owes to the past, and remind him that we stand on the shoulders of our forebears. Moreover it would be a fitting tribute to those pioneers.

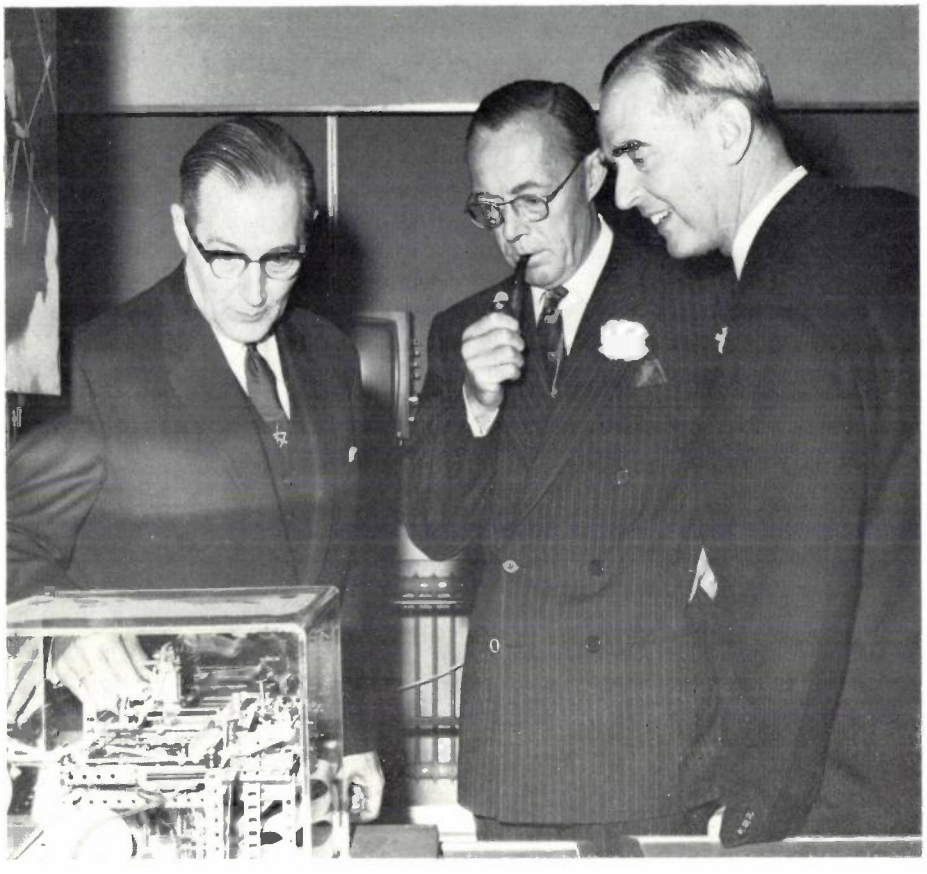
The composition of the working groups was interesting, because there was hardly any subsidiary theme to which widely different departments of the Philips organization were not able to make some contribution. A number of working groups had already been active for some time, but it was not until the official inauguration of the working groups in June 1965 that all the eighteen were complete and a full start could be made. It was a gigantic enterprise to bring into being, within fifteen months from the first confused discussions, a reasonably consolidated exhibition ready for opening on 24th September 1966.

Meanwhile all kinds of subsidiary projects took shape. Mr. Kleiboer, for example, kept insisting on the significance of the Evoluon for youth. This led to the installation of the "Prelude to Technology" display in the basement of the annexe building, originally intended for service use. The basement also provided accommodation for the small auditorium. The original plans for a large auditorium, to be built as a separate building beside the Evoluon, could not be carried out because of lack of time and money.



Sketch by James Gardner for one of the exhibits.

← The interior seen from gallery 3, in the opposite direction from that in the previous photograph. On gallery 1: the exhibit "People at Philips".



H.R.H. Prince Bernhard and Mr. Philips looking at an exhibit after the opening of the Evoluon on 24th September 1966.

"Prelude to Technology"

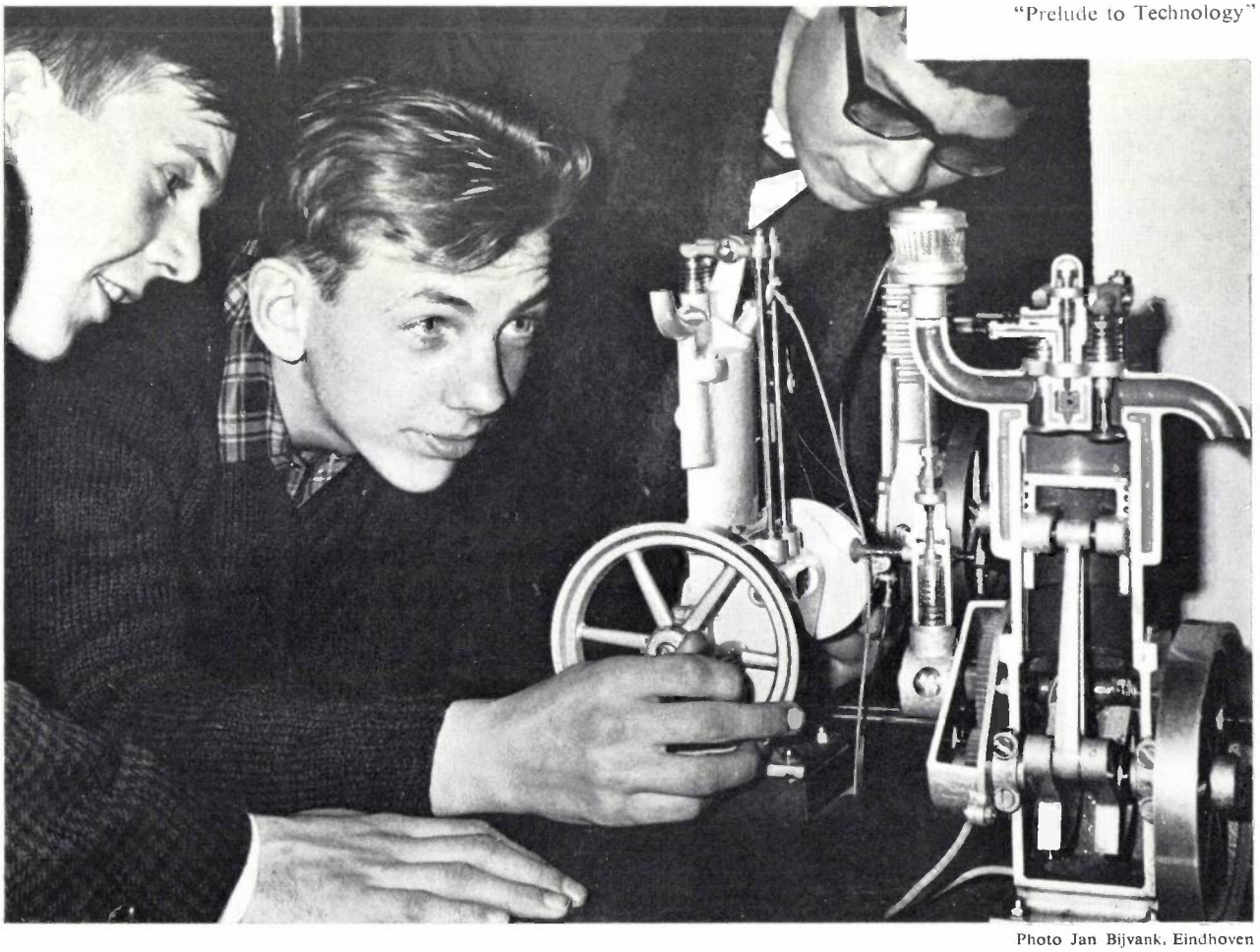


Photo Jan Bijvank, Eindhoven

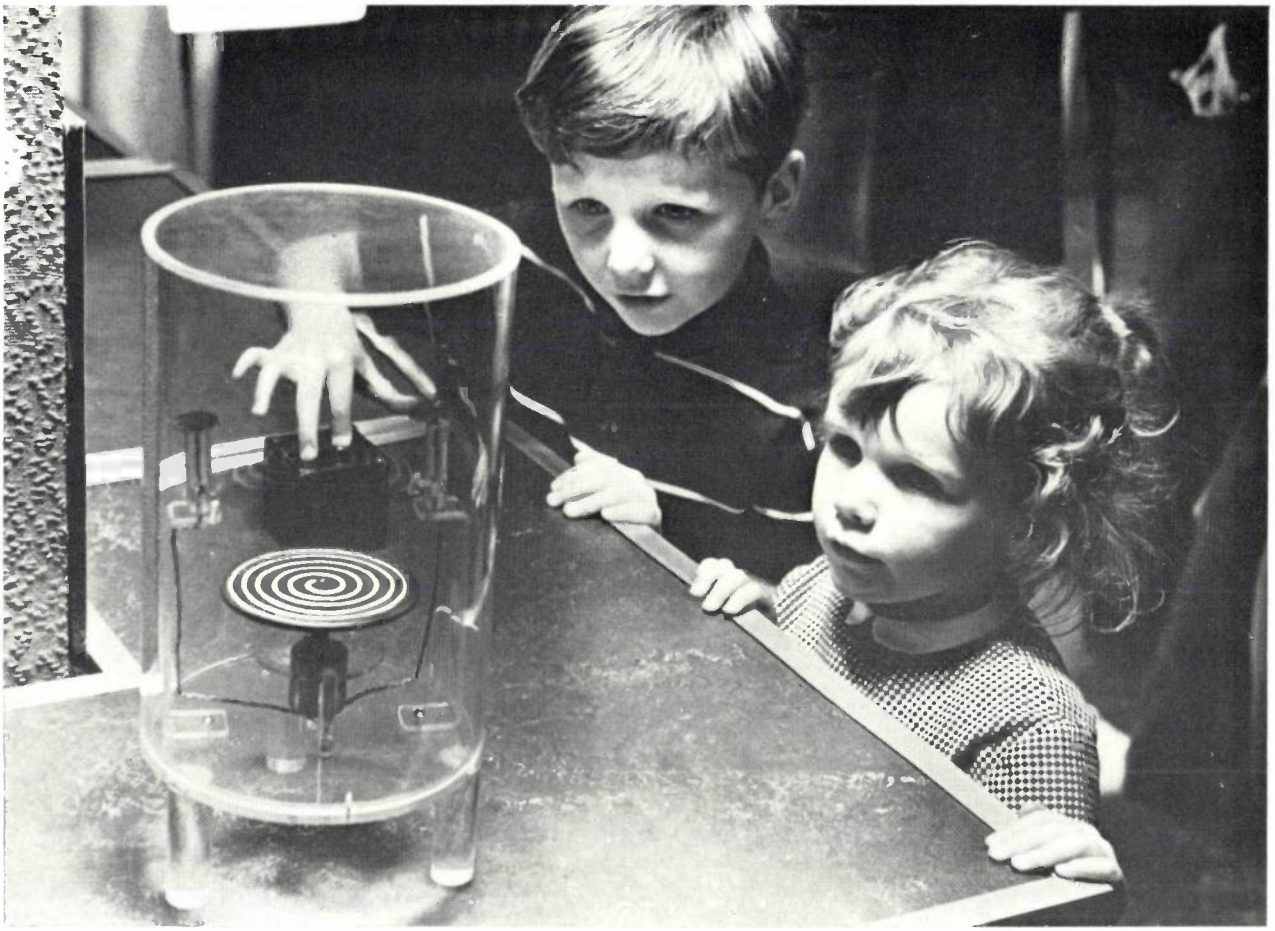


Photo Jan Bijvank, Eindhoven

The inauguration and the experience of the first years

The Evoluon was officially opened on Saturday 24th September 1966 by His Royal Highness Prince Bernhard. At that time 85% of the planned exhibition had been completed. It must be admitted that on the eve of its opening the Evoluon was decidedly unpopular with many people in Philips, except among the few hundred people who had contributed and who had their imaginations stirred by this adventure. This negative attitude changed abruptly during the opening weekend, when the visitors had the surprising experience of not finding themselves in the super Philips showroom they had expected. It was of course our intention that the Evoluon should contribute to the prestige of Philips. Our belief that this prestige would best be achieved by an indirect approach was confirmed beyond all expectations.

This welcome success had two main origins. First, the Philips staff engaged in this project possessed so little exhibition experience that they were able to give full rein to their imagination. Secondly, and perhaps more important, Mr. Gardner and Mr. Kleiboer and the other expert advisers were only too ready to apply their enthusiasm to these unconventional ideas.

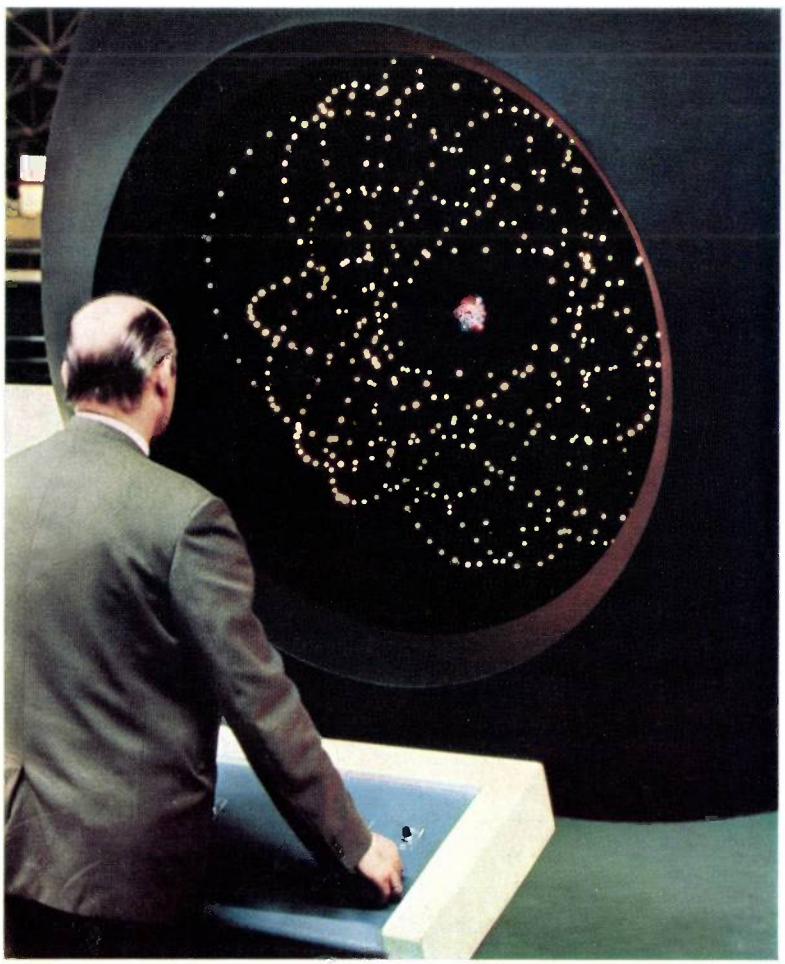
Although the number of visitors is not a direct measure of success, it was nevertheless extremely gratifying that the most optimistic estimate — 300 000 visitors a year — was well surpassed. The annual number of visitors steadily rose from about 440 000 to nearly half a million, and not long ago the two-millionth visitor was welcomed.

Something that far exceeded our wildest dream was the way in which the Evoluon has been found to appeal to the young. That well-known phrase "young people from 8 to 80" often came up in our discussions. Indeed, it must conceal a deeper truth: something with a direct appeal to the young may also appeal to their elders through its rejuvenating qualities.

Greatly aided through the personality of Ir. G. Ahsmann [*], the Evoluon's appeal to the young also made it into a natural springboard for a wide variety of new youth activities, such as study weekends for secondary-school children, for the Young Scientists Association (both nationally and internationally [**]) and a young people's laboratory.

[*] Ir. G. Ahsmann was the chief associate of Prof. Schouten in 1965 and 1966 and is now head of the scientific department of the Evoluon (*Ed.*).

[**] The International Coordinating Committee for the Presentation of Science and Development of Out-of-school Activities (ICC), Place St. Lazare 2, 10130, Brussels.



Top: Faraday in his laboratory.

Left: Model of the atom. The nucleus and the electron orbits for the first eleven elements of the periodic system can be made visible in the dark interior of the sphere.

Upper right: One of the exhibits on gallery 1: The organization of Philips.

Lower right: The colours, characterized by the three-dimensional colour space. The rectangular coordinates are chosen in accordance with Hering's complementary colours.

Far right: The exhibits differ not only in subject but also in the way they are presented. From top to bottom:

The living cell

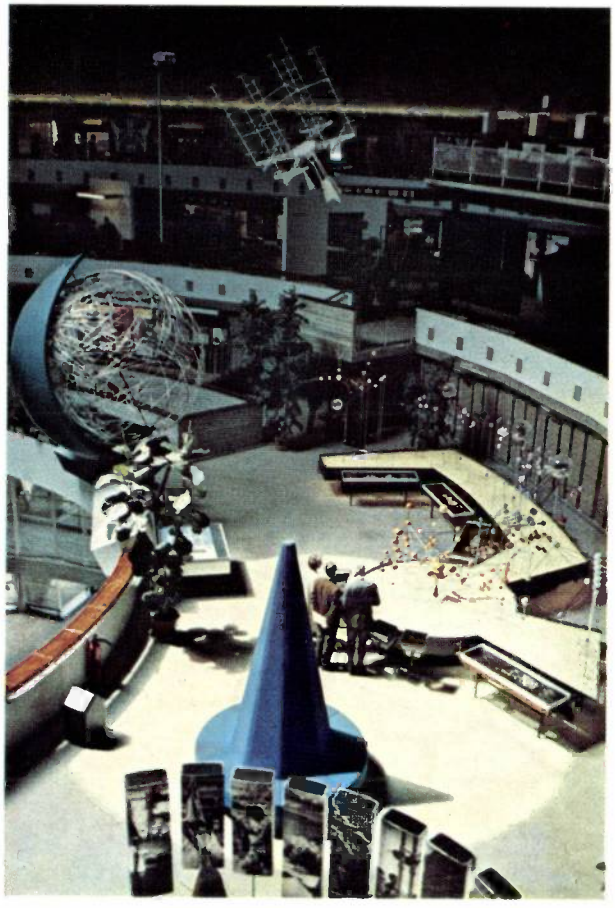
Perception, inside and outside the exhibit

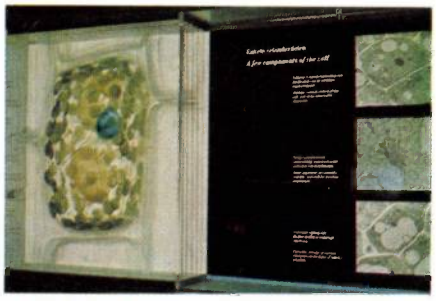
Sound spectroscope

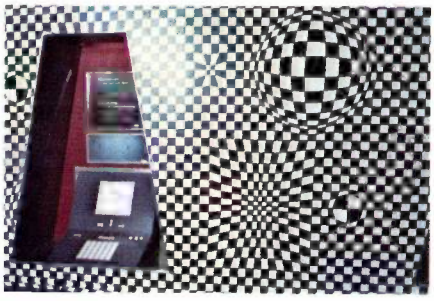
Interrupted gearwheels

Working model of a steam-engine, made in 1864 by E. H. Stuyver, an Amsterdam coppersmith

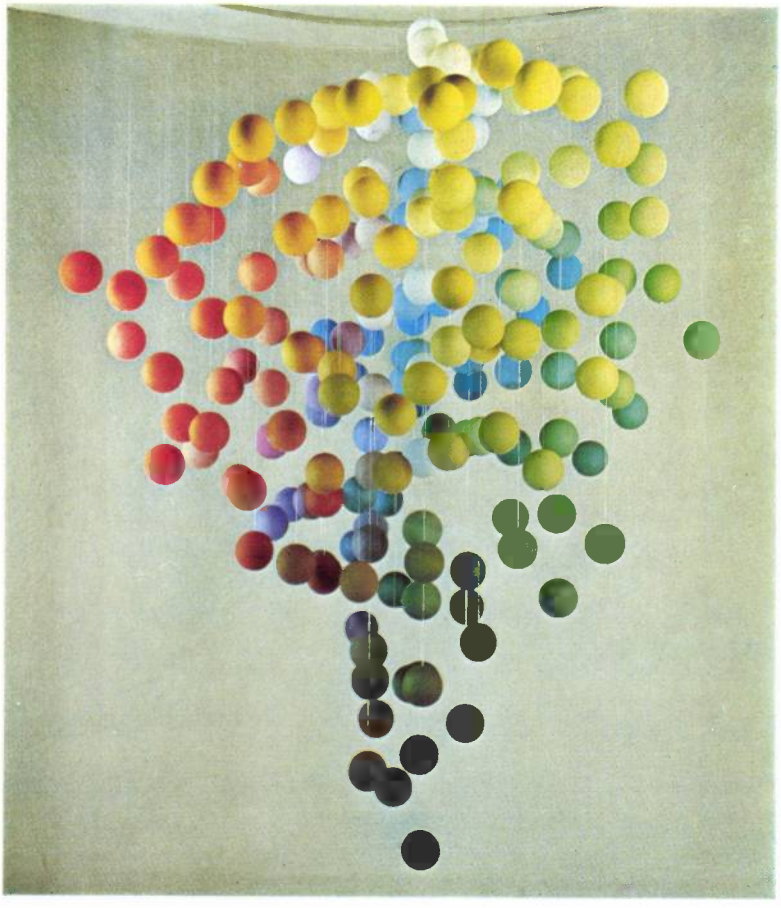
Business organization

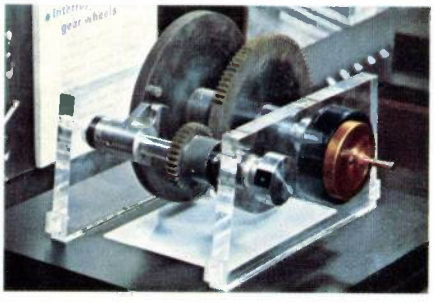


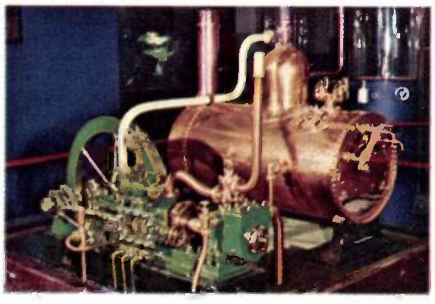





Geluidsspectroscoop
Soundspectroscope




gear wheels




Kas
Cash

Photo KLM Aerocarto N.V

The future

Now that we are at some distance from our work, we can see clearly the task emerging for us in the future. Gallery 3 was only partly successful in its concept. We ought to say here that Mr. Gardner originally took a very sceptical attitude to the depicting of such an abstract theme as "The significance of technology for society". Within this main theme the least successful was the subsidiary theme "Recreation", the very field in which Philips has the highest turnover. Many interesting ideas were proposed, but for one reason or another we failed to present this theme in a really satisfactory way.

Our current view is that although the six subsidiary themes should be represented on gallery 3, it is artificial to do so in separate sections. Where indeed can the line be drawn between recreation and education, and where in turn between these fields and communication?

This line of thought has led to an entirely new conception of gallery 3, in which the previous subsidiary themes repeatedly return in different forms. A start has already been made on carrying this new concept into effect, with a new and more flexible form of presentation.

Up to now we have referred solely to the planning, the construction and the success of the exhibition. We must also remember that such an exhibition cannot function without permanent care. This applies in particular to working exhibits and to those which the public can operate themselves. There must be an untiring team of guides and continuous renovation of exhibits. Every year at least 10% of the exhibits are improved or replaced. The Evoluon, in order to go on living, must itself evolve.

Finally, a word or two about the appeal of the Evoluon to young people. We believe that the educational value of an exhibition like that in the Evoluon lies in the fact that everyone is at liberty to skip something that at school they would be compelled to grapple with. Unfortunately things cannot be done in this way at school, but in the Evoluon everyone has the delightful freedom to concern himself with just what happens to take his interest at that moment. A lot can be learnt in this way. What appeals to youth, we trust, will also appeal to the youthful element in every visitor. And what constantly renews itself will never cease to fascinate.

Photo Bart Hofmeester, Rotterdam

Volume 31, 1970, No. 5/6 pages 125-204 Published 27th May 1971

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No. 7/8/9

Until quite recently, anyone speaking of a "transistor" would have been referring to the junction transistor, which has replaced the thermionic valve in so many fields in the last fifteen years. For individual circuit elements this would in the main still be quite correct. For integrated circuits, however, there is now another important contender in the field: this is the MOS transistor, which is a particular type of field-effect transistor.

Philips have been working intensively on these MOS transistors for several years, both in the laboratory and in production plants, in the Netherlands and in other countries. This issue of Philips Technical Review is completely given over to these activities. The first five articles are about the characteristics and technology of the MOS transistor; the next-to-last article is concerned with a medium-sized integrated circuit, and the last one with a large-scale integrated circuit. The other seven articles discuss small integrated circuits as well as particular transistors that have been made as part of the research programme to enable various special aspects such as high power or high cut-off frequency to be investigated.

Although many facets of the MOS transistors and the related Philips work are discussed, the contents of this issue by no means cover the whole field. There are for example no general considerations of MOS transistor circuits nor is there anything about complementary MOS transistors. Of course, not all of the contributions to a combined issue like this can be equally topical: fairly recent developments such as the LOCOS technique are accompanied by other topics that are not so new, but not necessarily therefore of lesser interest.

MOS transistors

L. J. Tummers

With the advances in semiconductor technology of a few years ago it became possible to produce a new solid-state amplifier, the MOS transistor (MOS for metal/oxide/semiconductor). When the term transistor is used, what is generally meant is the junction transistor. And in fact the junction transistor remains the only type to be widely used as a discrete element. A transistor of this type has three layers (*NPN* or *PNP*) and charge-carriers of both polarities play an important part in its operation; because of this they are sometimes called "bipolar transistors". In this type of transistor, and particularly in the middle layer (the base), the behaviour of the minority charge carriers is of great importance; in a *P*-type layer the minority carriers are the electrons, and in an *N*-type layer they are the holes.

Unipolar transistors depend for their operation on charge-carriers of only one polarity. This class includes field-effect transistors (sometimes called FETs). Here the current-flow in a strip of semiconducting material is modulated by applying a voltage to a control electrode, called the gate, which is insulated from the strip.

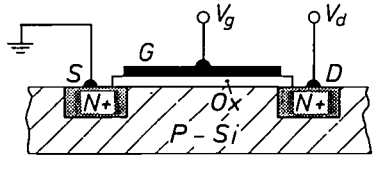
The idea underlying the operation of the field-effect transistor goes back to 1925 [1], but it was not until about 1950 that devices could be made that worked successfully [2]. In the first field-effect transistors to be produced the gate was not insulated from the semiconductor by a dielectric; it was itself a semiconducting layer, but of the opposite conduction type, and biased in the reverse direction with respect to the other material (this was the junction FET). This approach avoided difficulties arising from surface effects at the interface between the semiconductor and the insulating layer, which had prevented other types from working properly.

In about 1960 a new situation was created by the advent of the planar technique for making bipolar silicon transistors [3]. In doping of the silicon by diffusion this technique makes use of the masking properties of a layer of SiO_2 applied to the silicon. It was found that this SiO_2 layer could also act as an insulator between the semiconductor and the gate electrode of a field-effect transistor, and in this way the MOS transistor was born [5]. These devices were also not entirely free from difficulties due to surface effects, but with time they have increasingly been overcome [6]. In fact, the metal/oxide/semiconductor configuration has shown itself to be a wonderfully sensitive measuring instrument for investigating silicon surfaces.

From its very nature the MOS configuration permits a wide variation in device geometry. For example, MOS transistors with a very short channel have been made, which is an advantage for high-frequency operation, and others have been made with a very wide channel, which is necessary if the device is required to give a high output power. Examples of such devices will be found in other articles in this issue.

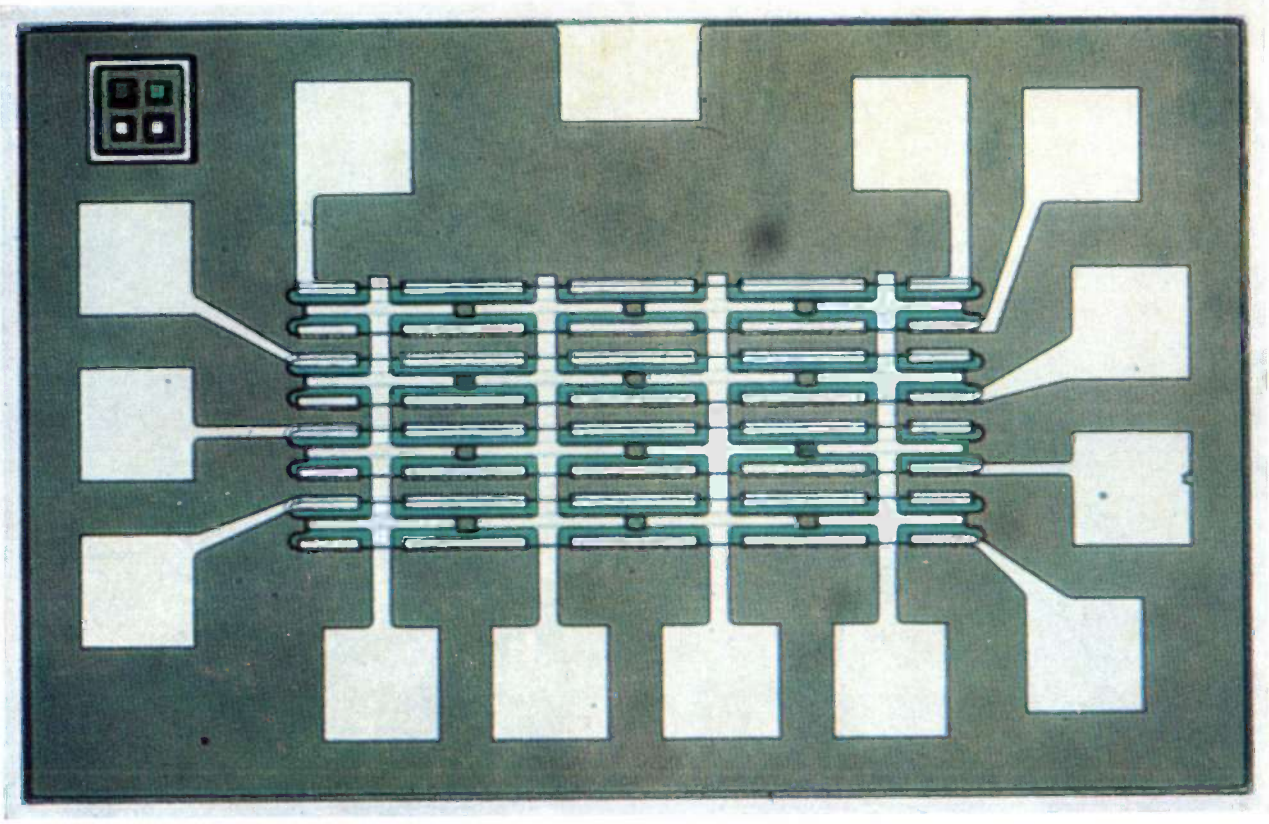
It is of course important to consider the practical advantages and disadvantages of the MOS transistor compared with the bipolar transistor. A striking difference is the very high input impedance of the MOS transistor, due to the presence of the oxide layer. This means that in applications where a very high input impedance is required the MOS transistor is clearly preferable. But there are other features that may favour the choice of a MOS transistor. Its characteristics are more linear than those of bipolar transistors and therefore introduce less distortion; less feedback can be used; simpler circuits can sometimes be used, and the device has good thermal stability. Another point in its favour is that the MOS transistor is a little easier to manufacture than a bipolar transistor.

Set against these good features there are others which



Schematic cross-section of a MOS transistor on a substrate of *P*-type silicon. *S* source. *D* drain. *S* and *D* are regions of strongly doped *N*-type silicon. *G* metal gate, insulated from the substrate by the SiO_2 layer *Ox*. The voltage between *D* and *S* is the drain voltage V_d, that between *G* and *S* is the gate voltage V_g. The electrons flow through the substrate from *S* to *D* in an extremely thin layer called the channel, situated under the oxide. Owing to the presence of the negative charge induced by V_g, the channel is in fact a layer of *N*-type silicon [4].

Prof. Ir. L. J. Tummers is a Deputy Director of Philips Research Laboratories, Eindhoven, and Professor Extraordinary in Transistor Technology at Eindhoven Technical University.



Example of an experimental integrated MOS circuit which includes very recent developments in both geometry and materials [8]. The circuit is a 16-bit store in which each bit is a MOS transistor with an adjustable threshold voltage. This voltage represents the information stored in the bit. The MOS transistors in this circuit differ from the conventional ones in that the insulation between the substrate and the gate is a sandwich formed by an SiO_2 and an Si_3N_4 layer. The circuit was made with the aid of the LOCOS technique.

make the MOS transistor less suitable for certain types of application. They explain why the MOS transistor has until now only been used on a very limited scale as a discrete circuit element. For operation at very high frequencies the bipolar transistor is superior because of its higher transconductance. While MOS transistors can be made with a high transconductance, the gate then has to be large, giving a high capacitance. The ratio g_m/C of the transconductance and the capacitance, which is in many cases a figure of merit for high-frequency behaviour, is therefore usually smaller for the MOS transistor than for the bipolar transistor. However, it is not completely impossible to make a MOS transistor with a high g_m/C ratio, but to achieve this the current I through the transistor has to be made rather large. This reduces the ratio g_m/I, which is a measure of the voltage gain, so that in effect we have only exchanged one disadvantage for another. Generally speaking, a MOS transistor at a given current setting requires a much higher load resistance to give a particular voltage gain than a bipolar transistor [7].

The high input impedance of the MOS transistor, which is a very attractive feature for certain applications, also gives rise to a difficulty. Electrostatic charge build-up may easily increase the voltage on the gate to a value high enough to cause breakdown in the oxide layer, destroying the device. It is often necessary to protect the gate with a diode or resistor.

Another disadvantage of the MOS transistor — although one that is diminishing with the advance of technology — is that its characteristics tend to be more susceptible to variations in the production process than those of the bipolar transistor. This applies particularly

[1] See for example H. C. de Graaff and H. Koelmans, The thin-film transistor, Philips tech. Rev. **27**, 200-206, 1966.
[2] W. Shockley, A unipolar "field-effect" transistor, Proc. I.R.E. **40**, 1365-1376, 1952.
[3] See A. Schmitz, Philips tech. Rev. **27**, 192, 1966.
[4] A more detailed treatment will be found in the article by J. A. van Nielen in this issue, page 209.
[5] D. Kahng and M. M. Atalla, Silicon-silicondioxide field-induced surface devices, IRE Solid-State Device Research Conference, Pittsburgh 1960.
[6] See the article by J. A. Appels, H. Kalter and E. Kooi in this issue, page 225.
[7] See, for example, the article by J. A. van Nielen [4].
[8] This circuit was made by Ir. R. H. W. Salters of Philips Laboratories, Eindhoven.

to the threshold voltage, i.e. the gate voltage at which the transistor just starts to conduct.

Although at present there may be little likelihood of the large-scale use of MOS transistors as individual circuit elements, the situation is entirely different for their use in integrated circuits. In integrated circuits using bipolar transistors the individual circuit elements comprised in a single silicon chip have to be isolated from one another, which involves a number of separate steps in the production process. No such isolation is required in MOS circuits, and this not only makes the fabrication process simpler but also saves space on the silicon chip. A further saving of space is achieved by using the MOS transistor as a load resistor. In this way complex integrated circuits can be made that consist entirely of MOS transistors.

With their high impedance levels, integrated MOS circuits are not particularly fast, but for very large integrated arrays where speed requirements are not critical the MOS transistor has a slight advantage. Research is still going on in both fields, however, and it is therefore difficult to make any forecast.

The most advanced investigations now in progress or recently completed relate to refinements in device geometry and materials. The hope is that these refinements will allow MOS transistors to be made that can be used up to higher frequencies, that have a very low or adjustable treshold voltage, etc. In the refinements of the geometry one of the aims is the more accurate registration of the gate with the source and drain regions, and experiments are being carried out both with variants of the classical methods and with entirely new methods, such as doping the material by bombardment with fast ions (ion implantation). In the ion-implantation technique, and in other techniques as well, the gate electrode itself is used as a mask, thus simplifying registration.

In the study of the materials, efforts are being made to find insulators that can be used instead of SiO_2, or in combination with it, and also other materials for the gate. Here the search is not only for metals suitable to replace the aluminium used at present — the nature of the metal partly determines the threshold voltage — but devices are also being made with non-metals, such as polycrystalline silicon.

In the development of the integrated bipolar or MOS circuits there has been a fair amount of cross-fertilization. Not only have the potentialities of the MOS circuits stimulated the search for new ideas in the field of bipolar circuits, but progress in the one field has also frequently proved of value to the other. For example, the research on silicon surfaces that made the MOS structure possible has also given valuable help in the development of the bipolar transistor.

There is every indication of this pattern of cross-fertilization between the two fields continuing in the near future.

Summary. The MOS transistor (MOS = metal/oxide/semiconductor) originated in about 1960. It is a field-effect transistor whose metal gate electrode is isolated from the silicon substrate by an oxide layer. Previous efforts to make field-effect transistors with an isolated gate failed because of undesirable effects at the semiconductor/insulator interface. Advantages of the MOST compared with the bipolar transistor are its very high input impedance, better linear characteristics and lower feedback. It has high thermal stability and is a little easier to make than the bipolar, but the latter is better at high frequencies. MOST research is now directed towards refining the geometry and improving the materials. In integrated circuits using MOST devices the elements do not have to be isolated. For large integrated circuits that do not have to be particularly fast there is some advantage to be gained by using the MOST.

Operation and d.c. behaviour of MOS transistors

J. A. van Nielen

The MOS transistor (MOS = metal/oxide/semiconductor) is a type of field-effect transistor. It may be regarded as a resistor of semiconducting material whose conductivity is determined by the potential of a control electrode (the gate) situated outside the current path. The gate of a field-effect transistor is isolated from the current path either by an insulating layer or by a reverse biased *P-N* junction. In the MOS transistor there is an insulating layer.

MOS transistors are made on a relatively thick substrate of monocrystalline, lightly doped silicon by means of the planar technique; a schematic cross-section can be seen in the previous article (page 206). The insulator between semiconductor and gate is a layer of SiO_2, obtained by oxidation of the silicon. The source and drain electrodes are heavily doped zones of the opposite conduction type from that of the substrate. They are produced by diffusion, sometimes with the aid of a bombardment by fast ions (ion implantation).

There are two types of MOS transistor: those on an *N*-type substrate and those on a *P*-type substrate. In the first type current conduction takes place by the flow of holes from the source to the drain, and the device is called a *P*-channel MOS transistor (*fig. 1a*). The other type, in which the conduction is due to electrons, is called an *N*-channel MOS transistor (fig. 1*b*). We shall return to the current conduction in an MOS transistor in more detail later.

The relation, at constant gate voltage V_g, between the current I_d flowing through a MOS transistor and the potential V_d of the drain — we assume the potential of the source to be zero — much resembles the I_a-V_a characteristic of a pentode valve. Characteristics of an *N*-channel MOS transistor are shown in *fig. 2a*. The I_d-V_g characteristic takes the form of a quadratic function (fig. 2*b*); in a MOS transistor I_d is zero when V_g is less than a certain voltage V_{th}, called the threshold voltage. The transconductance of MOS transistors is usually between 1 and 10 mA/V. A special feature of the device is its exceptionally high input resistance, of the order of 10^{14} ohms.

In this article we shall first deal at somewhat greater length with the operation of a MOS transistor and consider the factors that determine the threshold volt-

Ir. J. A. van Nielen is with Philips Research Laboratories, Eindhoven.

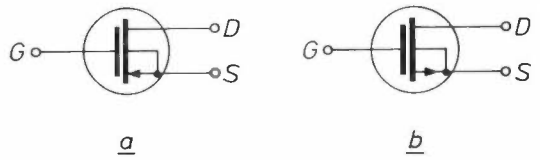


Fig. 1. Schematic representation of the MOS transistor, *a*) with *P*-channel, *b*) with *N*-channel. *S* source. *G* gate. *D* drain. The diagram shows that the substrate is connected to the source; this connection is not always present.

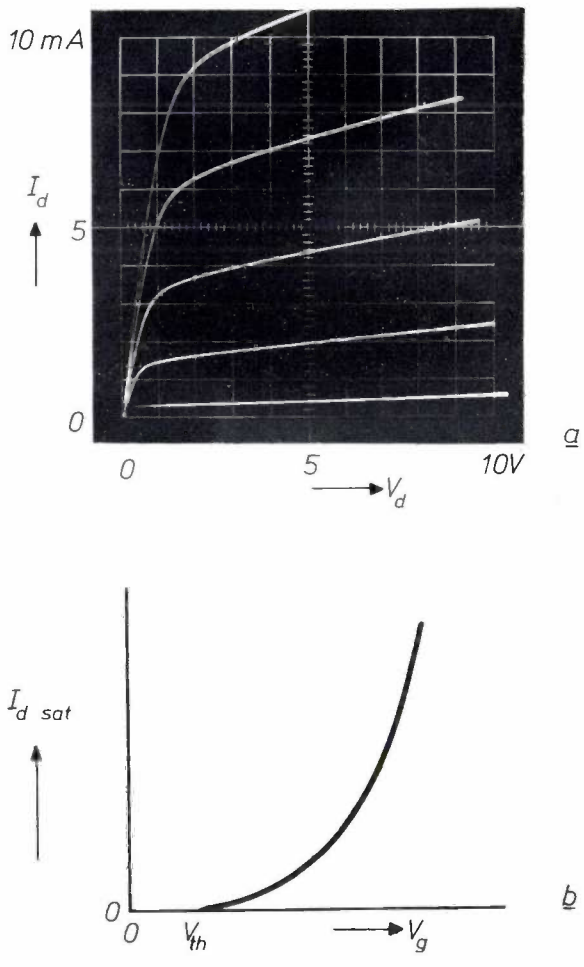


Fig. 2. *a*) Some I_d-V_d characteristics of a MOS transistor. Each characteristic relates to a particular value of V_g and consists of a curved region and an almost horizontal region (saturation region).
b) An I_d-V_g characteristic relating to a value of V_d in the saturation region. The variation of the saturation value $I_{d\,sat}$ of the current with V_g is a quadratic function and I_d is zero when V_g is lower than the threshold voltage V_{th}.

age V_{th}. We shall then derive some approximate equations for the d.c. voltage characteristics, and finally briefly discuss the way in which the behaviour of a MOS transistor depends on the doping and potential of the substrate.

Operation of an MOS transistor

Let us consider the case of a MOS transistor on a *P*-type substrate as illustrated in the previous article (page 206). The source *S* and the substrate form a diode and similarly the drain *D* and the substrate form a diode. When a voltage is applied between *S* and *D*, at $V_g = 0$, then one of these two diodes is biased in the reverse direction and only an extremely weak current flows through the transistor.

We shall now see what happens when V_g is gradually increased from zero, and we start with the simple case where no voltage has been applied between *S* and *D* ($V_d = 0$). The system constituted by gate, oxide layer and silicon can be regarded as a capacitor whose lower plate is not a metal, but a *P*-type semiconductor. As long as $V_g = 0$ the charge on both plates of the capacitor is zero and the semiconductor is everywhere electrically neutral. When $V_g > 0$, a positive charge appears on *G* and a negative charge of equal magnitude appears on the semiconductor in a layer next to the oxide. At first this charge is carried solely by the (immobile) acceptor ions. Since their density is determined by doping of the substrate, and is thus fixed, this layer is thicker the higher the positive charge, i.e. the greater the magnitude of V_g. Since the positive, mobile charge carriers are driven out of this layer, it is referred to as a depletion layer.

When V_g is further increased, the negative charge in the semiconductor is then no longer carried by the acceptor ions alone, but also by electrons which now appear in a very thin layer of the depletion region next to the oxide. This layer, which may for example be 10^{-2} μm thick, forms a conducting connection between the source and drain and is therefore referred to as the channel.

In *fig. 3* this picture is presented in the form of an energy-band diagram. The bands are of course curved in the depletion region. The curve for the Fermi energy E_F, however, will be a horizontal straight line, as the semiconductor is everywhere in thermodynamic equilibrium. When V_g is sufficiently high, E_F at the interface of the silicon and silicon dioxide may therefore come to lie above E_i, the centre of the forbidden zone. This means that the silicon at that position has changed to the other conduction type. This change, which has nothing to do with a change in doping, but is only present when V_g is high enough, is called inversion. The inverted layer is the channel. Since V_g determines the electron concentration in the channel, it also determines the conductivity of the channel and hence the current flowing through the transistor when $V_d \neq 0$. The minimum gate voltage needed at $V_d = 0$ to bring about inversion is the threshold voltage V_{th} mentioned above.

When a gate voltage sufficient to give inversion is applied the electron concentration in the channel very quickly adjusts itself to the value appropriate to the new situation, since most of the electrons required are supplied from *S* and *D*, where there are large numbers of electrons available. The new equilibrium thus comes about much more quickly than if the electrons were to come into the conduction band as a result of thermal excitation alone.

We now consider the situation that actually applies in an operating MOS transistor: here $V_d \neq 0$. The potential $V(x)$ in the channel is then a function of the coordinate x along the channel, and gradually increases going from *S* to *D*. The extent of the inversion, or the conductivity of the channel, therefore gradually decreases from *S* to *D*. The thickness of the depletion layer, on the other hand, gradually increases, because

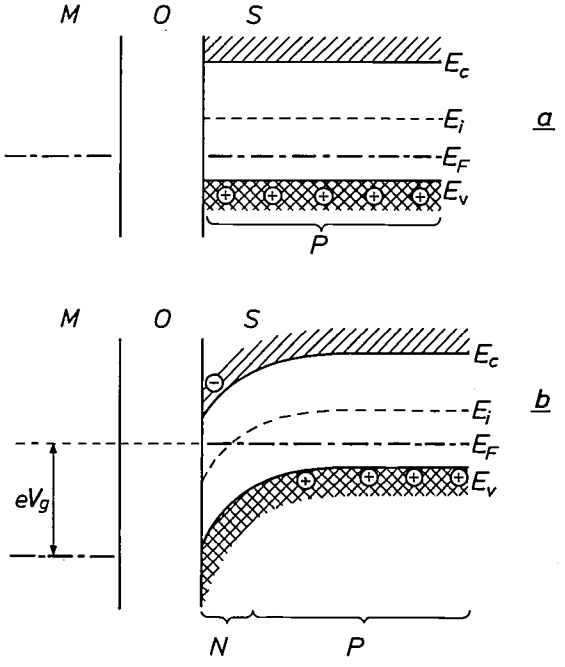


Fig. 3. *a*) Energy-band diagram of the metal/oxide/silicon system (*M*, *O* and *S*) when the silicon is *P*-type and there is no natural band curvature near the interface of the semiconductor and the oxide. The Fermi energy E_F in the silicon is lower than the middle E_i of the forbidden zone (E_v is the upper edge of the valence band, E_c the lower edge of the conduction band). Since metal and silicon are at the same potential, E_F has the same value in both.
b) When the metal (the gate electrode) is raised to a potential V_g, the holes in the silicon are driven from the zone at the interface, giving rise there to a negative space charge, carried by ionized acceptor ions. The corresponding band curvature may be so great that E_i at the interface comes below E_F, so that an *N*-type layer is formed there (an inversion), the "channel".

the voltage across the *N-P* junction formed by the channel and the rest of the substrate rises from *S* to *D* (*fig. 4a*).

If we now gradually raise the potential of *D*, at fixed V_g, then the current I_d also rises, but owing to the decrease in the conductivity of the channel the increase of I_d gradually becomes less steep as V_d becomes higher. At a particular value of V_d the effective gate voltage $V_g - V_d$ at the end of the channel, where $V(x) = V_d$, has decreased to V_{th}. At that position the condition for inversion is no longer fulfilled; the channel is said to be "pinched off" (fig. 4*b*). This value of V_d, which is equal to $V_g - V_{th}$ and is denoted by $V_{d\,sat}$, is called the "pinch-off voltage". If V_d is increased further the point in the channel where $V(x) = V_{d\,sat}$, the pinch-off point, is shifted in the direction of *S*. In the region of V_d values at which the channel is pinched off, the current I_d varies much less strongly with V_d (saturation region; see fig. 2*a*).

Later in this article, when we present an approximate theory of the d.c. current behaviour of the MOS transistor, we shall return to the current-saturation effect. First, however, we shall take a closer look at the threshold voltage V_{th}.

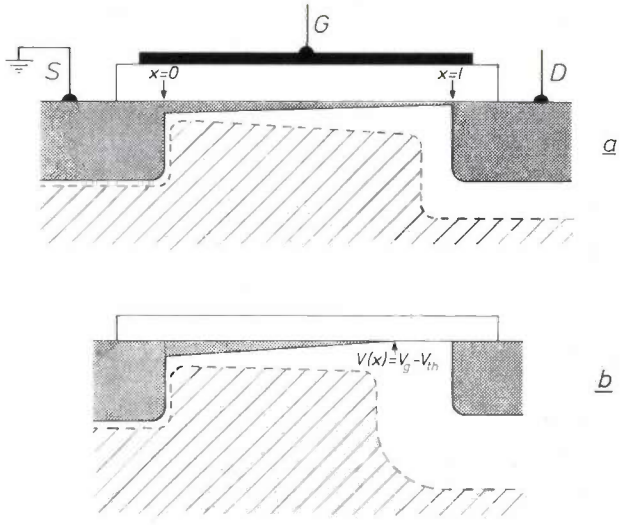


Fig. 4. *a*) Schematic picture of the situation in a non-saturated MOS transistor on a *P*-type substrate. Situated immediately under the oxide is an extremely thin, inverted layer of the substrate, through which the current flows and whose conductivity gradually decreases from *S* to *D*. The latter is schematically represented by diminishing thickness. Under the channel, and also under the drain, is a layer in which there are no mobile carriers, the depletion layer (*white*); its thickness increases going from *S* ($x = 0$) to *D* ($x = l$).
b) The same for the case of current saturation. In the last part of the channel $V(x) > V_g - V_{th}$.

The threshold voltage V_{th}; four types of characteristic

In the simple case outlined above, where the energy bands at the boundary surface were not curved in the semiconductor at $V_g = 0$ and $V_d = 0$, the gate voltage V_{th} required to produce inversion could only be positive. In practice the situation is usually not as simple as that, because the structure of the MOS transistor may already contain built-in charges. In MOS transistors on a *P*-type substrate V_{th} may be negative. Such transistors conduct even at $V_g = 0$. For MOS transistors on an *N*-type substrate it is also possible in principle for V_{th} to be either positive or negative. Four types of characteristic may therefore be encountered (*fig. 5*). The MOS transistors that conduct at $V_g = 0$ — i.e. those in which a channel is naturally present — are called *depletion-type* MOS transistors, and the others are said to be of the *enhancement type*. These names come from the use of the MOS transistor as switching devices in digital circuits. Devices of the depletion type are normally open and require to be closed by a gate voltage, which depletes the conducting channel; those of the enhancement type are normally closed and require to be opened by a gate voltage.

Apart from the relative positions of E_i and E_F, i.e. the doping of the substrate, there are three other factors that determine the value of V_{th} found in a MOS transistor. The first is that, during the oxidation of the silicon, a quantity of positive charge q_{ox} enters the oxide, which has the effect of shifting V_{th} in the negative direction. Because of this effect *P*-channel transistors usually belong to the enhancement type and *N*-channel transistors to the depletion type.

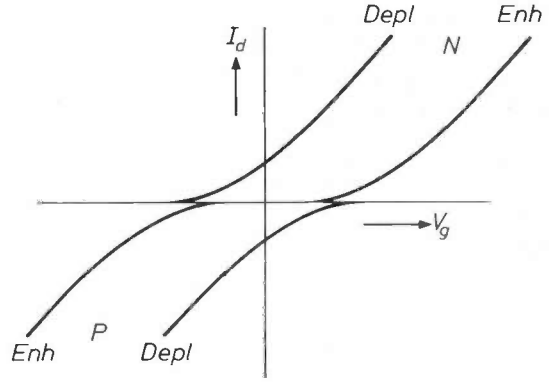


Fig. 5. In MOS transistors on either a *P*-type or an *N*-type substrate the threshold voltage V_{th} may in principle be either positive or negative, so that there are four types of I_d-V_g characteristic. MOS transistors which conduct when $V_g = 0$ are "depletion-type" devices (*Depl*) and the others "enhancement-type" devices (*Enh*).

In the second place V_{th} depends to some extent on the metal from which the gate electrode is made. The difference between the work function of this metal and that of the substrate, the contact potential Φ_{ms}, acts as a built-in contribution to the gate voltage (*fig. 6a, b*). Finally there may be effects from surface states whose energy levels lie in the forbidden band. These surface

states may trap free charge carriers, which can then make no contribution to the conduction. The larger the number of surfaces states, the higher the gate voltage needed to produce a particular concentration of free charge carriers, in other words the threshold potential $|V_{th}|$ is higher. Advances in recent years in the technology of manufacturing MOS transistors have made it possible to keep the concentration of surface states so low that they no longer have any significant effect [1].

If we include all these charges then the charge on the two plates of the capacitor formed by the MOS transistor is:

$$q_g + q_{ox} = -(q_{inv} + q_{ss} + q_{depl}), \quad \ldots \quad (1)$$

where q_g is the charge per unit area on the gate, q_{inv} the charge in the channel, q_{ss} that in the surface states and q_{depl} that in the depletion layer. If we take the usual definition of the threshold voltage V_{th} as the value of V_g at which the energy difference $E_i - E_F$ at the interface between the silicon and the SiO_2 is equal to that in the bulk of the substrate but of opposite sign (fig. 6c), then:

$$V_{th} = \Phi_m + 2\,\Phi_F - (q_{inv} + q_{ss} + q_{ox} + q_{depl})/C_{ox}. \quad \ldots \quad (2)$$

Here C_{ox} is the capacitance per unit area of the capacitor. The term $2\,\Phi_F$ is in this case the band curvature caused by V_g. The charge q_{inv} is usually so much smaller than the other charges that it may be neglected in establishing a value for the threshold potential.

Approximate theory of the d.c. current behaviour

Let us again consider the case of an MOS transistor on a substrate of P-type silicon, i.e. with a channel in which the conduction is by electrons. To calculate the characteristics we introduce two simplifications. In the first place we assume that everywhere in the layer of the substrate adjacent to the oxide — the lower face of the capacitor — the charge per unit area $q(x)$ is determined by the difference between V_g and the potential $V(x)$ at the position x in the channel, the relationship being:

$$q(x) = -C_{ox}\{V_g - V(x)\}. \quad \ldots \quad (3)$$

(Expressed in the symbols used in equation (1), $q(x) = q_{inv} + q_{depl} + q_{ss}$.) Let h be the thickness of the oxide layer and ε_{ox} its relative dielectric constant, then $C_{ox} = \varepsilon_0\varepsilon_{ox}/h$. The minus sign in (3) indicates that the charge $q(x)$ is negative if $V_g - V(x)$ is positive.

Equation (3) would apply exactly if the lines of force in the dielectric were perpendicular to the surface. This situation is approximated when $V(x)$ in the channel does not vary too greatly with x, that is to say when $dV(x)/dx$ and $d^2V(x)/dx^2$ are small (this is Shockley's gradual approximation [2]); in practice this is the case in the greater part of the channel.

The second assumption is that the density $q_{inv}(x)$ of the negative charge carried by mobile electrons in the channel is given by:

$$q_{inv}(x) = -C_{ox}\{V_g' - V(x)\}, \quad \ldots \quad (4)$$

where $V_g' = V_g - V_{th}$.

If this equation can also be used, the current, which

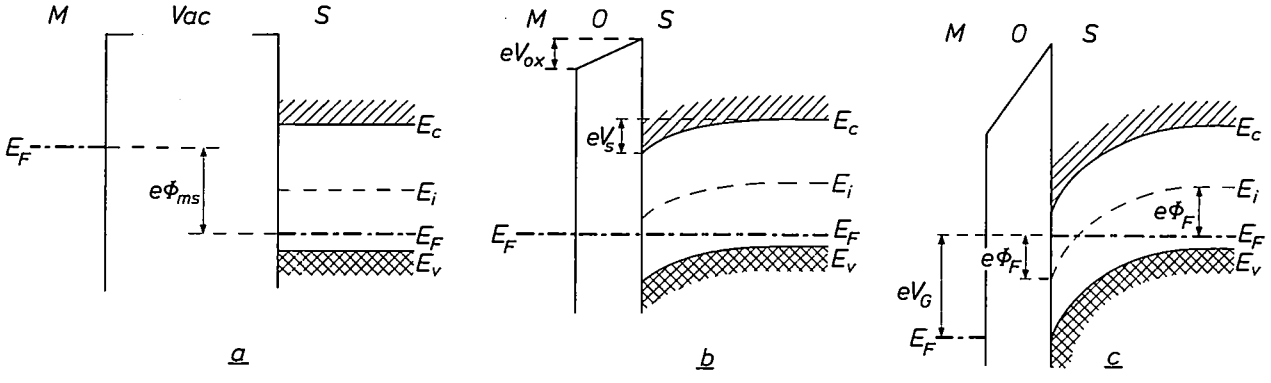


Fig. 6. *a*) Energy-band diagram of a metal M and a semiconductor S which are not in contact with each other and have a different work function (i.e. the work that must be performed in order to remove an electron with an energy E_F from the metal or the semiconductor into a vacuum). The difference in work function is $e\Phi_{ms}$; Φ_{ms} is the contact potential.
b) If the metal and semiconductor in (*a*) are joined by an oxide layer to form an MOS structure, then E_F has the same value everywhere and the original difference $e\Phi_{ms}$ brings into existence a voltage V_{ox} across the oxide and a band curvature eV_s in the semiconductor; $V_{ox} + V_s = \Phi_{ms}$. Because of the band curvature, V_{th} has here a value different from that in the case illustrated in fig. 3.
c) As (*b*), but for the case where a gate voltage V_g is applied such that $E_i - E_F$ at the surface is exactly equal to $-e\Phi_F$.

is equal to the product of $q_{inv}(x)$, the width w of the channel, the mobility μ of the charge carriers and the field strength $-dV(x)/dx$, is given by:

$$I_d(x) = \mu C_{ox} w\{V_g' - V(x)\}dV(x)/dx. \quad . . \quad (5)$$

We can now find an expression for the steady-state current by integrating equation (5) over the whole length l of the channel. Since I_d is independent of x in the steady state, we may place I_d in front of the integral sign:

$$I_d \int_0^l dx = \mu C_{ox} w \int_0^{V_d} \{V_g' - V(x)\}dV(x).$$

From this we derive:

$$I_d l = \mu C_{ox} w\{\tfrac{1}{2} V_g'^2 - \tfrac{1}{2}(V_g' - V_d)^2\}$$

or

$$I_d = \beta(V_g' V_d - \tfrac{1}{2}V_d^2), \quad . . . . \quad (6)$$

where

$$\beta = \mu C_{ox} w/l. \quad . . . . . \quad (7)$$

The curve corresponding to equation (6) is a parabola with the apex upwards.

In these calculations we have tacitly made a third assumption, which is that μ depends on none of the other quantities. This is not entirely true, because if V_g is high, i.e. if there is a strong transverse field, μ is somewhat smaller than when V_g is low [3].

In the special case where $V_d = 0$, we may deduce from (6) the following expression for the conductivity G_0 of the channel:

$$G_0 = \lim_{V_d \to 0} I_d/V_d = \beta V_g'. \quad . . . \quad (8)$$

Within the limits of our approximation the conductivity thus varies linearly with V_g (*fig. 7*).

Let us now return for a moment to equations (4) and (5). The situation in the channel is that, going from S to D, the charge density gradually decreases; the field strength, on the other hand, increases, with the effect that $I_d(x)$ has the same value everywhere in the channel. At the value of V_g where the right-hand side of (6) reaches its maximum — i.e. where $V_g' = V_d$, or where $V_g = V_d + V_{th}$ — equation (4) shows that the charge density $q_{inv}(x)$ at the drain ($x = l$) is equal to zero, which means that the field strength there would be infinitely high. The gradual approximation is therefore no longer valid here; only the rising part of the parabola represents a part of the I_d-V_d characteristic. The maximum current can readily be shown from (6) to be equal to $\frac{1}{2}\beta V_g'^2$.

In the region $V_d > V_g'$ we may expect as a first approximation that the current will be independent of V_d and equal to this maximum value, on the following grounds. If we let the voltage V_d increase above V_g', then the potential $V(x)$ in the channel reaches the value V_g' at a point just before the end of the channel; the remainder of V_d appears across the part that lies between this point and the drain. Since the conductivity of this part is low, and yet the potential difference $V_d - V_g'$ still causes a current flow there, the length of this part will be relatively small. The length of the highly conducting part of the channel thus differs relatively little from l, and the current will be approximately equal to the maximum current (the current at saturation). $I_{d\,sat}$ denotes the saturation value of the current, we may thus write:

$$I_{d\,sat} = \tfrac{1}{2}\beta V_g'^2. \quad . . . . . \quad (9)$$

This equation shows that the current through a MOS transistor operating in the saturation region is a quadratic function of the gate voltage (see fig. 2*b*). The

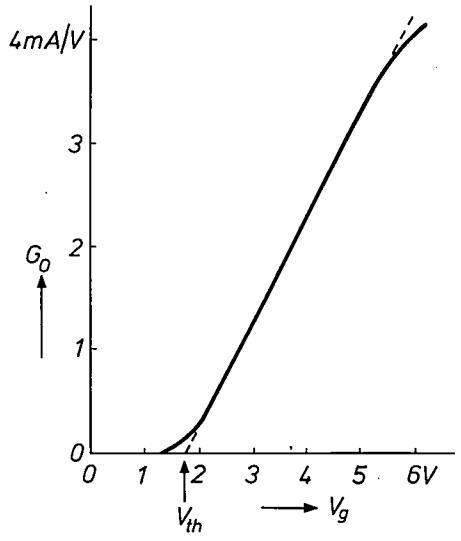


Fig. 7. The conductivity G_0 of a MOS transistor with $V_d = 0$ varies linearly with V_g over a wide range. The solid curve relates to the results of measurements, the dashed line relates to equation (8).

transconductance g_m in this region is given by:

$$g_m = (\partial I_{d\,sat}/\partial V_g)_{V_d} = \beta V_g'. \quad . . . \quad (10)$$

The transconductance is not constant but varies linearly with V_g'. The calculation given here also shows that g_m is equal to G_0 (see equation 8).

As could be seen from fig. 2*a*, the current in the saturation region is in reality not entirely independent

[1] See the article by J. A. Appels, H. Kalter and E. Kooi in this issue, page 225, and also H. C. de Graaff and J. A. van Nielen, Electronics Letters **3**, 195, 1967.

[2] W. Shockley, A unipolar "field-effect" transistor, Proc. I.R.E. **40**, 1365-1376, 1952.
C. T. Sah, Characteristics of the metal-oxide-semiconductor transistors, IEEE Trans. **ED-11**, 324-345, 1964.

[3] See the article by N. St. J. Murphy, F. Berz and I. Flinn in this issue, page 237.

of V_d. One of the reasons for this is that near the drain the lines of force from the gate are no longer perpendicular to the interface between oxide and channel, which is the assumption made in the gradual approximation. On increasing V_d the distribution of the lines of force in insulator and substrate around the pinch-off point of the channel varies in such a way that the pinch-off shifts slightly towards the source, making the channel shorter. This increases the transconductance of the device (see equations 10 and 7) and also, since there is no change in the gate voltage, it increases the current as well.

Transconductance and gain

The characteristics and relations arrived at in our theoretical treatment are of significance in the practical application of the MOS transistor. Very often a maximum voltage gain is required from the MOS transistor. The voltage gain $|\Delta V_d/\Delta V_g|$ in the amplifier circuit of *fig. 8* is approximately equal to the product of the transconductance g_m and the load resistance R_l. Now g_m in the MOS transistor increases with the d.c. current I_d; equations (9) and (10) show that the relation is:

$$g_m = (2\beta I_d)^{\frac{1}{2}}. \quad \ldots\ldots \quad (11)$$

This means that to obtain a high voltage gain the user will be inclined to bias the transistor to a high current. At a given supply voltage V_{dd} a limit is set to this, however, by the maximum permissible voltage drop across R_l, i.e. $I_d R_l$. Therefore not g_m, but the quantity g_m/I_d is a measure of the available voltage gain in any given circuit [4]. As a rule the transconductance of MOS transistors is smaller than that of bipolar transistors, and a higher load resistance is therefore needed to obtain an equally high voltage gain.

The purely second-order characteristic of the MOS transistor (see equation 9) is an advantage for applications in receiver input stages. If the valve or transistor in this stage has a non-linear characteristic, strong signals outside the passband of the receiver can introduce spurious signals (intermodulation products) by interaction with the desired signal. With a selective receiver, most of these intermodulation products do not appear within the passband of the following stages and introduce no interference. If however the expression for the characteristic of the valve or transistor contains higher terms in odd powers of the input voltage, the intermodulation products will include a signal at the carrier frequency of the desired station, but with its amplitude determined by the modulation of the interfering station. The programme from the interfering station appears to be modulating the carrier from the desired station: this effect is called "cross-modulation". A MOS transistor, whose characteristic does not contain such higher odd terms, does not introduce cross-modulation [5].

The substrate

To conclude this article we shall examine how the behaviour of a MOS transistor is affected by the doping of the substrate and by a substrate potential V_b differing from zero [6]. We again consider a transistor on a *P*-type substrate and again start with the case where $V_d = 0$ and a gate voltage V_g is applied to produce an inversion layer. We have already seen that this layer is extremely thin, and that the depletion layer is relatively thick (e.g. 1 μm) because the charge density in it cannot be greater than the density N of the acceptor ions present in the silicon. The charge density in the depletion layer is therefore constant over a large part

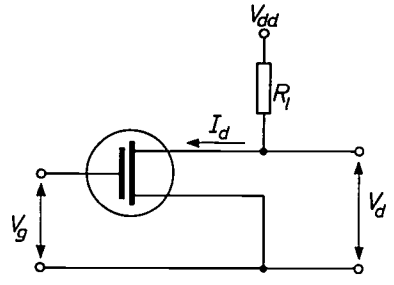


Fig. 8. Circuit for voltage amplification with a MOS transistor. R_l load resistance. V_{dd} supply voltage.

of the thickness, and there is a fairly sharp boundary between this layer and the rest of the substrate.

To calculate the space charge in the depletion layer we may regard the channel and depletion layer together as an abrupt N^+-P junction and apply the appropriate equations. If a particular V_g produces a voltage V_s across the N^+-P junction — the band curvature is then $-eV_s$ (fig. 6*b*) — then the thickness d of the depletion layer is:

$$d = (2\,\varepsilon_0 \varepsilon V_s/eN)^{\frac{1}{2}}, \quad \ldots\ldots \quad (12)$$

and the depletion charge per unit area q_{depl} is:

$$q_{depl} = -(2\,\varepsilon_0 \varepsilon e N)^{\frac{1}{2}}\, V_s^{\frac{1}{2}} = a V_s^{\frac{1}{2}}. \quad \ldots \quad (13)$$

In the approximate theory leading to equations (4) and (6), q_{depl} is taken to be zero or regarded as constant, i.e. independent of V_g and x. In calculating the effect of V_g on q_{depl} this is a very good approximation, because provided the inversion is not unduly small — i.e. provided the charge density in the channel is greater than that in the depletion layer — a variation of V_g influences the electron concentration in the channel very much more than the depletion charge [7]: the first varies exponentially with V_s, whereas q_{depl} only

varies with $V_s^{\frac{1}{2}}$, as we have seen. This means that when the gate voltage V_g is made high enough, the depletion charge remains practically constant, whereas the mobile charge increases linearly with V_g.

When $V_d \neq 0$, the potential $V(x)$ in the channel goes from 0 to V_d, and similarly the reverse voltage across the induced N^+-P junction also goes from 0 to V_d. This reverse voltage can be further increased by giving the substrate an additional reverse bias V_b with respect to the source. The total depletion charge per unit area is then $a\{V_s + V(x) + V_b\}^{\frac{1}{2}}$ and is thus dependent on x. The increase in the depletion charge due to the contributions from $V(x)$ and V_b takes place at the expense of the mobile charge $q_{inv}(x)$, since the total charge is still that given by equation (3). The direct consequence of this is that the current reaches saturation at a drain voltage lower than $V_g - V_{th}$ and the saturation value $I_{d\,sat}$ is also smaller than that given by equation (9). The transconductance is now also lower and no longer equal to G_0.

The effect of applying V_b is thus to change the amount of mobile charge, and with it the current. The substrate contact may therefore be regarded as a second gate electrode. If the doping N of the substrate is 10^{16} per cm^3 the transconductance can in fact be just as high when the device is driven via the substrate as when it is driven via the insulated gate electrode G. We should also note that when $V_b \neq 0$ the value of the threshold potential is of course different:

$$V_{th}(V_b) = V_{th}(0) + \{(V_b + 2\Phi_F)^{\frac{1}{2}} - (2\Phi_F)^{\frac{1}{2}}\}a/C_{ox}.$$

As can be seen from (13), q_{depl} is proportional to $N^{\frac{1}{2}}$. Equations (6), (9) and (10) for I_d, $I_{d\,sat}$ and the transconductance g_m, respectively, will thus become more accurate as N decreases, i.e. as the doping is reduced; the deviation is already very small for $N = 10^{14}$/cm^3. In spite of the deviations from these equations when the substrate is more strongly doped, it is still fairly accurate to take the variation in transconductance with V_g as linear and the variation of the saturation current $I_{d\,sat}$ with V_g as a quadratic function.

Punch-through

We should note here an unwanted effect that may occur in the substrate when the drain voltage is too high, particularly when the substrate is weakly doped. The higher the voltage of the drain with respect to the substrate, the wider becomes the depletion zone around the drain. In MOS transistors with a short channel this zone may become so wide that it touches the source. The electric field in the depletion region then acts directly on the charge carriers in the source diffusion, causing the carriers to move outside the channel to the drain. This effect is known as *punch-through*. It sets a limit to the improvement in the high-frequency characteristics of a MOS transistor that can be achieved by making it with a shorter channel; it also limits the maximum drain voltage in high-frequency power transistors [8].

Summary. A MOS transistor is a field-effect transistor in which the gate is insulated from the semiconductor (silicon) by a layer of SiO_2. If the semiconductor is P-type the source and drain are strongly doped zones of N-type silicon. The current flows through a thin layer at the surface of the oxide, called the channel, in which the silicon is changed by the field of the gate from P-type to N-type. The input resistance is very high (about 10^{14} Ω). The current-voltage characteristic (when the gate voltage is constant) shows saturation, just as in the case of a pentode valve. The transconductance varies linearly with the gate voltage and is of about the same magnitude as in thermionic valves. If the gate voltage has a magnitude lower than a certain threshold voltage, no current flows. This threshold voltage in both an N-type and P-type transistor may be either positive or negative. The doping of the silicon substrate makes the behaviour of the MOS transistor deviate from the theoretical description given in the article; with a doping level of $N = 10^{14}$/cm^3 or less, however, the deviations are small and can be neglected.

[4] G. Klein and H. Koelmans, Active thin film devices, Festkörperprobleme 7, 183-199, 1967.
[5] R. J. Nienhuis, A MOS tetrode for the UHF band with a channel 1.5 μm long; this issue, page 259.
[6] J. A. van Nielen and O. W. Memelink, The influence of the substrate upon the DC characteristics of silicon MOS transistors, Philips Res. Repts. **22**, 55-71, 1967.
[7] A. S. Grove, B. E. Deal, E. H. Snow and C. T. Sah, Solid-State Electronics **8**, 145, 1965, and T. I. Kamins and R. S. Muller, Solid-State Electronics **10**, 423, 1967.
[8] See the article by R. D. Josephy in this issue, page 251.

The MOS transistor as a small-signal amplifier

P. A. H. Hart and F. M. Klaassen

Introduction

Under the usual conditions of operation the MOS transistor operates in saturation and is a square-law device: the characteristic curve of the drain current as a function of the gate voltage is parabolic in shape (see the preceding article [1], fig. 2*b*). However, if the MOS transistor is biased to bring the operating point on to the slope of the parabola, then the amplification will be practically linear for small signals, since a small portion of the slope of the parabola near the operating point approximates to a straight line.

The very high input impedance of the MOS transistor makes it an attractive device for use in amplifier circuits, especially in the input stages. It has applications both in the audio frequency range [2] and at high frequencies [3].

For these applications it is necessary to know what gain the MOS transistor will give at these frequencies, and also its noise characteristics. This information can be obtained from a model of the MOS transistor in the form of a network of electrical elements (*fig. 1*). For a fairly wide range of frequencies the network can be greatly simplified to give a useful equivalent circuit. This can be used as the basis for designing the circuit in which the MOS transistor is to be included. For more general calculations of the maximum available gain of the MOS transistor, this equivalent circuit can be reduced to a representation of the MOS transistor as a linear four-terminal network (or two-port), characterized by four complex quantities. The general theory of linear four-terminal networks can then be applied.

The noise generated in a MOS transistor is mainly thermal noise, originating in the conducting channel. The magnitude of this noise can readily be derived from the elements of the equivalent circuit or from the equivalent four-terminal network. This is not the case for flicker noise, which is predominant at low frequencies and appears to be connected with the behaviour of the charge carriers at the interface between the silicon and the oxide surface layer. The level of the flicker noise is inversely proportional to the frequency, and it is therefore known as $1/f$ noise. We shall give an approximate expression for the magnitude of this $1/f$ noise, which has been derived by applying a generally accepted theory of its origins.

Dr. Ir. P. A. H. Hart and Dr. F. M. Klaassen are with Philips Research Laboratories, Eindhoven.

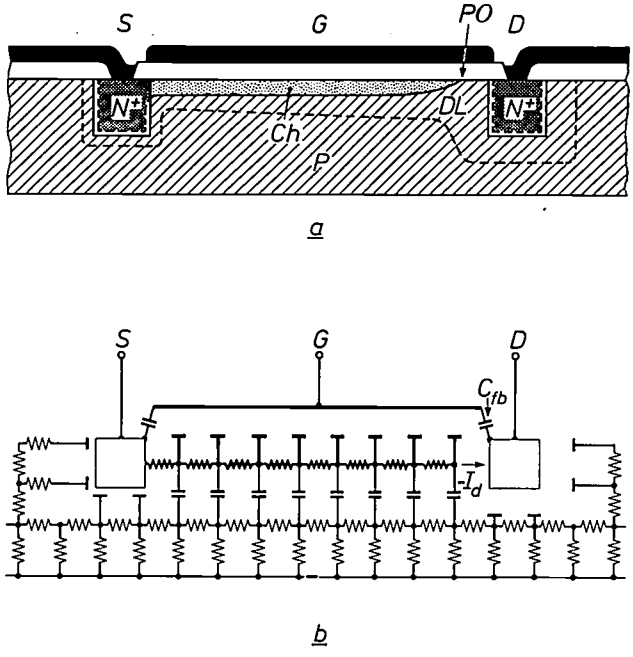


Fig. 1. *a*) An *N*-channel MOS transistor in cross-section. *P* is a *P*-type silicon substrate, *S* a N^+-type source. *G* gate. *D* drain. *Ox* oxide layer. *Ch* channel. *PO* pinch-off point. *DL* depletion layer.
b) Electrical model of a MOS transistor. The channel has a distributed capacitance to the gate *G*. This capacitance and the resistance of the channel belong fundamentally to the mechanism of the MOS transistor and are therefore drawn in heavy lines. The distributed capacitance of source, drain and channel to the semiconducting substrate is a stray element, like the capacitances between the source and drain and the gate *G*. The arrow $-I_d$ indicates the direction in which the electrons move.

First of all we shall take a closer look at the magnitude and frequency-dependence of the linear amplification of the MOS transistor. Here we shall make use of the equivalent circuit and the equivalent linear four-terminal network. The noise will be dealt with in the last part of this article, beginning on page 222.

Linear amplification, equivalent circuit

The description of the characteristics of the MOS transistor as a linear amplifier is subdivided into two parts: the development and description of the equivalent circuit, and a discussion of the gain and stability in terms of the general theory of linear four-terminal networks. The construction of the MOS transistor is

illustrated schematically in fig. 1*a*. We assume that the MOS transistor is made of *P*-type silicon, so that electrons are responsible for the charge transport in the conducting channel produced by inversion at the surface. (In what follows, however, it makes no essential difference whether we start from *P*-type or *N*-type material; all that need be done is to substitute *N* for *P*, and to read "hole" for "electron".) In fig. 1*a*, *Ch* stands for the channel that extends from the source *S* to the pinch-off point *PO*. (If the MOS transistor is not biased for operation in saturation, the channel extends to the drain *D*.) Fig. 1*b* shows an electrical network which is directly derived from the structure of the MOS transistor. Here the channel is represented by a resistance between *S* and *PO*, which is coupled to the gate by a distributed capacitance, thus forming an *RC* ladder network. The arrow marked $-I_d$ indicates the direction of the electrons coming from the channel and moving through the pinched-off part of the channel under the influence of the electrical field between the drain *D* and the point *PO*. The whole system of source, channel and drain is surrounded by a depletion layer (*DL* in fig. 1*a*) and is therefore isolated from the substrate, apart from a small leakage current. It is however capacitively coupled to the substrate. In fig. 1*b* this capacitive coupling is represented by a distributed capacitance between the system and a resistance network (shown by thin lines) which represents the substrate. Finally, since there is overlapping of the electrodes, there will be a fringing capacitance between source and gate, and also between gate and drain. The capacitance between source and gate is often negligible compared with that between channel and gate; the capacitance C_{fb} between drain and gate is of considerable importance, since it gives signal feedback from drain to gate.

It has been found that for a fairly wide frequency range the network with distributed elements in fig. 1*b* can be simplified to a network with lumped elements, as shown in *fig. 2*. The resistance R_1 and the capacitance C_1 represent the impedance between the gate *G* and the source *S*. The current source *i* — by definition a current source supplies a current whose magnitude is independent of the load — represents the a.c. component of I_d. The sign of *i* is such that the transconductance g_m is positive in the expression $i = g_m v_{gs}$, which gives the relation to the a.c. voltage v_{gs} on the gate. The capacitance C_{fb} represents the feedback capacitance referred to above. To describe the feedback from the drain via the substrate to the channel a further impedance should really be shown in parallel with C_{fb}, consisting in its simplest form of a resistance and capacitance in series. This has not been done because the effect of C_{fb} is usually predominant. The resistance R_2 and the capacitance C_2 represent the

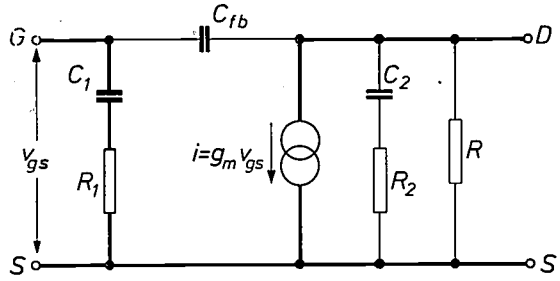


Fig. 2. Equivalent circuit of the MOS transistor. As in fig. 1*b*, the principal elements of the MOS transistor are shown in heavy lines. The network of distributed capacitances and resistances that indicate the gate and channel in fig. 1*b* has been simplified here to a series arrangement of C_1 and R_1, and the stray effect of the substrate is represented by C_2 and R_2.

substrate impedance between source and drain. Finally, fig. 2 includes the resistance *R*; this bears no direct relation to fig. 1*a* and *b*, but indicates that the saturation of the MOS transistor is not complete. In the saturated state I_d still increases slightly with the drain voltage V_d. This can be seen in fig. 2*a* of the preceding article [1]. It is also explained there (page 214) that when V_d is increased the channel becomes shorter since the distribution of the lines of force changes, and this causes the current to increase because the transconductance is inversely proportional to the length of the channel. At low frequencies the effect of *R* is greater than that of R_2 and C_2, and at high frequencies it is usually the other way about.

When the MOS transistor is not operated in saturation, V_d has a much greater effect on I_d. This means that we must substitute a much smaller resistance for *R* in the equivalent circuit. Now the amplification factor of the MOS transistor, like that of the thermionic triode, is equal to the product of the internal resistance *R* and the transconductance g_m. Because of its lower internal resistance the unsaturated MOS transistor has a lower amplification factor, which explains why it is usual to operate a MOS transistor in saturation if it is to be used as an amplifier. The difference between an unsaturated MOS transistor and a saturated one is rather like that between a triode and a pentode in valve circuits; the pentode gives a higher voltage amplification.

Now that we have established the form of the equivalent circuit, we shall discuss the various elements in quantitative terms. The elements C_1, R_1 and *i* are essential to the operation of the MOS transistor and are shown in heavy lines in fig. 1*b* and fig. 2. All the other elements are strays, and are found to be very dependent on the method of manufacture. Since C_1 and R_1 in fact represent an *RC* ladder network, their value is no

[1] J. A. van Nielen, Operation and d.c. behaviour of MOS transistors; this issue, page 209.

[2] R. J. Nienhuis, Integrated audio amplifiers with high input impedance and low noise; this issue, page 245.

[3] R. J. Nienhuis, A MOS tetrode for the UHF band with a channel 1.5 μm long; this issue, page 259. See also T. Okumura, The MOS tetrode, Philips tech. Rev. 30, 134-141, 1969 (No. 5).

longer constant at frequencies that are so high that the distributed nature of the channel, which is only a few microns long, starts to become significant. The same applies to the transconductance g_m. Although the presence of stray capacitance usually makes it impossible to use the MOS transistor at such high frequencies, it will nevertheless be useful to look a little more closely at this frequency limitation which is inherent in the mechanism of the MOS transistor.

The essential elements

Solutions to the differential equations that describe the operation of the MOS transistor when the strays can be neglected have been obtained at Philips Research Laboratories and the numerical results have been published [4]. It would take us too far to discuss these solutions in detail here, and we shall therefore simply present the results.

To express the frequency-dependence of the transconductance g_m, we must express g_m in the following form (see equations (7) and (10) in the preceding article [1]):

$$g_m = \frac{\mu C_{ox} w}{l} V_{sat} H(\omega) = g_{m0} H(\omega), \quad . \quad (1)$$

where μ is the mobility of the charge carriers (assumed to be constant), l is the length and w the width of the channel, C_{ox} is the capacitance of unit area of the gate to the channel, and V_{sat} is the voltage between the point PO of the channel and the source S; $H(\omega)$ is a complex function of the frequency that approaches unity at low frequencies. If there is no charge in the oxide, if the oxide layer is relatively thin (say 0.1 μm) and if the impurity concentration in the substrate is sufficiently low (say $< 10^{14}/cm^3$), then $V_{sat} = V_{gs} - V_{th} = V_{gs}'$ (V_{th} is a threshold voltage); in other cases V_{sat} is less than V_{gs}' and should be calculated taking these effects into account [5]. The numerical approximations given here for H, C_1 and R_1 as a function of frequency are derived for the case where $V_{sat} = V_{gs}'$. Computer calculations show, however, that even when $V_{sat} < V_{gs}'$, the approximations are still reasonably good.

The frequency-dependent function $H(\omega)$ is shown in *fig. 3* with $\omega\tau$ as variable; τ is given by $\tau = C/g_{m0}$, where C is the total capacitance of the gate to the channel. The quantity τ can be taken as the relaxation time of the RC ladder network in fig. 1*b*, since the total resistance of the channel is equal to $1/g_{m0}$, as explained in the preceding article [1] on page 213. We can approximate $H(\omega)$ by an analytical form:

$$H(\omega) \approx \frac{e^{0.104\, j\omega\tau}}{1 + 0.164\, j\omega\tau} \approx \frac{1}{1 + 0.267\, j\omega\tau}; \quad . \quad . \quad . \quad (2)$$

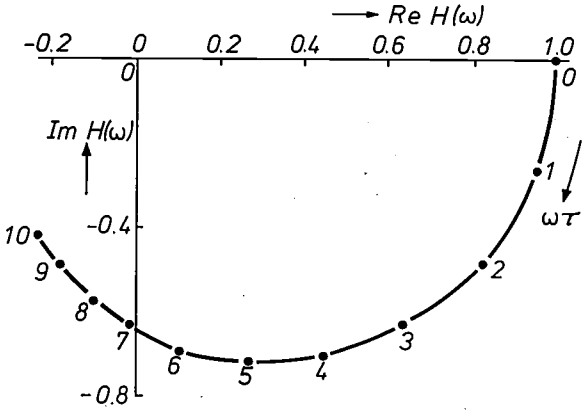


Fig. 3. The complex function $H(\omega)$ describing the frequency dependence of the transconductance of the MOS transistor. τ is the relaxation time of the RC ladder network formed by gate and channel.

the first approximation applies for $\omega\tau < 10$, the second is adequate if $\omega\tau < 1$.

Not only the transconductance but also R_1 and C_1 are dependent on frequency. The approximations [4] derived for $\omega\tau < 10$ are:

$$R_1 = 0.2 \frac{1}{g_{m0}} (1 - 10^{-3}\, \omega^2\tau^2), \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

$$C_1 = 0.67\, C (1 - 5.8 \times 10^{-3}\, \omega\tau - 1.62 \times 10^{-3}\, \omega^2\tau^2). \quad (4)$$

We see that R_1 and C_1 are virtually constant at frequencies below about $\omega\tau = 4$, and gradually decrease at higher frequencies. The reason for this is that R_1 and C_1 are lumped elements that represent the impedance of a ladder network (fig. 1*b*). If the frequency rises, the impedance of each partial capacitance decreases, while the channel resistance remains unchanged. The a.c. current flowing via the gate to the source electrode will thus take a shorter and shorter path. Since it thus "sees" a shorter and shorter portion of the RC ladder network, the values of C_1 and R_1 will have to decrease.

The behaviour of $H(\omega)$ can also be explained in a similar manner. The a.c. current i in the channel undergoes an increasing phase shift with increasing frequency, and the effect of this is enhanced because, as the frequency rises, the current in the channel becomes increasingly modulated in the neighbourhood of the source and decreasingly modulated in the neighbourhood of the pinch-off.

It has already been pointed out that this distributed nature of the channel only becomes apparent at very high frequencies (in an existing MOS transistor, which will later be taken as an example, $\omega\tau = 4$ corresponds to about 4 GHz). For practical purposes R_1 and C_1, and often g_m as well, may be treated as independent of frequency.

Stray elements

The magnitude of the stray elements, which depends to a great extent on the structure of the MOS transistor and on the method of manufacture, is often impossible to calculate with sufficient accuracy and usually has to be determined experimentally. For example, to calculate C_{fb} the formula for the parallel-plate capacitor can be applied to the overlapping parts of the gate and drain, with the oxide layer as the dielectric. If there is a relatively large overlap, this gives the correct expression for C_{fb}; if the overlap is small, however, as it is for transistors made by ion implantation [6], then this expression is incorrect since C_{fb} is then determined entirely by fringing effects. In high-frequency MOS transistors the value of C_{fb} is between 0.01 and 1 pF.

Actual values found for R_2 and C_2 are 50-10 000 Ω and 0.1-1 pF. Very high values of R_2 and low values for C_2 can be obtained in MOS transistors made in a thin silicon layer on an insulating substrate [7]; the substrate is then largely eliminated. The same applies to MOS transistors made by ion implantation in a weakly doped substrate; a large depletion zone forms in the weakly doped silicon around the drain, so that the influence of the substrate is very small [6].

The value of the internal resistance R cannot be calculated exactly from the transistor configuration either. As we saw above R is related to the displacement of the pinch-off point that occurs when the drain voltage V_d is changed, and this displacement is due to a rearrangement of the electric lines of force in the oxide layer, channel and substrate in the region of the drain. The calculation of this field variation is a two-dimensional potential problem, for which no analytical solution can be given [8] [9]. The values of R found in practice lie between 1 kΩ and 100 kΩ, depending on the configuration and on the conductivity of the substrate.

There are a number of approximate relations for R that express the relationship to the characteristics of the MOS transistor and the operating conditions. If the substrate is heavily doped, the behaviour of the pinched-off part of the channel can be approximated to that of a one-dimensional P-N junction with a reverse bias $V_{ds} - V_{sat}$ across it [9]. If, in addition, the pinched-off part of the channel is short compared with the total channel length l, then it can be shown that:

$$R \approx \frac{2l}{g_{m0} V_{sat}} \sqrt{\frac{2eN(V_{ds} - V_{sat})}{\varepsilon_{ox}}}.$$

Here e is the elementary charge, N the number of donors or acceptors per unit volume of the substrate, and ε_{ox} the dielectric constant of the oxide.

On lightly doped substrates the depletion zone around the drain is much more extensive. In this case, provided the channel length l is not too large compared with the thicknes of the depletion layer, the following approximation gives a better fit [10]:

$$R \approx \frac{1}{g_{m0}} \frac{l}{h} \frac{\varepsilon_{ox}}{\varepsilon_{Si}} \frac{V_{sat}}{V_{gs}'}.$$

In this expression h is the thickness of the oxide layer and ε_{Si} is the dielectric constant of the silicon. The influence of the geometrical ratio h/l can be seen here; high-frequency transistors with a short channel have a low internal resistance R.

The Y-parameters

The behaviour of a MOS transistor can be more generally described by treating the device as a linear active four-terminal network (or two-port; see *fig. 4*). A MOS transistor has only three terminals, but may nevertheless be treated as a four-terminal network by assuming that input and output have one terminal in common. We adopt the four-terminal network system of notation in which the currents I_1 and I_2 at the input and output are expressed in the voltages V_1 and V_2 across input and output. The parameters that characterize the four-terminal network then have the dimensions of admittance and are described, applying the

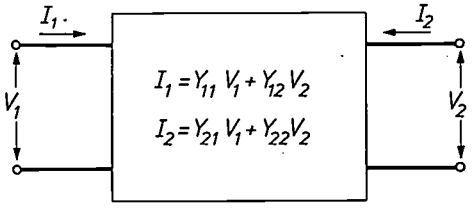


Fig. 4. An electrical four-terminal network (two-port) is completely described by four Y-parameters.

[4] J. A. Geurst, Calculation of high-frequency characteristics of thin-film transistors, Solid-State Electronics **8**, 88-90, 1965. J. A. Geurst and H. J. C. A. Nunnink, Numerical data on the high-frequency characteristics of thin-film transistors, Solid-State Electronics **8**, 769-771, 1965. The equations for the MOS transistor and for the thin-film transistor are identical provided the effect of the substrate can be neglected. A more easily handled approximate solution in the form of a series expansion has been given by D. H. Treleaven and F. N. Trofimenkoff, MOS FET equivalent circuit at pinch-off, Proc. IEEE **54**, 1223-1224, 1966. A more accurate approximation has recently been derived at Philips Research Laboratories by J. A. van Nielen, A simple and accurate approximation to the high-frequency characteristics of insulated-gate field-effect transistors, Solid-State Electronics **12**, 826-829, 1969 (No. 10).

[5] J. A. van Nielen and O. W. Memelink, The influence of the substrate upon the DC characteristics of silicon MOS transistors, Philips Res. Repts. **22**, 55-71, 1967.

[6] J. M. Shannon, Ion-implanted high-frequency MOS transistors; this issue, page 267.

[7] J. A. van Nielen, M. J. J. Theunissen and J. A. Appels, MOS transistors in thin monocrystalline silicon layers; this issue, page 271.

[8] J. A. Geurst, Theory of insulated-gate field-effect transistors near and beyond pinch-off, Solid-State Electronics **9**, 129-142, 1966.

[9] V. G. K. Reddi and C. T. Sah, Source to drain resistance beyond pinch-off in metal-oxide-semiconductor transistors (MOST), IEEE Trans. **ED-12**, 139-141, 1965. A multidimensional analysis was recently published by D. Frohman-Bentchkowsky and A. S. Grove, Conductance of MOS transistors in saturation, IEEE Trans. **ED-16**, 108-113, 1969 (No. 1). A numerical calculation has been given by H. W. Loeb, R. Andrew and W. Love in Electronics Letters **4**, 352, 1968 and by J. E. Schroeder and R. S. Muller, IGFET analysis through numerical solution of Poisson's equation, IEEE Trans. **ED-15**, 954-961, 1968.

[10] S. R. Hofstein and F. P. Heiman, The silicon insulated-gate field-effect transistor, Proc. IEEE **51**, 1190-1202, 1963.

conventional symbol, as Y-parameters. These Y-parameters have the following physical significance: Y_{11} represents the input admittance when the output of the four-terminal network is short-circuited, whereas Y_{22} is the output admittance when the input is short-circuited; Y_{21} is called the transfer admittance, and Y_{12} the feedback admittance. *Table I* gives the Y-parameters calculated from the equivalent circuit for the three possible configurations: grounded-source, grounded-gate or grounded-drain. (These might be compared with the familiar grounded-cathode, grounded-grid and grounded-anode or cathode-follower configurations in valve circuits.)

The gain

Active four-terminal networks can be divided into two classes of different gain characteristics: stable networks and potentially unstable networks. In the first case the gain of the four-terminal network is finite for every value of the admittances of signal source and load. In the second case this is not so, and the gain will be infinitely high for certain combinations of source and load admittance. In the latter case, therefore, spontaneous oscillation is possible. (The real part of both the source and the load admittance is always assumed to be either positive or zero.)

The highest gain that can be obtained with stable four-terminal networks with appropriately matched source and load admittance is called the *maximum available power gain* G_m. The source and load admittances are then exactly equal to the conjugate complex value of the input and output admittances of the four-terminal network.

J. M. Rollett [11] has given the following convenient expression for G_m:

$$G_m = \frac{|Y_{21}|}{|Y_{12}|}\left(k - \sqrt{k^2-1}\right), \quad \ldots \quad (5)$$

where

$$k = \frac{2\,\mathrm{Re}\,Y_{11}\,\mathrm{Re}\,Y_{22} - \mathrm{Re}\,(Y_{12}Y_{21})}{|Y_{12}Y_{21}|}.$$

Rollett has shown that k indicates whether the four-terminal network is potentially unstable or not: if $k \geqq 1$ then the network is stable; if $k < 1$, the network is potentially unstable and equation (5) becomes meaningless.

To avoid misunderstanding we should emphasize that a potentially unstable four-terminal network is far from useless; however, a careful choice of source and load admittances must then be made to ensure stability [12]. A two-port of this type can always be made stable by introducing damping resistances at the input or output, preferably in such a way that the k of the "new" two-port is then exactly equal to unity. As equation (5) shows, G_m then has a maximum value equal to

$$G_{ms} = \frac{|Y_{21}|}{|Y_{12}|}. \quad \ldots \ldots \quad (6)$$

The quantity G_{ms} is called the *unconditionally stable available gain*. (It is also possible to obtain an unconditionally stable gain G_{ms} from stable two-ports with $k > 1$ by adding negative resistances, such as tunnel diodes.)

Another way of achieving stabilization is the method known as "neutrodyning". In this method feedback is introduced at such a level that the "new" two-port thus formed has a feedback admittance of zero. Expressed in the Y-parameters of the original two-port, the maximum available gain in this case is:

$$G_{mn} = \frac{|Y_{21} - Y_{12}|^2}{4\,\mathrm{Re}\,(Y_{11} - Y_{12})\,\mathrm{Re}\,(Y_{22} - Y_{12})}. \quad \ldots \ldots \quad (7)$$

Neutrodyning usually requires one of the elements in the newly introduced feedback path to be adjustable (a special "neutralizing capacitance" is often used). Since this is not a practical proposition in mass-produced circuits, especially with integrated circuits, neutrodyning is not widely employed. Nevertheless, G_{mn} is frequently quoted in the literature as a kind of figure of merit for a semiconductor device, indicating

Table I. Y-parameters of a MOS transistor ($\tau_1 = R_1C_1$, $\tau_2 = R_2C_2$).

	grounded source	grounded gate	grounded drain
Y_{11}	$j\omega C_{rb} + \frac{j\omega C_1}{1 + j\omega\tau_1}$	$g_m + \frac{j\omega C_1}{1 + j\omega\tau_1} + \frac{j\omega C_2}{1 + j\omega\tau_2} + \frac{1}{R}$	$j\omega C_{rb} + \frac{j\omega C_1}{1 + j\omega\tau_1}$
Y_{22}	$\frac{1}{R} + j\omega C_{rb} + \frac{j\omega C_2}{1 + j\omega\tau_2}$	$\frac{1}{R} + j\omega C_{rb} + \frac{j\omega C_2}{1 + j\omega\tau_2}$	$\frac{j\omega C_1}{1 + j\omega\tau_1} + \frac{j\omega C_2}{1 + j\omega\tau_2} + \frac{1}{R}$
Y_{12}	$-j\omega C_{rb}$	$-\frac{1}{R} - \frac{j\omega C_2}{1 + j\omega\tau_2}$	$-g_m - \frac{j\omega C_1}{1 + j\omega\tau_1}$
Y_{21}	$g_m - j\omega C_{rb}$	$-g_m - \frac{1}{R} - \frac{j\omega C_2}{1 + j\omega\tau_2}$	$-\frac{j\omega C_1}{1 + j\omega\tau_1}$

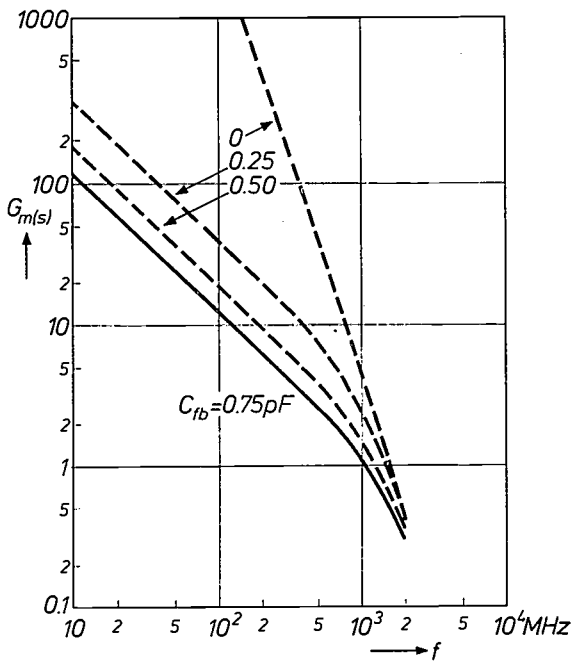


Fig. 5. Calculation of the power gain of an actual high-frequency MOS transistor. Below about 1000 MHz the transistor is potentially unstable ($k < 1$) and a plot is given of the unconditionally stable available gain G_{ms}; above about 1000 MHz the maximum available power gain G_m is given. The characteristics and operating conditions of the transistor are: $l = 6$ μm, $w = 740$ μm, the thickness of the oxide layer $h = 0.11$ μm, $I_d = 3$ mA, $g_{m0} = 5.8$ mA/V, $C_2 = 0.55$ pF, $R_2 = 140$ Ω, $R = 6.6$ kΩ. The feedback capacitance C_{fb} is 0.75 pF. To illustrate the effect of C_{fb}, calculations have also been made for the theoretical values 0.5 pF, 0.25 pF and 0 pF.

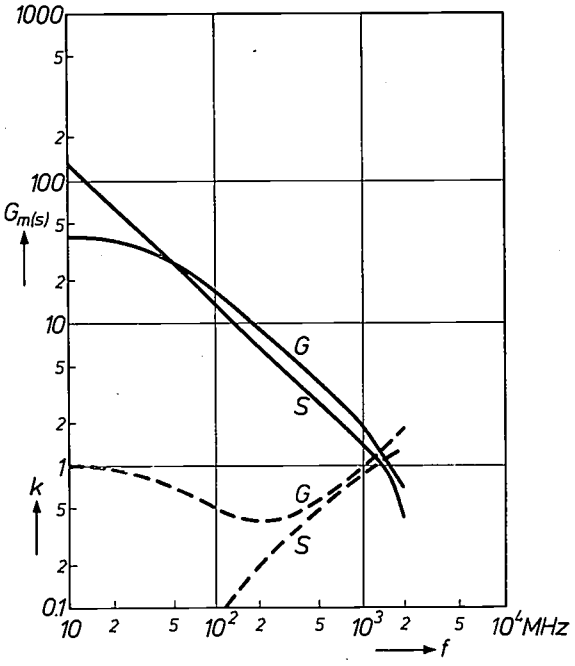


Fig. 6. The power gain (solid curves) and k (dashed curves), calculated for the same MOS transistor to which fig. 5 relates, but now with $C_{fb} = 0.75$ pF and for two configurations: with grounded source (S) and with grounded gate (G). Where $k < 1$, the circuit has to be stabilized, and here the unconditionally stable available gain G_{ms} is plotted; for other cases the maximum available gain G_m applies.

the gain that can theoretically be attained with the device.

In the MOS tetrode [3], and in the MOS transistor made by ion implantation [6], C_{fb} is extremely small, and consequently so is Y_{12}. If we neglect Y_{12}, the expression for G_{mn} becomes very simple and equal to that for G_m:

$$G_{mn} \approx G_m \approx \frac{|Y_{21}|^2}{4 \operatorname{Re} Y_{11} \operatorname{Re} Y_{22}}. \quad . \; . \quad (8)$$

To illustrate this, *fig. 5* gives calculated values of G_m for an actual MOS transistor [13], designed for high-frequency operation in a grounded-source configuration. To demonstrate the effect of C_{fb}, the calculations were carried out for four values of C_{fb}: the actual value 0.75 pF and the purely theoretical values 0.5 pF, 0.25 pF and 0 pF. In the region where $k < 1$ it is assumed that stabilization is achieved by means of damping resistances, so that $G_m = G_{ms}$. The transition from the region where $k < 1$ (low frequencies) to the region where $k > 1$ (high frequencies) can be seen as a bend in the curves. G_{ms} is inversely proportional to the frequency (see (6) and Table I); this explains the slope of the three curves for $C_{fb} \neq 0$ on the left of the bend in fig. 5. If C_{fb} is zero, G_m is given by equation (8) and the parameter k is meaningless. There should therefore be no bend in the curve for this value. For low frequencies, the expression (8) is $\propto \omega^{-2}$, for high frequencies it is $\propto \omega^{-4}$. For the transistor in question the transition is about 300 MHz, and the part of the curve shown in the figure for $C_{fb} = 0$ is therefore shown slightly bent.

The magnitude of k as a function of frequency is given in *fig. 6* for $C_{fb} = 0.75$ pF, and for two configurations, one with grounded source (S) and the other with grounded gate (G). It can be seen that k deviates less from unity for the grounded-gate case, and that at lower frequencies the gain is lower than for a grounded-source configuration. At higher frequencies however the gain is higher than with the grounded-source configuration.

A circuit of particular interest can be obtained by combining the two configurations and connecting a grounded-source MOS transistor in cascade with a grounded-gate MOS transistor. This combination is known as a cascade circuit or, in integrated form, as a MOS tetrode [3].

[11] J. M. Rollett, Stability and power-gain invariants of linear twoports, IRE Trans. **CT-9**, 29-32, 1962.
[12] H. W. Bode, Network analysis and feedback amplifier design, 13th printing, Van Nostrand, Princeton, N.J., 1959.
[13] This transistor was designed by J. A. van Nielen, who provided us with the data.

Noise in MOS transistors

To describe the noise in a MOS transistor it is helpful to distinguish between the frequency range where the noise is mainly of thermal origin and the region where $1/f$ noise dominates. This happens at the lowest frequencies; above a certain frequency limit only the thermal noise is significant. This limit depends on a variety of factors, such as the operating point and configuration of the MOS transistor, and also the fabrication process.

Thermal noise

The thermal noise mainly originates from the channel, i.e. from the distributed resistance shown in heavy lines in fig. 1*b*. Since this resistance does not have a constant value — it depends on the applied voltages — and since moreover it is capacitively coupled with the gate and the substrate, the Nyquist noise theorem cannot be applied directly to the thermal noise in the channel as a whole. To calculate the noise we divided up the channel resistance into a large number of short segments and applied Nyquist's theorem to each segment separately. We then found the sum of the noise contributions, using the equations for the MOS transistor. This is a complicated calculation, but the result is relatively simple and easy to represent in the equivalent circuit. This is done by connecting a noise-current source i_n parallel with the current source i in the equivalent circuit (fig. 2); the mean square value of the noise current is then given by [14]:

$$\langle i_n^2 \rangle = 4\,kT\,\alpha\,g_m\,\Delta f, \quad \ldots \quad (9)$$

where

$$\alpha = \tfrac{1}{2}\,V_{gs}'/V_{sat} + \tfrac{1}{6}.$$

In this expression k is Boltzmann's constant, T the absolute temperature in the channel, α a factor defining the effect of the substrate (without this effect, $V_{gs}' = V_{sat}$ and $\alpha = \frac{2}{3}$; as a rule $1 \leqslant \alpha \leqslant 3$) and Δf is the frequency interval within which the noise is to be calculated.

In addition to the current source i_n it is also necessary to include a voltage source e_{n1} in the equivalent circuit, in series with the resistance R_1, because the gate "sees" the channel via a capacitive coupling. The magnitude of the noise voltage e_{n1} is given by [15]:

$$\langle e_{n1}^2 \rangle = 4\,kT\,R_1\,\Delta f.$$

Since i_n and e_{n1} have the same physical origins, they will be correlated. As we shall see later, this correlation has very little effect.

Another source of noise in the MOS transistor is the thermal noise in the substrate. This can be taken into account by introducing a voltage source e_{n2} in series with R_2 in the equivalent circuit:

$$\langle e_{n2}^2 \rangle = 4\,kT\,R_2\,\Delta f.$$

In practice, however, it is found that e_{n2} plays no significant part, except at very high frequencies, where the gain is in any case low.

It appears at first sight that shot noise could arise in the transport of the charge carriers through the pinched-off part of the channel. Experiments have shown, however, that the noise has the magnitude that would be expected if it were thermal noise in the channel. If noise did arise in the transport of charge carriers through the pinched-off portion, the measured noise would be an order of magnitude greater. Moreover, shot noise is unlikely to occur, since in this part of the channel, just as at the base-collector junction in a transistor, the charge carriers have no potential barrier to overcome.

The leakage currents to the gate and the substrate are extremely low ($< 10^{-14}$ A), and their contribution to the noise is negligible.

Equivalent noise resistance and noise factor

In theory the noise sources i_n, e_{n1} and e_{n2} give a complete description of the thermal noise of the MOS transistor. To make the description more general, and to permit easier comparison with other amplifying elements, the MOS transistor is represented as a linear noise-free active four-terminal network, as in the Y-parameter representation, and the noise generated in the MOS transistor is accounted for by introducing equivalent external noise sources. It can be shown that it is in general necessary and sufficient to introduce *two* sources [16], for example a voltage source E in series with the input and a current source J in parallel with it. Other configurations are possible, e.g. a current source shunted across the input with a current source shunted across the output. These sources will usually be correlated with each other.

Neglecting the noise source e_{n2}, we can write for E and J:

$$E = \frac{i_n}{g_m}, \quad \ldots \quad (10)$$

and

$$J = Y_{11}\left(e_{n1} + \frac{i_n}{g_m}\right). \quad \ldots \quad (11)$$

If the impedance of the signal source is much smaller than the input impedance $1/Y_{11}$ of the MOS transistor, then J plays no significant role as in P-N junction field-effect transistors and valves and the signal-to-noise ratio is determined entirely by E. This is the case at low frequencies — certainly in the audio region and often above it; if C_1 is say 1.5 pF, then the input impedance of a MOS transistor at 1 MHz still has an absolute value of about $10^5\ \Omega$. It is therefore common practice to represent the noise voltage E as originating

from an imaginary resistance at the input of the MOS transistor, called the equivalent noise resistance R_n, defined by

$$\langle E^2 \rangle = 4\,kT_0R_n\Delta f, \quad \ldots\ldots \quad (12)$$

where $T_0 = 290$ °K. It is necessary here to fix a value T_0 for the temperature of the MOS transistor and the resistance, because the thermal noise of the MOS transistor varies with temperature in an entirely different way from that of a resistor. From equations (9), (10) and (12) it follows that

$$R_n = \alpha/g_m.$$

The noise resistance of the MOS transistor is thus inversely proportional to the transconductance, since for any given MOS transistor α is of course a constant. This implies that the noise resistance is related to the operating current, because the transconductance g_m increases as the current increases (see equation (1), which shows that the transconductance is proportional to the gate voltage). The connection with the operating current setting is shown in *fig. 7*, which gives the results of measurements on a low-frequency MOS transistor: at $f > 30$ kHz, R_n decreases with increasing I_d. The frequency range below $f = 30$ kHz is the range in which flicker noise predominates; we shall return to this presently. The measurements show that $\alpha = 1.1$ in this transistor.

When the noise source J also becomes significant, the noise can no longer be represented by a single noise resistance. This is the case at high frequencies, and here it is preferable to use the concept of noise figure. The noise figure F of a four-terminal network is defined by the expression:

$$F = \frac{P_{n0} + P_{n1}}{P_{n0}}.$$

Here $P_{n0} + P_{n1}$ is the total noise power at the output in a narrow frequency band Δf. P_{n0} is the contribution from the thermal noise of the impedance Z_i of the signal source connected between the input terminals, assuming that the temperature of Z_i is 290 °K. P_{n1} is therefore the part of the output noise added by the four-terminal network [17]. The noise figure is a function of the signal-source impedance and has a minimum F_{min} for a particular value of this impedance. For a MOS transistor at a temperature of 290 °K the following expression has been found to be a very good approximation [15]:

$$F_{min} = 1 + 2\beta^{\frac{1}{2}}\frac{\omega}{\omega_0} + 2\beta\left(\frac{\omega}{\omega_0}\right)^2,$$

where

$$\beta = R_1/R_n \quad \text{and} \quad \omega_0 = (R_nC_1)^{-1}.$$

For $\omega < \omega_0$ the variation of F_{min} with ω is substantially linear, but it becomes effectively square-law for $\omega > \omega_0$ [18].

Measurements of the minimum noise figure F_{min} at various frequencies are presented in *fig. 8*. These measurements were made with a MOS transistor very similar to the one assumed for the calculated curves of fig. 5; the figure shows that the gain of the transistor at about 1 GHz has dropped to unity. The useful frequency range of the device therefore lies below this fre-

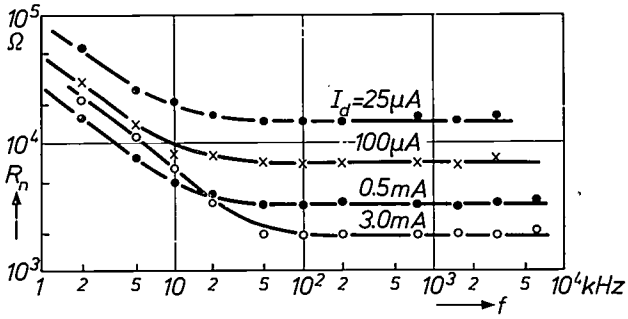


Fig. 7. The equivalent noise resistance R_n of a low-frequency MOS transistor as a function of frequency f. Below 30 kHz, $1/f$ noise dominates. The noise resistance R_n decreases on increasing I_d, i.e. with increasing transconductance of the MOS transistor.

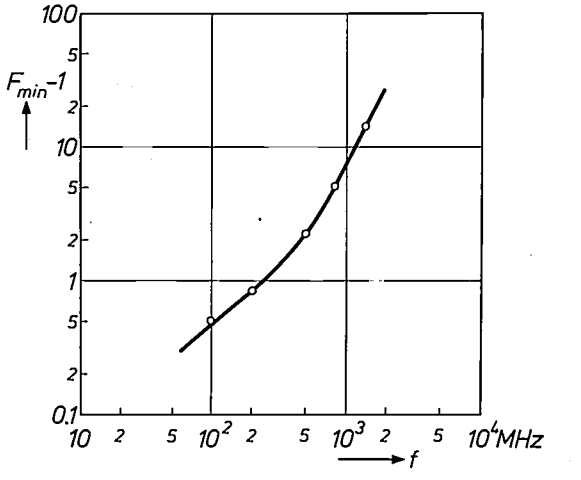


Fig. 8. Measurements and calculated frequency behaviour of the noise figure F_{min} at optimum signal-source admittance for the MOS transistor in figs. 5 and 6. The calculation is based on the data given in fig. 5 and the values $C_1 = 1.5$ pF, $\beta = R_1/R_n = 0.2$.

[14] F. M. Klaassen and J. Prins, Thermal noise of MOS transistors, Philips Res. Repts. **22**, 505-514, 1967.
[15] F. M. Klaassen and J. Prins, Noise of field-effect transistors at very high frequencies, IEEE Trans. **ED-16**, 952-957, 1969 (No. 11).
[16] A. G. Th. Becking, H. Groendijk and K. S. Knol, The noise factor of four-terminal networks, Philips Res. Repts. **10**, 349-357, 1955.
[17] P. A. H. Hart, Standard noise sources, Philips tech. Rev. **23**, 293-309, 1961/62.
[18] A numerical calculation has been given by F. M. Klaassen, A computation of the high-frequency noise quantities of a MOS-FET, Philips Res. Repts. **24**, 559-571, 1969 (No. 6).

quency, and fig. 8 shows the noise figures that will be encountered. Calculating F_{min} for this transistor using the above equation we obtain the solid curve shown in fig. 8, which fits the measured points very well.

Noise at low frequencies

No definitive and detailed theory has yet been given for the $1/f$ noise that predominates at low frequencies. It is assumed that this noise originates as a result of charge carriers from the channel entering the oxide layer through a tunnelling process and becoming temporarily trapped there [19]. The fluctuation in the number of free charge carriers in the channel appears as a fluctuation in current. Assuming that the noise does originate in the way we have just described, it can be shown that the current fluctuation ΔI_d is given by:

$$\langle \Delta I_d^2 \rangle = \frac{\gamma(N_s) e \mu}{l^2} \frac{df}{f} I_d V_d . \quad . . \quad (13)$$

Here γ is an empirical factor that varies with the number of traps N_s, and e is the elementary charge. To some extent, γ depends on the fabrication process used for the MOS transistor. It has been confirmed experimentally that $\langle \Delta I_d^2 \rangle$ is proportional to the product $I_d V_d$ [20].

From equation (13) an equivalent noise resistance can be defined, as was done for thermal noise. In this case the equivalent noise resistance is a function of frequency, and will be designated R_{nf}. Confining ourselves to the practical case of a MOS transistor operating in saturation, then

$$R_{nf} = c_n \frac{h V_{sat}}{wl} \frac{1}{f} , \quad . . . . . \quad (14)$$

where h is the thickness of the oxide layer and

$$c_n = \frac{e\gamma(N_s)}{8\, kT_0 \varepsilon_{ox}}$$

(ε_{ox} is the dielectric contant of the oxide). For different MOS transistors made by the same process, c_n has the same value, so that R_n is then only proportional to the geometrical factor h/wl [21]. (A condition is that the measurements should always be made at the same value of V_{sat}, not only because V_{sat} occurs in the numerator of equation (14), but also because $\gamma(N_s)$ varies with V_{sat}.) It has been found experimentally that as a first approximation c_n is proportional to N_s [22]. Values of c_n between 10^8 mA^{-1}s^{-1} and 5×10^8 mA^{-1}s^{-1}, depending on the fabrication process, have been found for N-channel transistors with fewer than 10^{10} traps per cm^2.

It can be seen from equation (14) that the short channel required for high-frequency transistors gives considerable flicker noise, so that MOS transistors of this kind will give a relatively large amount of flicker noise.

To indicate the limit of the low-frequency region in which flicker noise is significant, we define a transition frequency f_0, at which the flicker noise is equal to the thermal noise. There is no correlation between these two, so that the total noise may be represented as originating from a total noise resistance $R_{nt} = R_n + R_{nf}$. At f_0 the values of R_n and R_{nf} are identical, and therefore:

$$R_{nt} = R_n \left(1 + \frac{f_0}{f}\right) .$$

It can be shown that the transition frequency is given by the relation:

$$f_0 = \frac{c_n}{\alpha} \frac{\mu \varepsilon_{ox} V_{sat}^2}{l^2} .$$

This shows that the transition frequency depends on the square of the channel length. With the values $c_n = 10^8$ mA^{-1}s^{-1}, $\mu = 6 \times 10^{-2}$ m^{2}V^{-1}s^{-1}, $\varepsilon_{ox} = 3.6\varepsilon_0$, $V_{sat} = 1.5$ V, and $\alpha = 1.1$, the transition frequency is calculated to be 3 MHz for $l = 12$ μm, and 30 kHz for $l = 120$ μm. The latter case corresponds to the low-frequency transistor of fig. 7 (in particular the curve for $I_d = 3$ mA).

Summary. When the MOS transistor is operated on the slope of its parabolic characteristic, it can be used as a linear amplifier for small signals. A lumped-element equivalent circuit can be derived from the structure of the MOS transistor. Since the channel is in reality a distributed circuit the values of the input elements and of the transconductance change at high frequencies; numerical expressions are given which show the values as a function of frequency. Up to the frequencies at which stray capacitances limit the use of the MOS transistor, this frequency dependence is often of no practical significance. A general four-terminal-network representation of the MOS transistor is given, from which the maximum available gain and a stability criterion are calculated. Above a transition frequency f_0 the noise of the MOS transistor is mainly thermal noise, whose magnitude is expressed by an equivalent noise resistance or, at high frequencies, by a noise figure. Below this transition frequency, $1/f$ noise predominates; this can also be represented by an equivalent noise resistance. Expressions for the two noise resistances, the noise figure, and f_0 show the relationship between the geometry of the MOS transistor and its characteristics.

[19] A. L. McWhorter, $1/f$ noise and germanium surface properties, in: Semiconductor surface physics, editor R. H. Kingston, University of Pennsylvania Press 1957, pp. 207-228. N. St. J. Murphy, F. Berz and I. Flinn, Carrier mobility in MOS transistors; this issue, page 237.

[20] I. Flinn, G. Bew and F. Berz, Low frequency noise in MOS field effect transistors, Solid-State Electronics **10**, 833-845, 1967.

[21] For an experimental confirmation, see F. M. Klaassen, On the geometrical dependence of $1/f$ noise in MOS transistors, Philips Res. Repts. **25**, 171-174, 1970 (No. 3).

[22] G. Abowitz, E. Arnold and E. A. Leventhal, Surface states and $1/f$ noise in MOS transistors, IEEE Trans. **ED-14**, 775-777, 1967.

Some problems of MOS technology

J. A. Appels, H. Kalter and E. Kooi

Introduction

Scientists and engineers working in MOS transistor technology are charged with the production of MOS transistors and integrated circuits that possess certain specified characteristics, are stable in behaviour, and give high production yields. The specified requirements determine the various steps in the production process: from the design geometry to the choice and techniques of oxidation, etching, diffusion and other processes in the manufacture of a MOS transistor [1]. Some of the problems which this involves are described in this article; the structure and operation of the MOS transistor, which are dealt with elsewhere in this issue [2], are assumed to be generally familiar to the reader.

A typical example of a quantity that is determined by design geometry and technological processes is the transconductance of the MOS transistor. In the article just noted [2] it is shown that the transconductance — and hence the current that the transistor can carry at the maximum permissible gate voltage — is proportional to

$$\beta = \mu C_{ox} w/l. \qquad (1)$$

Here μ is the mobility of the charge carriers in the channel, C_{ox} the capacitance of the gate per unit area, w is the width and l the length of the channel (*fig. 1*).

The mobility μ depends on the semiconductor material of which the MOS transistor is made. For practical reasons this is almost invariably silicon. One of these reasons is that it is relatively simple to apply effective isolating layers to silicon by oxidation. Although impurity centres or defects may be present at the Si/SiO_2 interface, the nature and concentrations of these impurities can now be controlled, and they can in fact be used to alter the behaviour of a MOS transistor in a desired direction. Much of this article will be concerned with the Si/SiO_2 interface.

In the bulk of the silicon the mobility μ may be regarded as a constant of the material. At the surface the mobility is usually appreciably lower than in the bulk. Not only may it be affected here by the impurities or defects, but it is also found that μ decreases with increasing gate voltage, and therefore depends on the magnitude of the charge induced in the channel. A theoretical analysis based on detailed physical considerations has shown that this is to be expected [3].

It has also been found that the surface mobility is dependent on the crystal orientation at the surface. For electrons the mobility is greatest for the (100) plane of silicon; the surface mobility in this plane can even approach the value of μ in the bulk. For the holes the mobility is greatest for the (111) orientation, but hole mobility is substantially less than electron mobility. To achieve the maximum carrier mobility, and hence the maximum transconductance, the best choice is an N-channel transistor on a silicon chip whose surface is oriented in the (100) plane.

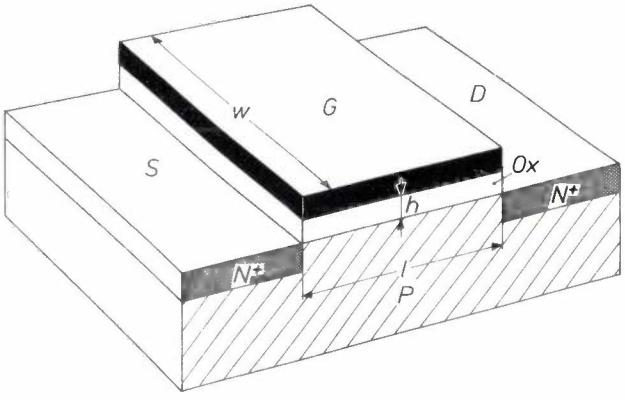


Fig. 1. Schematic diagram of a MOS transistor, made on a P-type silicon substrate. Two diffused zones of N^+ silicon constitute the source S and the drain D. Between them, isolated by an oxide layer Ox, is a metal control electrode, the gate G. If G is sufficiently positive, a concentration of free electrons occurs under the gate, forming an N-type conducting channel between source and drain. The length l and width w of the channel, and the thickness h of the oxide are the chief factors that determine the characteristics of the MOS transistor.

If a high value of β is desired it is also necessary to have a high C_{ox} (see equation 1), and for this purpose the oxide layer under the gate is made as thin as possible. The minimum thickness is mainly determined by

J. A. Appels, Drs. H. Kalter and Dr. E. Kooi are with Philips Research Laboratories, Eindhoven.

[1] A description of the photo-etching and diffusion processes is given in: A. Schmitz, Solid circuits, Philips tech. Rev. **27**, 192-199, 1966.

[2] J. A. van Nielen, Operation and d.c. behaviour of MOS transistors; this issue, page 209.

[3] N. St. J. Murphy, F. Berz and I. Flinn, Carrier mobility in MOS transistors; this issue, page 237.

the breakdown field-strength (about 10^3 V/μm); practical values frequently lie between 0.05 and 0.25 μm.

The dimensions of the silicon chips set an upper limit to the width w of the channel, and of course the chance of a defect increases with increasing w. A width of a few millimetres is fairly easy to achieve, and special techniques can be applied to give a channel with a width of a few centimetres [4].

The length l of the channel cannot be made very small without running the risk of "punch-through", i.e. a flow of current between source and drain outside the channel. The length l is usually a few microns, but special methods can be used to bring it down to about 1 micron. A very short channel is particularly important in MOS transistors for the UHF band [5].

In addition to the transconductance β, the parasitic capacitances play an important part in fast transistors. The most detrimental one is usually the feedback capacitance between drain and gate [6]. This capacitance depends on the amount of overlap between drain and gate: it can be reduced by bringing the gate into accurate register with the channel region. Various useful methods that we have developed for this will be discussed in this article.

The speed of integrated circuits made with MOS transistors is mainly limited by the parasitic capacitance between wiring and substrate. MOS transistors are therefore made with thick oxide layers under the wiring but with thin oxide layers at the active regions. This approach also tends to prevent the formation of parasitic MOS transistors; these can be formed when a voltage applied to a conductor induces a conducting channel in the substrate underneath the conductor. In the transition from the thick oxide to the thin oxide there has to be a step in the metallization; this has often proved to be a weak spot. We have therefore developed a process in which the thicker oxide is embedded deeper in the silicon substrate, so that any steps above the surface are smaller. This is known as the LOCOS process (local oxidation of silicon), and will also be described in this article.

First of all we shall take a closer look at the silicon/silicon-dioxide interface. The surface defects present there and the contact potential of the gate metal and the substrate doping all have an important effect on the threshold voltage, i.e. the minimum gate voltage needed to form a channel [2]. In fact these defects can have a much greater influence than the contact potential and substrate doping. They can change the threshold voltage by tens of volts, whereas the changes due to differences in contact potential between dissimilar metals and the variation of substrate doping that occurs in practice amount to only a few volts. The presence of mobile ions can result in a slow change in the threshold voltage. Control of the threshold voltage and making sure that it is stable are the main factors that decide which technology should be followed.

The silicon/silicon-dioxide interface

The theoretical treatment given here of the silicon/silicon-dioxide interface makes no pretence at being complete, but is a simple model that is nevertheless capable of explaining many experimental results, and one that has also been found useful for qualitatively predicting the behaviour of the Si/SiO_2 system from the processing conditions that were used when it was made. In this model we distinguish between defects of two kinds:

a) Surface states — states that can exchange charge with the silicon, and which can be described in physical terms as quantum states with an energy level between the valence and conduction band;
b) Oxide charge — fixed positive charges (ionized donors) near the interface and presumably in the oxide.

We shall now consider both types of defect in turn.

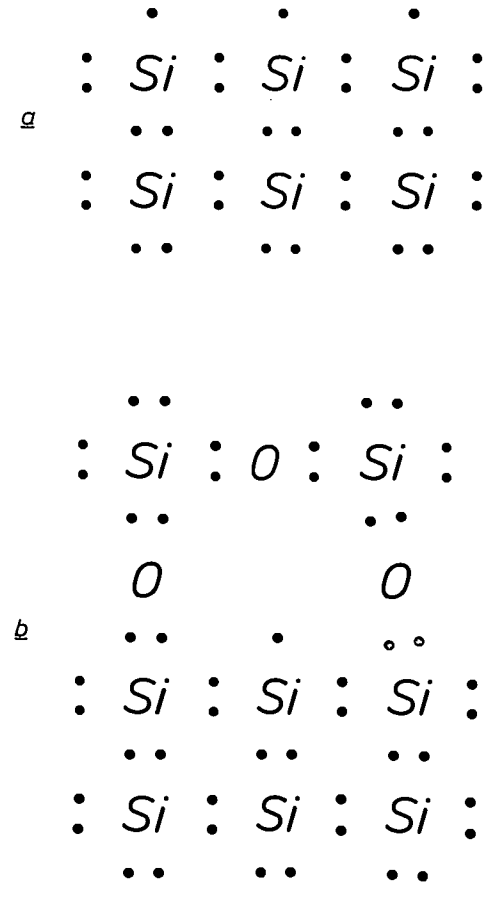


Fig. 2. *a*) Crystal lattice of silicon. At the surface of the crystal (top of the figure) each atom has an unpaired electron. *b*) Where the surface of the silicon crystal is covered with silicon dioxide, the lattices of the two substances do not exactly match. As a result silicon bonds remain unsaturated in places.

Surface states

If the crystal lattice terminates abruptly at the surface of a silicon chip, then a large number of unsaturated silicon bonds are to be expected, i.e. each atom in the outside layer of silicon atoms should have an unpaired electron (*fig. 2a*). Since there are about 10^{15} Si atoms per cm^2 at the surface, one would expect about the same number of unsaturated bonds on a "clean" surface. If the silicon is oxidized, as it is in the case under consideration, then the number of unsaturated bonds is of course lower, but it is not equal to zero because there will probably not be an exact fit between the Si and SiO_2 networks (fig. 2*b*). We shall now consider what electrical effects can result from the unsaturated silicon bonds.

It is very probable that it will take less energy to raise an unpaired electron into the conduction band than to raise a paired valence electron; in other words, the unpaired electron possesses an energy level that lies in the forbidden band. A silicon atom to which such an electron is bound may give up this electron or take up a hole, but in both cases the atom itself becomes positively charged (*fig. 3*):

$$:\underset{\cdot\cdot}{\overset{\cdot}{\mathrm{Si}}}: \rightleftarrows :\underset{\cdot\cdot}{\overset{+}{\mathrm{Si}}}: + e^-, \quad \ldots\ldots \quad (2)$$

$$e^+ + :\underset{\cdot\cdot}{\overset{\cdot}{\mathrm{Si}}}: \rightleftarrows :\underset{\cdot\cdot}{\overset{+}{\mathrm{Si}}}: \quad \ldots\ldots\ldots \quad (3)$$

If we are dealing, for example, with *P*-type silicon, then there are many holes and the equilibria (2) and (3) shift to the right. If moreover the energy gap $E_{ss} - E_v$ is small, a number of holes from the silicon may be trapped, and therefore the hole conduction near the surface of the crystal is not so good as in the bulk of the material.

It is also conceivable that the unsaturated silicon bonds may act not only as electron donors or traps for holes, but also as traps for electrons, since trapping an electron changes a silicon atom with an unpaired electron into an atom with eight electrons in its outer shell. This is the inert-gas configuration:

$$:\underset{\cdot\cdot}{\overset{\cdot}{\mathrm{Si}}}: + e^- \rightleftarrows :\underset{\cdot\cdot}{\overset{\cdot\cdot}{\mathrm{Si}}}:^- \quad \ldots\ldots \quad (4)$$

The effect may also be seen as the giving-up of a hole:

$$:\underset{\cdot\cdot}{\overset{\cdot}{\mathrm{Si}}}: \rightleftarrows :\underset{\cdot\cdot}{\overset{\cdot\cdot}{\mathrm{Si}}}:^- + e^+, \quad \ldots\ldots \quad (5)$$

and we may then conclude that the relevant energy level must lie in the forbidden band.

From a wide variety of measurements [7] it has been found that energy levels do in fact occur in the forbidden band, and that broadly two groups may be distinguished: a group near the conduction band — these are probably acceptor levels — and a group near the valence band — probably donor levels. Depending on the voltages applied in the measurements, there is a tendency for electrons or holes to concentrate at the Si/SiO_2 interface; if there is a high electron concentration the defects act mainly as acceptor levels, but at a high hole concentration mainly as donor levels. On the same sample the number of acceptor levels found in one measurement is invariably almost equal

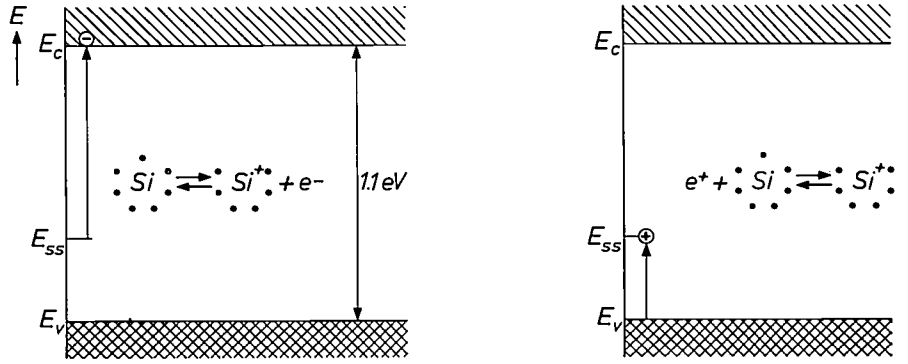


Fig. 3. The unpaired electron of a silicon atom with an unsaturated bond has an energy E_{ss} which lies in the forbidden band between valence band (energy E_v) and conduction band (energy E_c). The atom may occur as a donor; by giving up an electron (*on the left*) or taking up a hole (*on the right*) it then acquires a positive charge. If there is a high electron concentration the atom may also occur as an acceptor and acquire a negative charge.

[4] R. D. Josephy, MOS transistors for power amplification in the HF band; this issue, page 251.

[5] R. J. Nienhuis, A MOS tetrode for the UHF band with a channel 1.5 μm long; this issue, page 259.

[6] P. A. H. Hart and F. M. Klaassen, The MOS transistor as a small-signal amplifier; this issue, page 216.

[7] E. Kooi, The surface properties of oxidized silicon, Thesis, Eindhoven 1967.
M. V. Whelan, Influence of charge interactions on capacitance versus voltage curves in MOS structures, Philips Res. Repts. **20**, 562-577, 1965; Electrical behaviour of defects at a thermally oxidized silicon surface, Thesis, Eindhoven 1970.

to the number of donor levels found in another measurement; this lends plausibility to our assumption that the same trapping centres are involved in both cases.

The assumption that the centres are related to unsaturated silicon bonds explains why the number of surface states depends on the crystal orientation of the silicon surface. If this is a (111) plane, then there are usually 3 to 5 times as many surface states as on a (100) plane. This suggests that the oxide network fits better on a (100) crystal plane than on a (111) plane. Other crystal orientations give various numbers of surface states that lie between those of the (100) and (111) planes.

The way in which the surface states can affect the characteristics of a MOS transistor will be demonstrated by means of a number of experimental transistors on a *P*-type silicon substrate, i.e. with an *N*-type channel. This channel would have to be induced by applying a positive voltage to the gate. Since the effect of this is a decrease in the concentration of holes near the silicon surface and an increase in the electron concentration, the equilibria (2) and (3) shift to the left and the equilibria (4) and (5) to the right. This means that the donor states tend to become neutral (if they were not neutral already) and the acceptor states negative. The build-up of negative charge in the surface states means that the mobile charge entering the bulk of the silicon is less than the total induced charge. Consequently the threshold voltage, required for inversion, is higher than expected, and on increasing the gate voltage the subsequent increase in the inversion charge (and hence in the current through the transistor) is lower, and the transconductance is therefore affected.

The effect of the surface states is illustrated in *fig. 4*, which shows the I_d-V_{gs} curves for the experimental MOS transistors that all have the same dimensions but were annealed in different gas atmospheres after forming the gate oxide in an extremely dry atmosphere at about 1100 °C. During the anneal, the temperature was kept low (450 °C) compared with the normal growth temperature of SiO_2 on Si (1000 °C or higher), so that the processing steps could cause no difference in oxide thickness. They did, however, give rise to differences in the numbers of surface states, as may be shown from the threshold voltages and transconductances. In fact a hydrogen atmosphere and water vapour in an atmosphere of wet nitrogen even lead to negative threshold voltages and thus appear to remove the surface states for the most part. Water vapour in an oxygen atmosphere has considerably less effect. This suggests that a reduction of water to hydrogen plays an important part in the process. This hypothesis seems to be confirmed by the experience that the treatment in wet nitrogen is most effective when the chip is heated to a high temperature in an inert gas immediately after the silicon is oxidized. This treatment reduces the oxygen content of the SiO_2 through the influence of the silicon beneath it.

The simplest explanation for the disappearance of the surface states is a chemical reaction of hydrogen with the centres involved, i.e. the formation of SiH groups in our model. This explanation has been confirmed by infra-red absorption measurements [8]. With the aid of a sensitive method of measurement it has been shown that the SiO_2 almost invariably contains a certain number of SiH groups, whose concentration is particularly high when the oxidized surface is subjected to operations which reduce the number of surface states.

Often very little water vapour is sufficient to reduce the number of surface states; a heat treatment in an inert gas (e.g. nitrogen or helium) which is not extremely dry (containing a few ppm of water) may be effective. It is also found that treatment in a fairly dry environment may also be highly effective if there is a base-metal electrode (e.g. of aluminium) on the silicon surface. Here again the surface states under the electrode disappear upon heating. It is assumed that in this case a reaction of the metal with traces of water produces sufficient hydrogen.

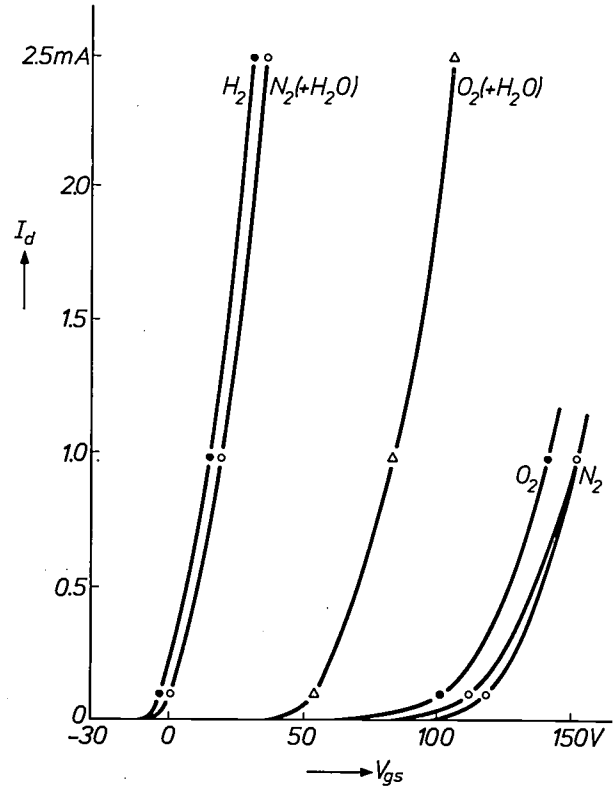


Fig. 4. The I_d-V_{gs} characteristics of a number of geometrically identical MOS transistors which have been processed in different gas atmospheres after the chips had been oxidized.

We may therefore conclude that the structure of the interface is generally very dependent on the crystal orientation of the silicon, on the method of growing the oxide and on the subsequent treatments. Many experiments can be explained on the assumption that the silicon bonds are or are not saturated with hydrogen. We can be certain, however, that this does not give a complete description of the interface. A more exact theory would have to take into account, for example, the occurrence of SiOH groups and particularly the influence of other impurities (whether deliberately introduced or not) on the interface structure. We shall return to this in the next section.

Positive charge at the oxide|silicon interface

Anyone assuming that all the difficulties are resolved by a suitable after-treatment that reduces the number of surface states to a negligible value will be surprised by the result that, although the I_d-V_{gs} curve has approximately the theoretically expected shape after such a treatment, the threshold voltage often has a value less positive (or more negative) than was expected.

Indeed, the *N*-channel MOS transistors of fig. 4 have a negative threshold voltage after treatment in hydrogen or wet nitrogen; in other words, they already have an inversion channel when the gate voltage is zero. This effect cannot be explained in terms of the surface states, since they have the very effect of opposing the inversion.

We must therefore assume that there are other centres present in addition to the ones we have mentioned. It is usually supposed that the effect is caused by the presence of positively charged centres in the oxide immediately adjacent to the silicon surface, although these are difficult to distinguish experimentally from ionized donor centres in the silicon near the surface. The amount of oxide charge, like the number of surface states described above, is connected with the interface structure. Again, with identical processing, the oxidized (100) plane is found to give the lowest oxide charge, and the (111) plane the highest. Impurities have an important effect, particularly sodium. It has been clearly demonstrated [7] that the presence of sodium during oxidation can have a marked effect on the amount of charge, although the crystal orientation still remains important. It has been shown by neutron-activation analysis that the sodium has a distribution in the oxide like that illustrated in *fig. 5*. Most of the sodium can be seen to lie in the top layer of the oxide, but there is also an accumulation at the interface with the silicon. The position of sodium in the oxide structure may perhaps best be represented as in *fig. 6a*. This structure may be regarded as a somewhat reduced oxide structure, which is also to be expected on the silicon side of the oxide layer. The sodium atom breaks the bond between an oxygen and a silicon atom, and itself forms a bond with the oxygen atom. As a result, one of the valence electrons of the silicon loses its bond, and as this electron is easily released, a positively charged centre is formed.

The sodium at the interface may conceivably be replaced by other alkali metals and even by hydrogen. This may perhaps explain why, even under fairly clean conditions, oxidation in steam gives rise to more oxide charge than oxidation in dry oxygen. On the other hand, it has also been observed that heating in hydro-

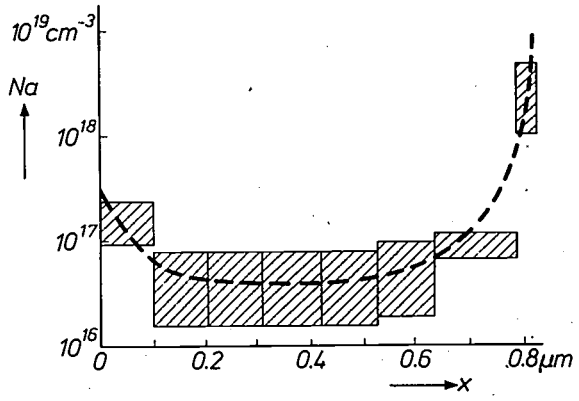


Fig. 5. Distribution of the concentration of Na atoms in the oxide as a function of the distance x from the silicon; the hatched area indicates the scatter of the measuring results.

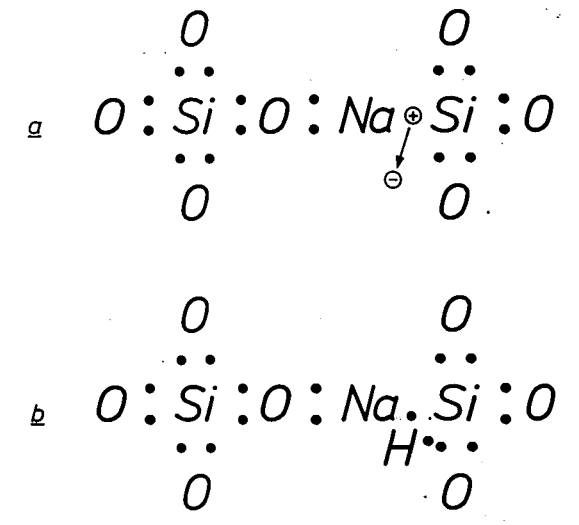


Fig. 6. *a*) The location of a sodium atom in SiO_2. The sodium atom breaks the bond between a silicon and an oxygen atom, and as a result one of the valence electrons of the silicon loses its bond; this electron is easily released and leaves behind a positive charge. *b*) A hydrogen atom can introduce an SiH group in SiO_2. In this group the hydrogen atom forms a homopolar bond with the silicon and there is no longer an unpaired electron.

[8] These measurements were carried out by Dr. K. H. Beckmann and T. Tempelmann of the Philips Hamburg laboratories: see K. H. Beckmann and N. J. Harrick, J. Electrochem. Soc. **118**, 614-619, 1971 (No. 4).

gen or in water vapour — particularly at less elevated temperature (600 °C) — may cause the charge to decrease. Here again, the formation of SiH may be expected, resulting for example in the structure illustrated in fig. 6*b*. The hydrogen atom now forms a homopolar bond with the silicon atom, and this no longer has an unpaired electron.

Fixed negative charge has sometimes been found. It can be caused by at least one impurity — gold, which is often present in small quantities. Gold is also readily made radioactive by neutron activation and its presence demonstrated in this way.

The presence of sodium and of other impurities may have a variety of causes. The impurities may come from the chemicals used or from the quartz glass tubes in which the oxidations are carried out. Another important source may be dust; if this settles on hot quartz tubes, sodium ions may easily enter the tube by diffusion and thus mix with the oxidizing gas. To achieve good process control it is therefore important to use pure chemicals and to protect the quartz tube from dust. Where extremely clean oxides are required, water-cooled quartz tubes may be used, and the silicon chip may then be heated by induction heating.

For the control of the oxide charge, cleanliness is not the only important consideration, and indeed it may not always be necessary; what is particularly important, as in the case of the surface states, is the gas atmosphere. An oxidizing atmosphere increases the oxide charge, especially when the temperature is relatively low (the effect is shown for example in fig. 4, where heating at 450 °C in oxygen does not in fact give a transistor of the depletion type — because there are so many surface states opposing inversion — but it does clearly alter the threshold voltage in the negative direction, by 15 volts). This effect of oxygen is not yet sufficiently understood. It is undoubtedly related to the oxidation mechanism: perhaps the oxygen at the upper surface attracts electrons which are then generated by structural change of the oxide/silicon interface. It is also conceivable that traces of impurities again play an important part: the transport of hydrogen from the interface towards the supplied oxygen is a likely possibility, which could cause the structure in fig. 6*b* to change for example to that in fig. 6*a*. In any case the significance for the technologist is that to obtain a low oxide charge he will have to end the oxidation in one way or another by tempering in a non-oxidizing gas.

Depending on the process used, the oxide charge is found to have a value ranging from less than 10^{10} to more than 10^{13} unit charges per cm^2. At an oxide thickness of say 0.2 μm, this means a change in the threshold voltage with respect to the theoretical value ranging from less than 0.1 V to more than 100 V. Process control has now advanced to a stage where variation of the oxide charge with a tolerance of 10^{10} charges per cm^2 is quite feasible. One of the results of the presence of the positive oxide charge was that it was originally very difficult to make *N*-channel MOS transitors that did not already have an inversion channel at zero gate voltage in the absence of surface states. This is the main reason why most MOS circuits have been (and still are) made with transistors of the *P*-channel type: here of course the presence of oxide charge only means that the threshold voltage is rather more negative, since *P*-channel transitors are always of the enhancement type.

Determination of oxide charge and surface states by measurement of the MOS capacitance

Information about the nature and number of the surface states can be obtained by a.c. circuit measurements that determine how the effective capacitance of the capacitor formed by gate, oxide layer and substrate depends on the applied d.c. voltage. A number of measurements have been made on MOS configurations specially designed for the purpose, with the metal contact on the oxide much greater than the gate of a transistor but small with respect to the dimensions of the silicon wafer, which was entirely covered on the other side by a metal substrate contact. These configurations might be referred to as MOS capacitors (*fig. 7*). The

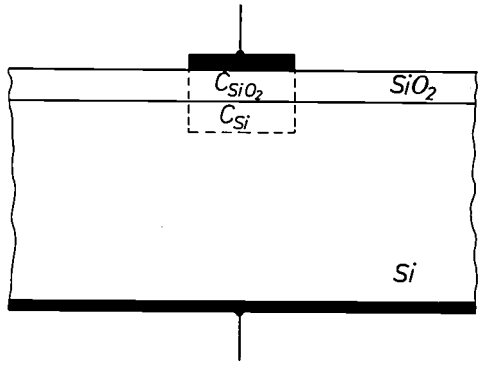


Fig. 7. MOS capacitor. The capacitance measured at the terminals is that of the series arrangement of C_{SiO_2} and C_{Si}. The magnitude of C_{Si} depends on the applied d.c. voltage.

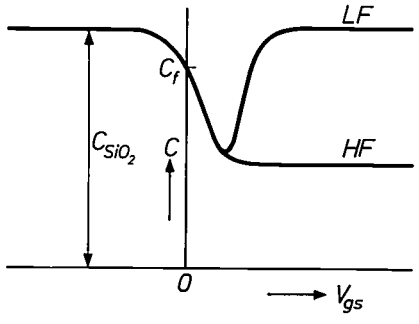


Fig. 8. Variation of the capacitance *C* of a MOS capacitor (on *P*-type silicon) with the applied d.c. voltage V_{gs} in the theoretical case where there are no surface states and no oxide charge. In this case the energy bands at $V_{gs} = 0$ are not bent and the capacitance measured is the flat-band capacitance C_f. The MOS capacitor has a capacitance equal to C_{SiO_2} when the silicon directly beneath the oxide is a good conductor. This can be demonstrated quite clearly with low-frequency a.c. voltages (curve *LF*), but at higher frequencies the agreement is not so good (*HF*).

capacitance C measured at a particular frequency may be regarded as the resultant of the series configuration of a capacitance C_{SiO_2} across the oxide layer and a capacitance C_{Si}, which is related to the fact that the charge on the lower "plate" of the capacitor has the form of a space-charge cloud in the silicon. We may write:

$$\frac{1}{C} = \frac{1}{C_{SiO_2}} + \frac{1}{C_{Si}}.$$

If $C_{SiO_2} \gg C_{Si}$, then $C \approx C_{Si}$; if $C_{Si} \gg C_{SiO_2}$, then $C \approx C_{SiO_2}$. In these equations C_{SiO_2} is a constant, but C_{Si} depends on the thickness of the depletion layer, that is to say on the applied voltage and on the doping concentration of the silicon.

If there is no oxide charge and there are no surface states, we may expect the relation between C and V_{gs} in a MOS configuration on a P-type substrate to be represented by curves like those in *fig. 8*. This figure also applies to N-type material, provided the positive and negative V_{gs} scales are interchanged.

to about C_{SiO_2}, but at higher frequencies C remains small. This is because the supply and removal of charge carriers in the inverted layer of an MOS capacitor cannot take place fast enough at such high frequencies. In MOS transistors this effect is not so pronounced, the inverted layer here being connected with source and drain diffusion.

Two typical examples of the results of capacitance measurements can be seen in *fig. 9a* and *b*. These measurements were not made on MOS capacitors but on MOS transistors specially made for the purpose, since it was also required to measure the drain current I_d. This is also included in the figures. The transistors on which the measurements represented in fig. 9*a* and fig. 9*b* were carried out had the same shape and dimensions and were made on P-type substrates having the same conductivity (50 Ωcm). There was only a slight difference in the production process: after oxidation (16 hours at 1200 °C in oxygen) and a phosphorus diffusion (4 hours at 1150 °C in dry nitrogen) the transistor of

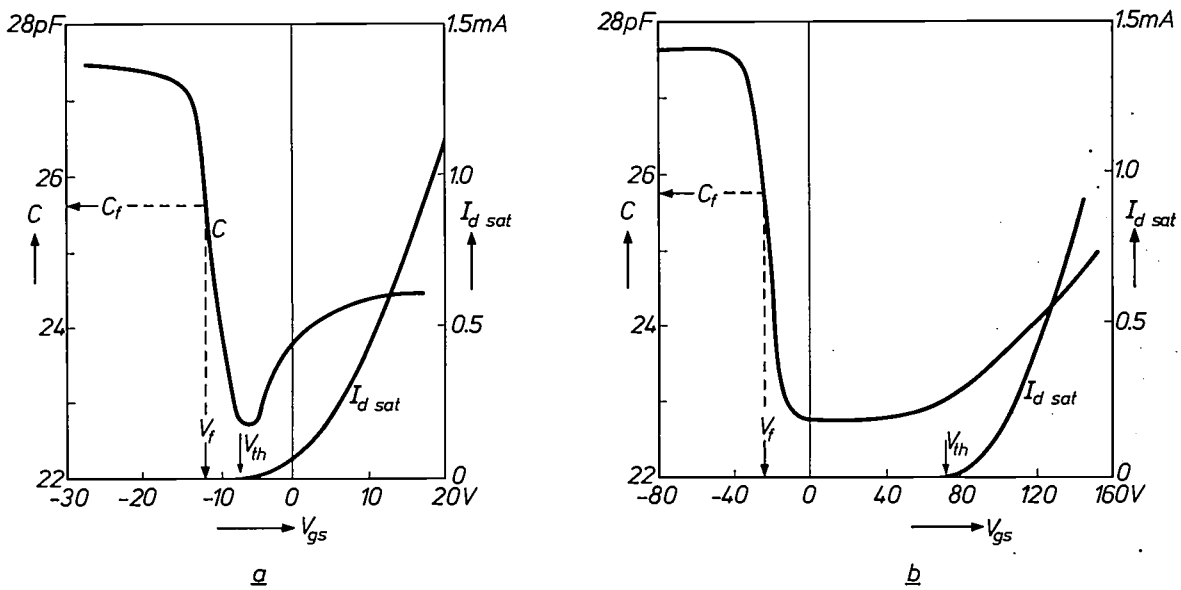


Fig. 9. Variation of the gate capacitance C with the d.c. voltage V_{gs} in an N-channel MOS transistor. *a*) After oxidation, the transistor was processed in wet nitrogen. From the calculated flat-band capacitance C_f it follows that the flat-band voltage V_f is —12 V. The threshold voltage V_{th} of this transistor is —6 V, as appears from the $I_{d\,sat}$-V_{gs} characteristic. *b*) Here the transistor has not undergone treatment in wet nitrogen, and consequently surface states are present at the SiO_2/Si interface. The flat-band voltage V_f is —30 V, the threshold voltage V_{th} is +80 V.

The nature of the curves may be explained in qualitative terms as follows. In the case of a P-type substrate a negative charge on the measuring electrode of the MOS capacitor would increase the hole concentration at the surface. Charge variations caused by the a.c. signal used for measurement then occur so close to the interface that C_{Si} is relatively large and C is approximately equal to C_{SiO_2}. If the negative voltage on the electrode is allowed to approach zero, then these charge variations gradually occur less close to the interface and C_{Si} becomes smaller. As a result the measured capacitance is also lower. If V_{gs} is raised to positive values, this process continues until the threshold voltage is reached and inversion takes place at the surface.

The values of C found when V_{gs} is raised still further depend on the frequency at which the measurement is made. At low frequencies (below about 100 Hz) C increases with rising V_{gs} up

fig. 9*a* was subjected to heat treatment at 450 °C for a further 30 minutes in wet nitrogen before the electrode metal was deposited.

The oxide layer was 1.2 μm thick in both transistors, and for the substrate conductivity of 50 Ωcm the threshold voltage V_{th} should be +6 V when all other effects are neglected. As can be seen in fig. 9*a*, the threshold voltage is —6 V, i.e. 12 V lower. If the sum of the oxide charge and the charge present in the surface states at the threshold voltage is put at N_t elementary charges e per cm², we can calculate N_t from the expression:

$$N_t e = -C_{ox} \Delta V_{th},$$

where ΔV_{th} is the change in V_{th} caused by the positive charge, i.e. —12 V in the present case. We then arrive at $N_t = 2 \times 10^{11}$ positive charges per cm².

From the shape of the curves of fig. 9*a* it can be shown that the positive charge in this transistor must be situated almost entirely in the oxide, and that the charge in the surface states is negligible in this case. Now let us return for a moment to fig. 8. This shows the variation of C with V_{gs} for the theoretical case in which there are no surface states and no oxide charges. In this case, for a gate voltage of $V_{gs} = 0$ the energy bands in the energy-band diagram of the MOS transistor are not curved [12], and C has a value C_f that can be calculated from the oxide thickness and the substrate doping. If oxide charges do exist, they cause band curvature at zero gate voltage, and a negative gate voltage V_f is then required to remove the band curvature and obtain the flat-band capacitance C_f. If we calculate this for a given transistor we can use the measured C-V_{gs} curve to find the magnitude of V_f for that transistor. The voltage V_f is a measure of the oxide charge without the charge in the surface states, because it is measured when there is as yet no question of inversion and the associated electron trapping. In fig. 9*a*, $V_f = -12$ V, which is therefore the same as ΔV_{th}. It is apparent, that ΔV_{th} is then entirely due to the oxide charge, and the surface states are negligible. The calculated number $N_t = 2 \times 10^{11}/cm^2$ therefore consists entirely of oxide charge.

In fig. 9*b* we have a different case. Here the value of V_f is about -30 V, and we see from the curve for I_d that V_{th} is about $+80$ V. If we compare the measured C-V_{gs} curve with the theoretical curve of fig. 8, we find that the fairly sharp minimum there has now become a broad region: the left-hand part of the curve has been displaced to the left (by 30 V) and the right-hand part to the right. The displacement to the left corresponds to V_f and is attributable to positive charge in the oxide, the displacement to the right must be due to the trapping of charge carriers, for the V_{gs} value at which C begins to increase coincides exactly with the value of the threshold voltage, as can be seen from the I_d curve.

Stability of the threshold voltage

It is found in practice that the threshold voltage of a MOS transistor may drift in use. This instability of the threshold voltage may be the result not only of a change in the number of surface states but also of a change in the oxide charge; in practice the change in the oxide charge is usually the cause [9].

A change can occur in the oxide layer through charge transport becoming possible in one way or another in the MOS system. There may for example be a transport of holes or electrons from the silicon to the oxide, where the carriers become trapped in defect centres. This effect is particularly noticeable at elevated temperatures (200 °C to 300 °C); it can be reduced to negligible proportions provided every effort is made to make a perfect interface structure (a (100) plane with a clean oxide and appropriate heat treatments).

A more serious effect is the transport of ions to the oxide. This occurs when the gate is positive with respect to the silicon. If no precautions are taken, this effect may be particularly marked at elevated temperature (e.g. 100 °C-300 °C). The threshold voltage may change by many volts, and always in the negative direction. This points to the build-up of a positive oxide charge near the interface. It is known that electric conduction in vitreous materials, such as SiO_2, is often due to alkali ions. Turning again to fig. 5, we see that there is a relatively large amount of sodium present in the oxide, particularly in the top layer. If all this sodium were to be driven towards the oxide/silicon interface when a positive voltage was applied, this would give rise to a concentration of about 10^{13} positive elementary charges per cm^2 (at an oxide thickness of 0.2 μm this would mean a change of 100 V in the threshold voltage). Such large changes are not generally found, however, and also the absence of any significant instability when the gate voltage is negative indicates that most of the sodium ions present in the oxide do not take part in the conduction. It is probable that reaction between the metal electrode and the oxide causes the release of ions at the oxide surface (not necessarily sodium ions alone), which then easily migrate through the oxide under the influence of the electric field.

This again demonstrates the need for clean conditions during oxide growth, but this alone is not sufficient. In the photo-etching techniques used after oxidation the oxide surface can again easily become contaminated, and contamination can also occur in the metallization stage. If all these operations are carried out with scrupulous care, MOS transistors can be made whose threshold voltage drifts by no more than a fraction of a volt when a field of 100 V/μm is applied, even at a temperature of 300 C. The chance of something going wrong is however so great that another solution is usually adopted, which is to cover the SiO_2 layer with another dielectric that is far more difficult for the ions to penetrate.

Other dielectrics

The main purpose of covering oxidized silicon with a second isolating layer is to prevent the migration of ions from the metal electrode into the oxide. A widely used procedure is to heat the oxidized silicon chip in a vapour of P_2O_5. This reacts with the SiO_2, whose top layer is converted into a vitreous mixed oxide (usual composition $SiO_2 : P_2O_5 \approx 10 : 1$). This phosphate glass prevents the migration of sodium ions very effectively, so that up to about 200 °C the instabilities can be kept to a fraction of a volt. On the other hand, it may itself introduce a slight instability, due to the occurrence of a polarization effect, which may cause some instability even at room temperature [10]. The magnitude of the effect depends on the composition and thickness of the phosphate glass in relation to the silicon dioxide beneath it. Given favourable ratios, the effect of the polarization can be restricted to a threshold voltage change of less than 0.1 V. It becomes more difficult when the oxide layers are very thin, because if

the phosphate glass is too thin (less than about 20 nm) there is a danger that the metal will react through the layer in places.

An interesting side effect of the application of phosphate glass is that at high temperature it absorbs impurities, such as sodium, from the SiO_2, which can be an advantage. A disadvantage of phosphate glass for thin layers is that the oxide charge is found to be a little higher after application of the layer, this resembles the effect found after heating in oxygen.

The best isolating layer known so far for blocking ion conduction is probably silicon nitride. This can be applied to the SiO_2 layer by reacting silane or silicon chloride with ammonia at a temperature of 800 to 1000 °C:

$$3\,SiH_4 + 4\,NH_3 \rightarrow Si_3N_4 + 12\,H_2.$$

If the reaction is made to take place in a hydrogen atmosphere, the immediate result is a radical reduction of surface states on the silicon. Traces of impurities are very important in this process also and cleanliness in processing is therefore necessary to keep the oxide charge low.

Since silicon nitride gives such effective masking, a thin layer is sufficient (10 nm to 20 nm). A disadvantage of silicon nitride is that there can be a small amount of electron conduction in it, which can give rise to slow instability effects (drift due to charge build-up at the nitride/oxide interface). The effect is more pronounced at high field strengths. To minimize the effect it is necessary to make the nitride much thinner than the oxide underneath it.

The effect can also be utilized to bring about a change in the threshold voltage: it then constitutes a storage effect [11], which can be made to last for days or weeks. The nitride should then be thick compared with the oxide. When the oxide is very thin (a few nanometres) charge can also be transferred from the silicon to the centres at the nitride/oxide interface by the tunnelling of carriers. This is also most effective at high field strengths; the momentary application of a high gate voltage (e.g. about 10^3 V/μm for 1 μs) causes a change in the threshold voltage that is maintained for a long time.

Aluminium oxide (Al_2O_3) is another good insulator which effectively blocks alkali ions. It is made at about 900 °C from $AlCl_3$ and CO_2 in a hydrogen atmosphere by the reaction:

$$CO_2 + H_2 \rightleftarrows H_2O + CO \text{ (the "water-gas reaction")},$$

$$3\,H_2O + 2\,AlCl_3 \rightarrow Al_2O_3 + 6\,HCl.$$

Since the atmosphere is again hydrogen, the result is an SiO_2 interface with very few surface states. It has been reported that the application of the Al_2O_3 layer causes a change of about 1.5 volts in the positive direction in the threshold voltage [12]. This is an interesting effect because it can be used in certain cases to counteract the unwanted effect of positive oxide charge.

To learn more about this effect we measured the flat-band voltage V_f in a MOS capacitor, by the method described on page 232. The insulation of the measuring electrode consisted of a coating of Al_2O_3 on a layer of SiO_2. The measuring electrode itself was a globule of mercury. After every measurement a thin layer of the insulating double layer was etched away and the measurement repeated. In this way we hoped to establish the distribution of the oxide charges over the thickness of the oxide. The results of the measurements are shown in *fig. 10*, in which the measured flat-band voltage V_f is plotted as a function of the oxide thickness h_{ox}. The flat sections of the curve indicate that there is no oxide charge present there (or at least less than 10^{10} charges per cm^2). The potential jump in the transition region from Al_2O_3 to SiO_2 might be explained by assuming the presence of an electric double layer of positive charges in the Al_2O_3 and

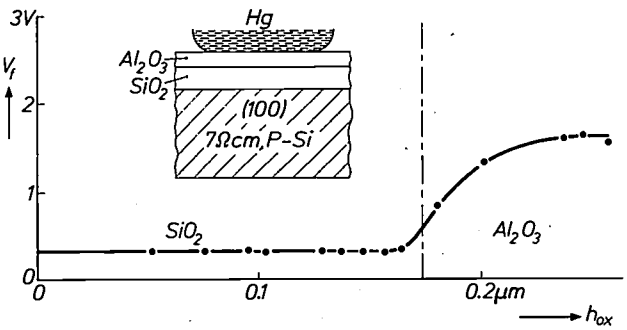


Fig. 10. The flat-band voltage V_f, measured for a MOS capacitor whose dielectric consisted of a layer of Al_2O_3 on a layer of SiO_2. During the experiment thin layers of the dielectric were etched away; after each etching operation renewed contact was made with a mercury globule and the magnitude of V_f was measured. The potential jump may be due to an electric double-layer in the junction between Al_2O_3 and SiO_2.

negative charges in the SiO_2. From measurements with other contact metals than mercury we found that the magnitude of the potential jump in fig. 10 depends on the metal used. To give a complete description of the metal/Al_2O_3/SiO_2/Si system we must therefore take into account not only the charge at the junction between the Al_2O_3 and the SiO_2 but also an exchange of charge between the metal and the isolating layer. This exchange causes a change in the effective contact potential of the metal.

Besides those in the Al_2O_3/SiO_2 system, potential jumps were also found at the interfaces of various other combinations of SiO_2, Si_3N_4, phosphate glass and Al_2O_3. The potential jump at the Al_2O_3/SiO_2 interface was found to be the best for obtaining the highest positive value of V_f.

---

[9] The sensitivity of the MOS transistor to ionizing radiation, which generates free electrons and holes in the oxide, is not dealt with here; see chapters 5 and 6 of the thesis by E. Kooi [7].

[10] E. H. Snow and B. E. Deal, Polarization phenomena and other properties of phosphosilicate glass films on silicon, J. Electrochem. Soc. **113**, 263-269, 1966.

[11] J. T. Wallmark and J. H. Scott, Switching and storage characteristics of MIS memory transistors, RCA Rev. 30, 335-365, 1969 (No. 2).

[12] H. E. Nigh, Some properties of vapour deposited aluminium oxide, Proc. Int. Conf. on the properties and use of M.I.S. structures, Grenoble 1969, pp. 77-87.

The use of Si_3N_4 or Al_2O_3 offers the additional advantage of a higher dielectric constant than that of SiO_2 ($\varepsilon_{Si_3N_4} = 6.5$; $\varepsilon_{Al_2O_3} = 9.5$; $\varepsilon_{SiO_2} = 3.8$). One can profit from this, for example, by making the isolating layer somewhat thicker for the same value of capacitance, which increases the breakdown voltage. In practice, however, the SiO_2 layer beneath the Si_3N_4 or Al_2O_3 is made relatively thick, in order to avoid the storage effects we mentioned earlier. The advantage of the dielectric constant is then minimal.

Silicon nitride and aluminium oxide are both difficult to etch with hydrofluoric acid, the etchant generally used in photo-etching techniques for SiO_2 layers, and have to be etched with hot phosphoric acid. However, this attacks most photo-lacquers. To overcome the difficulty, a layer of SiO_2 is grown (e.g. from SiH_4, CO_2 and H_2) on the Si_3N_4 or Al_2O_3, and a pattern is etched in this SiO_2 by the usual etching techniques. The pattern serves as a mask during the etching of the Si_3N_4 or Al_2O_3 layer by hot phosphoric acid.

Parasitic inversion channels; the LOCOS technique

Even when all charge transport through the isolating layer is avoided, MOS transistors can still give very troublesome instability effects. This is the case when charge is able to leak from the gate across the insulator surface. The surrounding area then acquires the same potential as the gate, and an inversion channel may also appear beside the gate. The result is that the drain current I_d gradually increases. When the current in the MOS transistor is switched off by a change in the gate voltage, the parasitic channel remains for some time and therefore I_d does not immediately go to zero.

In discrete devices this unwanted effect can easily be avoided by making the transistors in such a way that the drain region is entirely surrounded by the source region and the channel region is entirely covered by the gate. In integrated circuits, however, this is an undesirable arrangement, because the various electrodes have to be interconnected to other elements across the isolating layer. Another consequence of this may be the occurrence of inversion channels under the wiring. The way in which such parasitic transistors may occur is illustrated in *fig. 11a*.

The remedy is to ensure that at those places where parasitic channels are likely to occur the threshold voltage is higher than the potential that can appear across the oxide. It may be sufficient for this purpose to apply a thick isolating layer in the region concerned. This has the additional advantage of keeping down the capacitance of the wiring to the silicon. An unduly thick layer, however, also has serious disadvantages. Shallow windows have to be etched in the thick layer, and at the edge of these the metallization must make a fairly large step (fig. 11*b*). This has been found to be a weak spot in the metallization, which is detrimental to the yield and reliability.

We have developed a method which almost entirely overcomes this difficulty. The method is based on the local oxidation of a silicon surface masked with a layer of silicon nitride [13]. With this method, which we have called the LOCOS technique (local oxidation of silicon), structures can be made in the way shown in fig. 11. It also offers advantages for other MOS circuits.

The way in which the sunken layer of oxide is produced is shown in *fig. 12*. A silicon-nitride mask of the appropriate pattern is applied to the silicon (fig. 12*a*, *b*, *c*). Silicon is etched away from the parts not covered to a depth of about 1 μm (fig. 12*d*). Oxidation is now

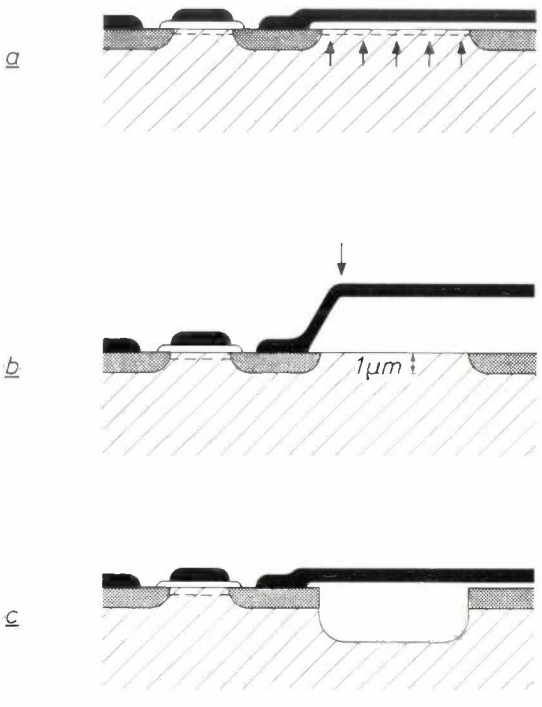


Fig. 11. *a*) In an integrated circuit with MOS transistors a parasitic inversion channel may form under the wiring (*see arrows*). *b*) To prevent this, the oxide beneath the wiring is generally made thick. The step which the metallization then has to make to the silicon surface forms a weak spot. *c*) By using the LOCOS technique with local oxidation the thick oxide (e.g. 2 μm) can be partly buried in the silicon, thus eliminating the large step in the metallization.

carried out until the recesses are completely filled with SiO_2. The oxide growth takes place at the expense of the silicon beneath it and the oxide layer formed is almost 2 μm thick (fig. 12*e*). Oxidation of the silicon nitride during this process is very slight, and the silicon nitride can be removed by etching in hot phosphoric acid (fig. 12*f*). The MOS transistors can now be made in the silicon islands formed on the surface.

Another method of avoiding parasitic inversion channels, or of at least splitting them up, is to increase the *N*-type or *P*-type doping of the substrate in the vulnerable zones. This has the effect of locally increasing the threshold voltage [2] [14]. The method is also used in combination with thick oxide in cases where the latter still offers insufficient protection.

We have already seen that the presence of surface states tends to prevent inversion (page 228). The parasitic channels might therefore also be avoided by incorporating a sufficient number of surface states at these places. This must be done locally, however, otherwise they would cause trouble in the MOS transistors. We have devised a simple method of doing this. Here again, use is made of the effective masking action of silicon nitride, although in a quite different way. The oxide is then covered with the nitride at the places where there is a risk of parasitic channels forming. A high-temperature heat treatment follows, which causes a large number of surface states to be formed because of the out-diffusion of hydrogen. In a second stage the surface states under the oxide can be removed by means of a treatment with water vapour, while those under the nitride/oxide system are maintained.

Parasitic capacitances

We have already mentioned the parasitic capacitances due to the wiring. The LOCOS technique provides a very effective means of reducing these.

In the MOS transistor a parasitic capacitance is often formed by the overlap of the gate metal across the drain region, setting a limit to the speed of the transistor. *Fig. 13* shows a number of methods that can be used for minimizing this capacitance.

For comparison, fig. 13*a* illustrates a conventional method for making an *N*-channel MOS transistor. After the N^+ diffusions the gate isolation is applied, and the metallization is then brought into register by means of photoetching techniques. An overlap of a few microns is difficult to avoid.

In fig. 13*b* the metal electrode is applied before the N^+ regions are made. These can be made either by diffusion or by ion implantation [15], using the masking effect of the metal (polycrystalline silicon can also be used in this case).

Often a thick oxide layer is used to reduce parasitic capacitances. The structure shown in fig. 13*c* was obtained by covering the whole surface with a thick layer of oxide after the N^+ diffusion. The thick oxide was then removed from the channel region and the thin gate isolation deposited in its place. The gate metal now lies partly on thick oxide, and the contribution of this part to the parasitic capacitance is therefore slight.

In fig. 13*d* the source and drain regions are diffused from a phosphorus-doped layer of silicon dioxide. During the diffusion a thin oxide film forms on the channel region [5]. Compared with fig. 13*c*, this has the

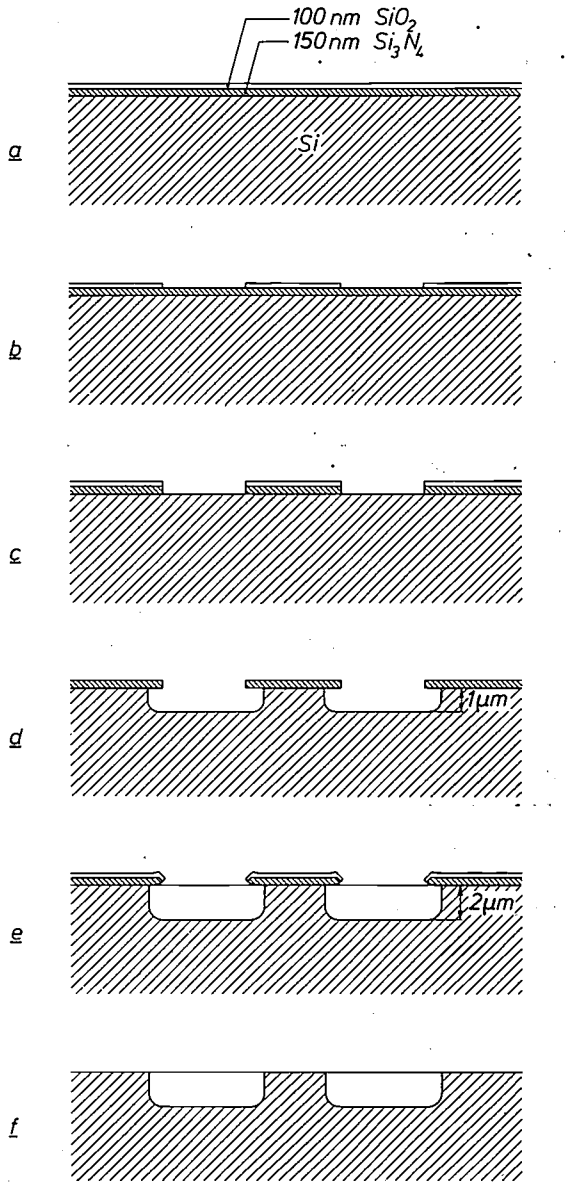


Fig. 12. The LOCOS technique.
a) A layer of silicon nitride and a layer of silicon dioxide are successively formed on the silicon.
b) A pattern of holes is etched in the SiO_2 using the conventional photo-etching technique.
c) The pattern is etched into the Si_3N_4 with hot phosphoric acid, the SiO_2 serving as a mask. SiO_2 masking is required because the photo-lacquers normally used for masking are attacked by hot phosphoric acid.
d) The silicon is etched away at the holes in the pattern to a depth of about 1 µm.
e) The silicon in the 1 µm deep holes is oxidized to a depth of about 2 µm. The holes are completely filled with the SiO_2 thus formed. The Si_3N_4 is only superficially oxidized.
f) All Si_3N_4 is etched away with hot phosphoric acid.

[13] 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 (No. 2).
J. A. Appels and M. M. Paffen, Local oxidation of silicon; new technological aspects, *ibid.*, in press.
E. Kooi *et al.*, LOCOS devices, *ibid.*, in press.
[14] L. M. van der Steen, Digital integrated circuits with MOS transistors; this issue, page 277.
[15] J. M. Shannon, Ion-implanted high-frequency MOS transistors; this issue, page 267.

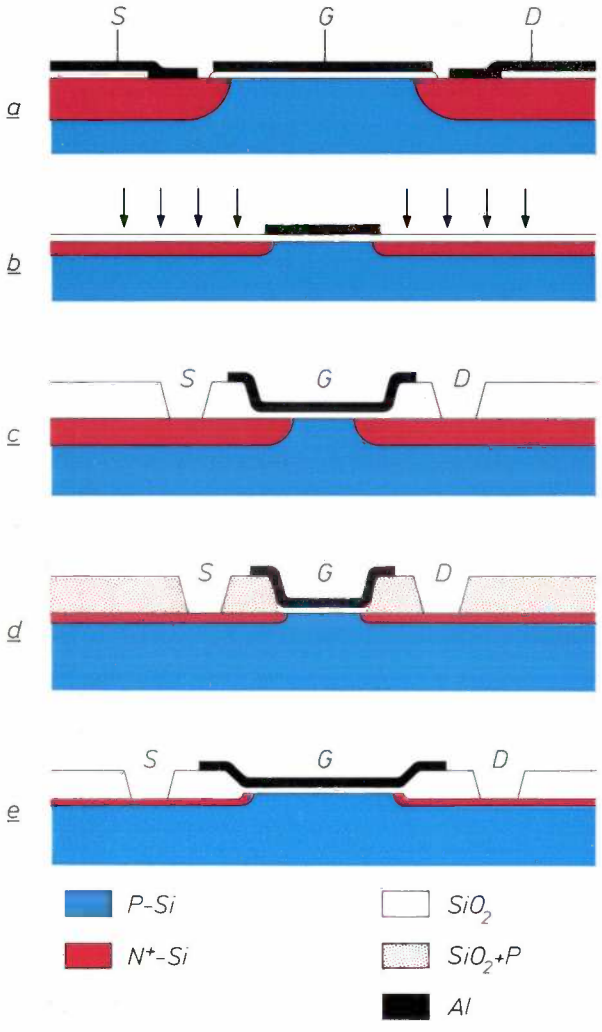


Fig. 13. *a*) Conventional MOS structure, in which the overlapping of the gate *G* causes fairly large parasitic capacitances to source *S* and drain *D*. *b*) to *e*) MOS structures with low parasitic capacitances. *b*) The overlap is small because the gate serves as a mask in the formation of the source and drain regions, formed by diffusion or ion implantation. *c*) Outside the channel region the gate lies on thick oxide. *d*) As *c*), but here the source and drain regions have been produced by diffusion from a doped oxide, so that their position in relation to the gate is more accurately determined. *e*) The thick oxide has been half buried in the silicon by the LOCOS technique. Exact control of the diffusion and oxidation processes gives diffusion regions that are only a fraction of a micron deep, thus minimizing the capacitance of the raised edges to the gate.

great advantage that it is not necessary to etch a hole at exactly the right place in thick oxide. The overlapping part of the metal electrode now lies almost entirely on thick oxide, and therefore makes very little contribution to the parasitic capacitance.

Fig. 13*e* illustrates how the LOCOS technique can be used with advantage here. After etching the channel-region pattern in the nitride layer, the diffusion of the source and drain regions is carried out, combined with oxidation, by briefly exposing the silicon chip to an atmosphere that contains the *N*-type dopant (arsenic or antimony) and then exposing it to an oxidizing atmosphere. During the oxidation process the *N*-type dopant diffuses further into the silicon. In the source and drain regions this results in a half-buried layer of oxide with a shallow N^+ diffused zone under it. Finally the nitride on the channel region is replaced by another insulator. Here again the overlapping part of the gate lies mainly on thick oxide, so that the parasitic capacitance to the drain remains small. The technique of diffusion and oxidation that we have described also helps to keep this capacitance small by limiting the diffusion depth to a fraction of a micron, which keeps the raised edges of the N^+ regions opposite the gate very narrow.

Summary. The threshold voltage of a MOS transistor depends critically on the surface states (silicon atoms with unsaturated bonds) at the Si/SiO_2 interface, and on positive charges in the SiO_2. The number of surface states is reduced by treatment in a gas atmosphere that promotes the formation of SiH groups. The positive oxide charge is mainly connected with the presence of Na^+ ions in the oxide. The oxide charge can also be reduced by the same treatment. Sodium ions in the oxide can cause positive oxide charge.

Migration of alkali ions through the oxide can cause slow drift of the threshold voltage. To avoid this, a layer of phosphor-silicate glass, silicon nitride or aluminium oxide is applied to the silicon dioxide. The electric charge at the interface, particularly that of Al_2O_3 and SiO_2, is sometimes utilized for controlling the threshold voltage.

Thick oxide is used for keeping down parasitic capacitances and for preventing the formation of parasitic channels under the wiring of integrated MOS circuits. The LOCOS technique developed by the authors makes it possible to embed the thicker oxide in the silicon, thus avoiding the steps or ridges in the metallization, which have been found to be weak spots.

Carrier mobility in MOS transistors

N. St. J. Murphy, F. Berz and I. Flinn

Introduction

The conductivity of a MOST channel is determined by the number of charge carriers in the channel and their mobility.

The channel is confined to a very thin layer (of the order of 1 to 10 nm), and the number of carriers in the channel is conveniently expressed by the *surface density* n (number of carriers per cm^2 of the layer). This quantity varies linearly with the gate voltage in the absence of trapping.

It is generally recognized that in the channel the carriers will undergo an extra scattering at the surface, so that they are characterized by a *surface mobility* μ that is smaller than the mobility as measured in the bulk. This surface mobility is usually treated as a quantity that is independent of the gate voltage V_{gs}, and also as a constant along the channel between source and drain. These assumptions are often not justified. The influence of the surface upon the mobility will depend upon V_{gs}, amongst other things, since it is affected by the channel thickness, and possibly also by the surface density of the carriers. Similarly, it will depend upon the potential difference between channel and substrate. This will vary along the channel, because of the potential variation along the channel between source and drain. It therefore seemed worthwhile to measure the carrier mobility as a function of various parameters, in particular of the gate voltage and of the bias between channel and substrate. These measurements are described in this article.

The surface mobility and the surface density are obtained by combining conductivity and Hall measurements [1] on a uniform channel, i.e. a channel in which variations of mobility and density along the channel can be neglected. This is ensured by keeping the drain voltage V_{ds} small compared to V_{gs}. A reverse bias V_{bs} may also be applied between channel and substrate. In the measurements V_{gs} and V_{bs} are the main parameters that have been varied. Other parameters are the temperature and the orientation of the crystal axes with respect to the surface.

Our most important result is the finding that the surface mobility μ varies strongly with the gate voltage V_{gs}. Generally μ increases steeply from a very small value at threshold to a maximum at gate voltages of the order of a few volts above threshold, decreasing slowly beyond this maximum. We are also led to the conclusion that trapping of carriers at the surface is insignificant at gate voltages greater than a few volts above threshold.

The results cannot be explained in terms of a simple kind of surface scattering, and they present a challenging problem for the theoretical physicist. Practically, the mobility variations have an appreciable effect on the characteristics of a MOST, and should be taken into account when for instance relating the performance of a MOST to the quality of preparation of the surface and the oxide layer. Finally it may be mentioned that data on surface mobility and trapping are important for the analysis of $1/f$ output noise in MOSTs [2].

Samples; method of measurement

The conductance of the channel is easily derived from the relation between drain current and drain voltage well below saturation. Hall-voltage measurements however need special specimens carrying Hall probes. *Fig. 1* shows an example of the Si devices that were especially prepared to our design. There are four MOSTs on each sample defined by the source and drain contacts *A-F*, *A-B* and *B-(H + H')*. Only the large square MOST *A-B* (250 μm × 250 μm) has been used for the measurements described here. This large MOST has diffused Hall contacts *D* and *E*. *G* is the gate contact and *C* is the contact to the substrate. (The MOST *A-F*, of dimensions typical for a commercial MOST, was used for measurements on noise [2]; the MOST *B-(H + H')* was designed for a different type of Hall measurement.)

N. St. J. Murphy, M.A., was formerly with Mullard Research Laboratories and is now with the Department of Electronic Engineering at Liverpool University; F. Berz, Ph.D., and I. Flinn, B.Sc., are with Mullard Research Laboratories, Redhill, Surrey, England.

[1] J. N. Zemel and R. L. Petritz, Phys. Rev. **110**, 1263, 1958; N. St. J. Murphy, Surface Sci. **2**, 86, 1964; A. B. Fowler, F. Fang and F. Hochberg, IBM J. Res. Devel. **8**, 427, 1964; D. Colman, R. T. Bate and J. P. Mize, J. appl. Phys. **39**, 1923, 1968; H. F. van Heek, Solid-State Electronics **11**, 459, 1968; F. F. Fang and A. B. Fowler, Phys. Rev. **169**, 619, 1968.

[2] I. Flinn, G. Bew and F. Berz, Solid-State Electronics **10**, 833, 1967; F. Berz and I. Flinn, Proc. Conf. on physical aspects of noise in electronic devices, Nottingham 1968, p. 135; F. Berz, Solid-State Electronics **13**, 631, 1970 (No. 5); C. T. Sah and F. H. Hielscher, Phys. Rev. Letters **17**, 956, 1966; G. Abowitz, E. Arnold and E. A. Leventhal, IEEE Trans. **ED-14**, 775, 1967.

Some of the samples were prepared at Mullard Research Laboratories, Salfords, Redhill, others were made at Hirst Research Centre, Wembley, and again others at Philips Research Laboratories, Eindhoven. They were all oxidized at 1200 °C. This was followed by a stabilizing phosphorus glass treatment and the deposition of an evaporated aluminium gate. The thickness of the oxides varied between 0.2 and 0.3 μm. The resistivity of the substrate was in the range of 5 to 20 Ωcm.

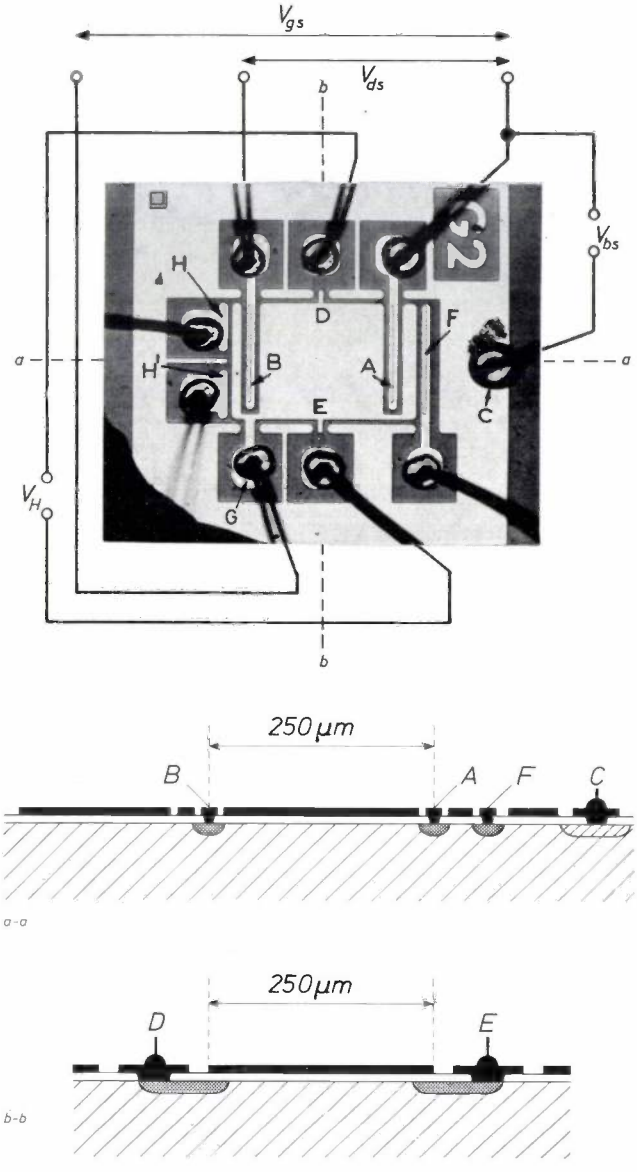


Fig. 1. MOST used for the measurements. In the cross-sections the substrate is hatched with a broad spacing, diffused contacts are shaded or closely hatched, the oxide layer is white and the metal layer is black. Only the large square MOST *A-B* was used for the present measurements. (*A-F* has been used for noise measurements and *B-(H + H')* was designed for a different type of Hall measurement.) Connections for the drain voltage V_{ds}, the gate voltage V_{gs}, the Hall voltage V_H and the bias voltage between channel and substrate V_{bs} are indicated schematically. The conductance is measured between the contacts *A* and *B*, the Hall voltage between the contacts *D* and *E*. In (*c*) the overlap of the contacts *D* and *E* with the gate extensions at *D* and *E* is indicated.

The following measurements were made simultaneously as a function of V_{gs}.

1) The source-to-drain current I_d, which gives the channel conductance g_s:

$$g_s = I_d/V_{ds}. \qquad (1)$$

2) The Hall voltage V_H between the Hall probes *D* and *E*, when a magnetic field creating a flux density *B* of about 13 kG was applied along the normal to the channel.

Since the channel is square, the channel conductance (in A/V) is equal to the surface conductivity and is therefore given in terms of the mobility μ_c (in cm²/Vs) and the surface carrier density n (in cm^{-2}) by:

$$g_s = ne\mu_c, \qquad (2)$$

where e is the electronic charge in coulombs.

The Hall voltage V_H is proportional to V_{ds} and B, and the proportionality factor is determined by the "Hall mobility" μ_H:

$$V_H = 10^{-8}\,\alpha\mu_H\,BV_{ds}, \qquad (3)$$

where α is a (dimensionless) geometric factor depending on the aspect ratio of the channel (in our case of a square channel, $\alpha = 0.68$); μ_H is in cm²/Vs and B in gauss. The Hall mobility μ_H may differ by a factor between 1 and 2 from the "conduction mobility" μ_c as introduced in (2).

In what follows we shall take $\mu_c = \mu_H$ when determining n. In terms of the measured quantities V_H and g_s we then have:

$$\mu_H = \frac{10^8}{\alpha B}\frac{V_H}{V_{ds}}, \qquad (4)$$

$$n = \frac{g_s}{e\mu_H}. \qquad (5)$$

Theoretically $\mu_H = r\mu_c$, where r is a statistical constant determined by the dependence of the average collision time of the carriers upon their energy [3]. In the bulk, $r = 1.18$ for scattering by thermal lattice vibrations (phonon scattering), and $r = 1.93$ for scattering by ionized impurities. As we shall see below (p. 241) we have found evidence that, for each of our samples, μ_H/μ_c is *substantially independent of* V_{gs}. Its value was found to be between 1.15 and 1.85, and varies from sample to sample. Therefore, taking $\mu_c = \mu_H$, our results for n will be correct for each sample within a *constant* factor between 1 and 2.

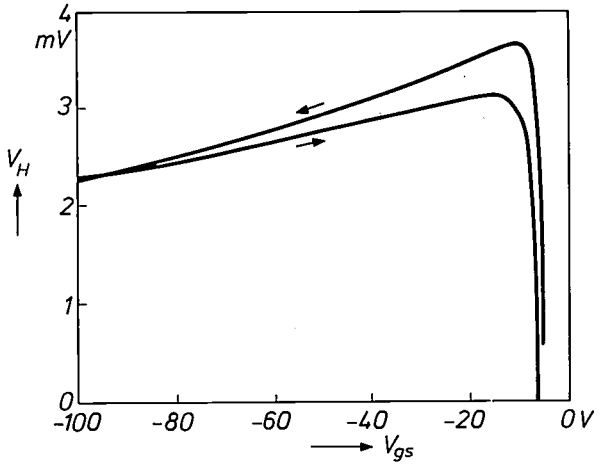
Measurements were carried out on some ten samples, both with *N* and *P* type Si channels, and with surfaces oriented in the {111} and {100} planes, at room temperature and at lower temperatures down to 150 °K. The source-to-drain voltage V_{ds} was kept small (0.2 V) with respect to V_{gs}, to ensure uniformity of the channel. In some measurements a reverse bias V_{bs} was applied

between channel and substrate. All samples showed similar qualitative characteristics, and only representative examples are quoted below.

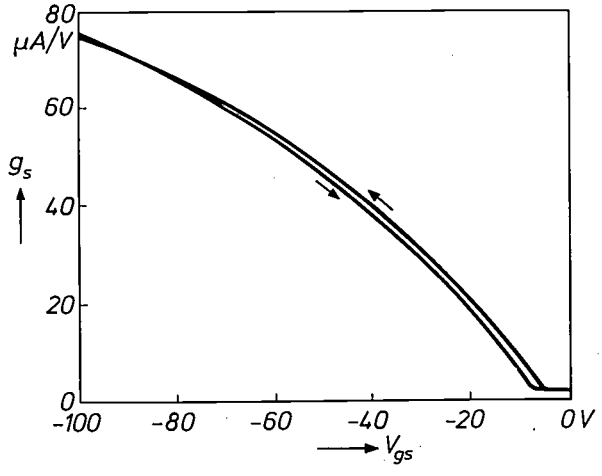
Results

The direct results of our measurements were recorder traces of V_H and g_s versus V_{gs}, as shown in *figs. 2a* and *b*. The presence of hysteresis can be seen in these graphs. (In this example it is larger than in most cases.) Equation (3) shows that the hysteresis in V_H is due to a hysteresis in μ_H. It has been checked that in fig. 2*b*, for $|V_{gs}| > 10$ V, the hysteresis of g_s is mainly due to a hysteresis in μ, and not, as is often assumed, to a hysteresis in n alone. These hysteresis effects are of great interest. They are probably linked with trapping at gate voltages close to threshold, or with ion mobility in the oxide. We have not however made any systematic study of them, and in what follows we shall only refer to curves taken for ascending values of $|V_{gs}|$.

From the direct results like those of fig. 2, the Hall



a



b

Fig. 2. *a*) Variation of the Hall voltage V_H and *b*) variation of the channel conductance g_s with gate voltage V_{gs}, for a *P*-type channel parallel to a {111} plane, at 285 °K.

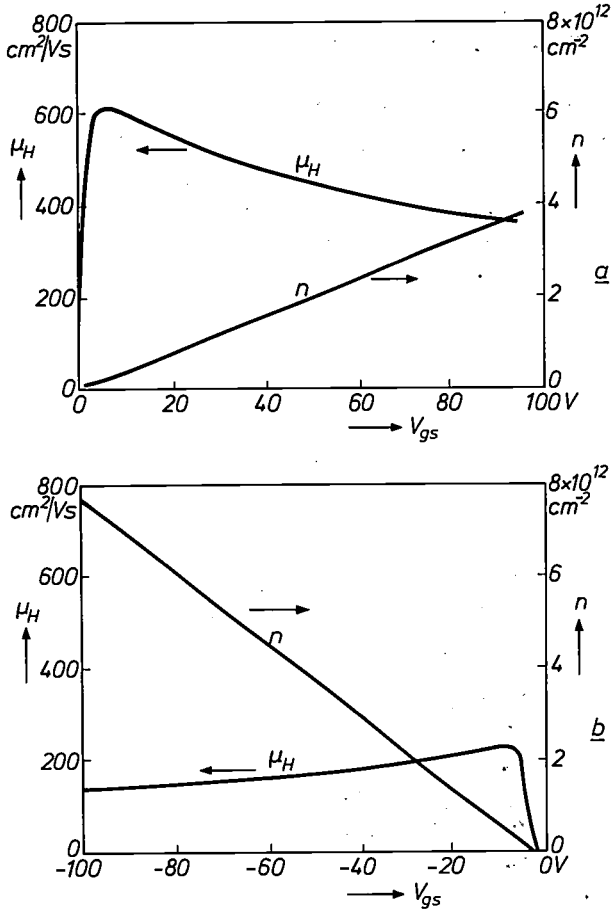


Fig. 3. Variation of Hall mobility μ_H and carrier density n with gate voltage V_{gs}, *a*) for an *N*-type channel, *b*) for a *P*-type channel, both oriented in a {111} plane and at 285 °K.

mobility μ_H and the surface carrier density n were derived, using eqs. (4) and (5). Typical results are shown in *figs. 3a* and *b*.

Variation of the Hall mobility with V_{gs} at room temperature

Most striking in our results is the strong variation of μ_H with V_{gs}. In all cases μ_H increases steeply with $|V_{gs}|$ near threshold. It reaches a maximum value μ_{Hm} for a value V_{gm} of V_{gs} of between 4 and 10 V above threshold according to sample. For $|V_{gs}| > |V_{gm}|$, μ_H decreases slowly with increasing values of $|V_{gs}|$. The value of μ_{Hm} is of the order of $\frac{1}{3}$ to $\frac{1}{2}$ of the bulk mobility. The value of n which corresponds to V_{gm} has been found to be below 10^{12} per cm² for all samples.

Values of μ_{Hm} for various crystal orientations are given in *Table I*. There is a large variation in the experimental values of μ_{Hm}, and the effect of crystal orientation is not clear. It has been found by other workers [4]

[3] See for example R. A. Smith, Semiconductors, Cambridge University Press 1961, pp. 117-123.

[4] E. Arnold and G. Abowitz, Appl. Phys. Letters 9, 344, 1966; see also the article by Colman *et al.* under [1].

Table I. The effect of crystal orientation on the maximum mobility μ_{Hm} at $T = 300$ °K. Each value shown is representative for a group of samples from the same silicon slice.

Channel type	N		P	
Surface orientation	{111}	{100}	{111}	{100}
μ_{Hm} in cm²/Vs	610	740 870	180 224	166

that for N-type channels $\mu_{Hm}\{111\} < \mu_{Hm}\{100\}$, whereas for P-type channels $\mu_{Hm}\{111\} > \mu_{Hm}\{100\}$, and our data tend to confirm these findings.

The mechanism of the variation of μ_H with V_{gs} is not understood. It has often been assumed, following a much simplified mathematical model used by J. R. Schrieffer [5], that the scattering of carriers at the surface is diffuse [6]. This, however, does not agree with our observations. As $|V_{gs}|$ increases, the carriers are drawn closer to the surface; therefore diffuse surface scattering would imply a continuous decrease of μ_H with increasing $|V_{gs}|$, in disagreement with the observed initial rise of μ_H.

It may be that in the region of $|V_{gs}| < |V_{gm}|$ scattering by ionized impurities plays a significant part. For increasing $|V_{gs}|$ the free-carrier density in the channel increases [7], resulting in a more effective screening of the charged impurities, which for impurity scattering would correspond to an increase in μ_H. In the region where $|V_{gs}| > |V_{gm}|$ the decrease of mobility with increasing $|V_{gs}|$ indicates some diffuse scattering at the surface. However, this decrease is much slower than predicted by Schrieffer [8] (see *fig. 4*). The scattering may be due to surface phonons associated with "Rayleigh surface waves" [9]. The scattering of carriers and their mobility in the channel must also be affected by the fact that for large values of $|V_{gs}|$ the carrier motion perpendicular to the surface will be hindered by quantization effects. At large gate voltages the carriers will constitute a two-dimensional gas in a plane parallel to the surface [10].

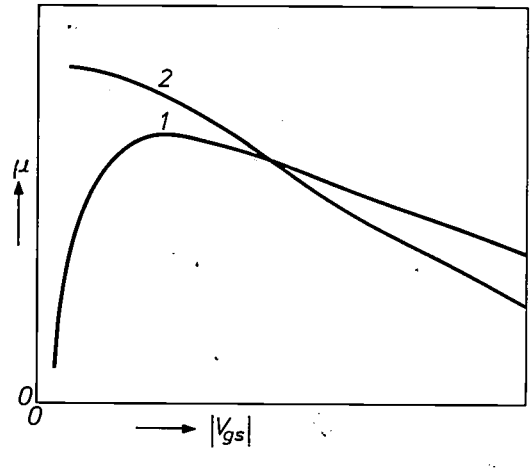


Fig. 4. The mobility as a function of V_{gs}. Curve *1* experimental. Curve *2* based on the Schrieffer theory of diffuse surface scattering (schematically).

This quantization effect may be understood by considering the simple case of a particle of mass m in a one-dimensional square potential-energy well of width a. In such a well the magnitude of the wave vector k of the particle wave function is restricted by boundary conditions to the values $\pi n/a$ ($n = 1, 2, \ldots$), corresponding to the energy values $E_n = \hbar^2k^2/2m = (\hbar^2/2m)(\pi n/a)^2$. The difference between successive energy levels increases when a is decreased.

In a MOST channel the carriers are attracted to the surface by the electric field produced by the gate voltage V_{gs}. This means that in the direction (z) normal to the surface they are contained within a potential well (though not a square one) which decreases in width when V_{gs} increases. As above, this leads to quantization effects in the z-direction [11]. More precisely, the carrier energy takes the form $E = E_{zn} + \frac{1}{2}m^*(v_x^2 + v_y^2)$, where m^* is the effective mass, v_x and v_y are the velocity components in the directions parallel to the surface, and E_{zn} is quantized ($n = 1, 2, \ldots$) (see *fig. 5*). Each value of n defines a two-dimensional sub-band, in which the energy minimum is E_{zn}. The difference $E_{zn} - E_{z(n-1)}$ increases with increasing $|V_{gs}|$. For large values of $|V_{gs}|$, $E_{z2} - E_{z1}$ will be larger than kT. At 285 °K this occurs for carrier densities of the order of 3 to 5×10^{11} per cm² when the carrier effective mass is half of the electron free mass. At these densities the carrier gas is non-degenerate; the majority of particles will occupy the lowest sub-band E_{z1} and remain mostly in this sub-band as the energy exchange during scattering by impurities or lattice vibrations is of the order of kT and thus not sufficient to reach E_{z2}. This implies that the velocity can only be altered by scattering in directions parallel to the surface. This restriction must have a strong influence on the mobility of carriers. Furthermore the effective mass m^* in the plane parallel to the surface is dependent upon surface orientation. This should also have an influence on the mobility values ($\mu = e\tau/m^*$, where τ is an average of the time between collisions). This effect has been tentatively mentioned [12] as a possible cause of the dependence of μ_H upon surface orientation (Table I).

Variation of free carrier density with V_{gs} at room temperature

In contrast to the behaviour of μ_H, the behaviour of n is very regular (see fig. 3). Although n is derived from μ_H (and g_s, see eq. 5) it varies linearly with V_{gs} well into the region where μ_H varies strongly. In fact, n varies linearly with V_{gs} within 1 or 2% from large values of $|V_{gs}|$ down to 2 to 4 V below $|V_{gm}|$, corresponding in some cases to a value of μ_H that is below μ_{Hm} by as much as 20%. For still lower values of $|V_{gs}|$ deviations from linearity seem to occur, but the results are inaccurate in this region.

A linear behaviour is expected in the absence of trapping. In this case the charge of the free carriers per cm², ne, must be equal to the product of the capacitance per cm² C and the voltage V_{gs} plus a constant:

$$ne = CV_{gs} + \text{constant}. \quad . \quad . \quad . \quad . \quad (8)$$

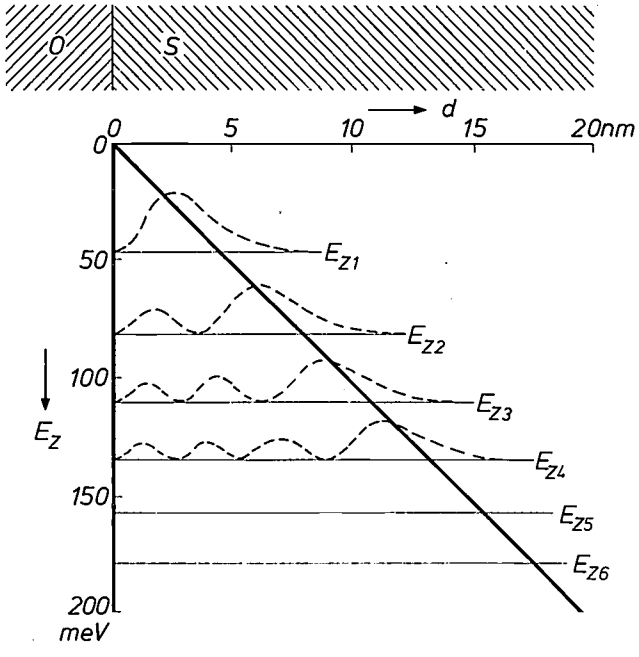


Fig. 5. Quantization of the energy of holes in a narrow *P*-type channel. The carrier effective mass is taken to be half of the electron free mass. The hole energy E_z is plotted downwards since electron energy is conventionally taken to increase upwards. d is the distance from the oxide-semiconductor interface. *O* oxide, *S* semiconductor. A potential-energy well for the holes (indicated by heavy lines) is created by the surface potential barrier (vertical line at $d = 0$) and by a field of 10^5 V/cm, assumed to be constant in the channel. The energy E_z of the holes corresponding to motion perpendicular to the channel is quantized, and the first six levels are indicated (thin lines). Ψ^2 is plotted (dashed curves) over each of the first four levels. Ψ is the normalized wave function corresponding to the level.

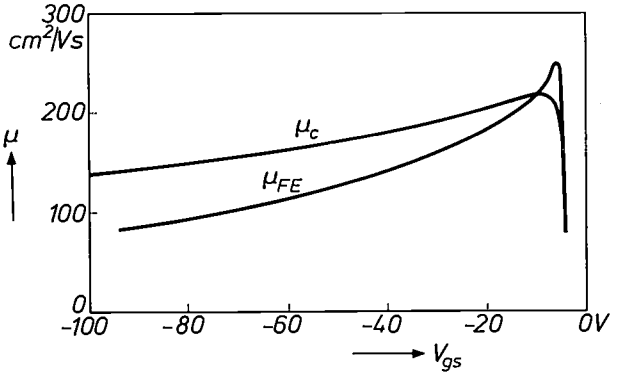


Fig. 6. Variation of μ_c and μ_{FE} with gate voltage V_{gs}, for a *P*-type channel oriented in a {111} plane at 285 °K.

[5] J. R. Schrieffer, Phys. Rev. **97**, 641, 1955.
[6] R. F. Pierret and C. T. Sah, Solid-State Electronics **11**, 279, 1968; F. Fang and S. Triebwasser, IBM J. Res. Devel. **8**, 410, 1964.
[7] E. Arnold and G. Abowitz, Appl. Phys. Letters **9**, 344, 1966; R. F. Greene and R. W. O'Donnell, Phys. Rev. **147**, 599, 1966; F. Stern and W. E. Howard, Phys. Rev. **163**, 816, 1967; see also the article by Murphy under [1].
[8] See the article by Pierret and Sah under [6].
[9] R. F. Wallis, Surface Sci. **2**, 146, 1964.
[10] See the articles by Murphy and by Colman *et al.* under [1], and by Stern and Howard under [7].
[11] See the articles by Murphy and by Colman *et al.* under [1].
[12] See the article by Colman *et al.* under [1].

Our analysis was in fact somewhat more detailed. Let us set aside for a moment our identification of μ_c with μ_H, and distinguish between the measured quantity $n_H = g_s/e\mu_H$, and the surface-carrier density proper $n_c = g_s/e\mu_c$. Then the observed linear variation of n_H with V_g for all samples, combined with the fact that, in the absence of trapping, n_c is expected to vary linearly with V_{gs}, can be considered as evidence that μ_c/μ_H is substantially independent of V_{gs}. This is the evidence that we mentioned before. Assuming this to be so, we have:

$$\frac{dn_c}{dV_{gs}} = \frac{\mu_H}{\mu_c}\frac{dn_H}{dV_{gs}}. \quad \ldots\ldots (9)$$

dn_c/dV_{gs} is given by the channel-gate capacitance C (which can be determined independently) so that μ_H/μ_c can be calculated from the observed variation of n_H with V_{gs}. In this way, as quoted earlier, values between 1.15 and 1.85 were obtained for μ_H/μ_c.

The field-effect mobility μ_{FE}

When discussing the performance of a MOST, the mobility is often defined as

$$\mu_{FE} = \frac{1}{C}\frac{dg_s}{dV_{gs}}, \quad \ldots\ldots (6)$$

which for the present we shall call the "field-effect mobility". C is the gate capacitance per cm². μ_{FE} and μ_c would be equal if μ_c were independent of V_{gs}, that is of n, and the incremental charge CdV_{gs} in the channel induced by dV_{gs} were equal to the incremental free-carrier charge edn. However, if μ_c varies with n we obtain from (6) and (2):

$$\mu_{FE} = \frac{e}{C}\left[\mu_c + n\frac{d\mu_c}{dn}\right]\frac{dn}{dV_{gs}}$$

$$= (1-\gamma)\left[\mu_c + n\frac{d\mu_c}{dn}\right], \quad \ldots (7)$$

where $\gamma = (CdV_{gs} - edn)/CdV_{gs}$ is the fraction of the induced charge that is trapped in surface states. It can be seen from equation (7) that when μ_c varies with n, it is not correct to assume, as is often done, that $\mu_{FE} = (1-\gamma)\mu_c$.

Fig. 6 shows an example of the variation of μ_c (taken equal to μ_H) and of the field-effect mobility μ_{FE}. It can be seen that, except near the maximum of μ_c, μ_{FE} is different from μ_c although there is no appreciable trapping over most of the range. The values differ because the mobility depends on V_{gs}, and therefore on n (eq. 7).

Effect of reverse bias between channel and substrate

The measurements described above were repeated with various values of reverse bias V_{bs} between channel and substrate (see fig. 1), for {111} and {100} channels. (V_{bs} is applied between source and substrate; V_{gs} is still measured between source and gate.) *Figs. 7* and *8*

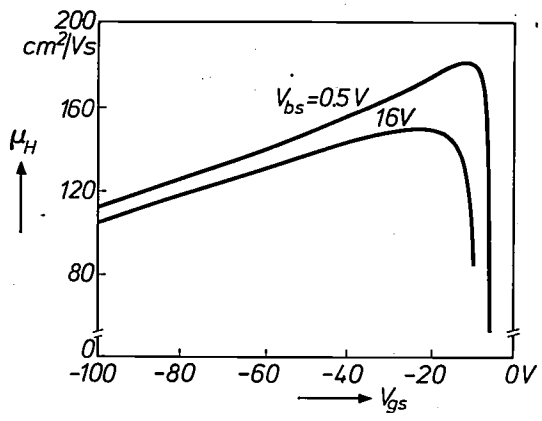


Fig. 7. Effect of a reverse bias V_{bs} on the Hall mobility as a function of V_{gs}, for a P-type channel in a {111} plane at 285 °K.

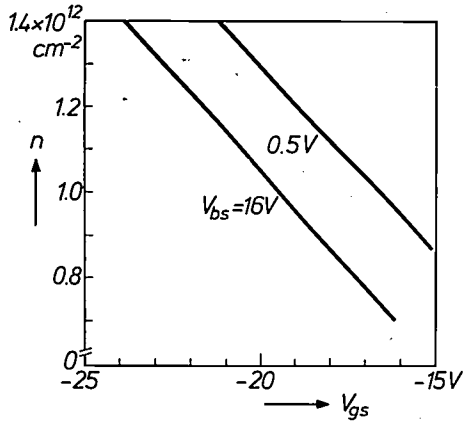


Fig. 8. Effect of a reverse bias V_{bs} on the carrier density n as a function of V_{gs} for the same channel as in fig. 7.

show an example of the results. It is always found that the mobility μ_H decreases with increasing V_{bs} for any given values of V_{gs} or n, at least in the range between threshold and V_{gm}. This may be due to scattering by charges in the depletion layer. Another possible factor is that the average field normal to the channel increases with V_{bs}. For $|V_{gs}| \gg |V_{gm}|$ the mobility, in some cases, does not appear to be affected by V_{bs}.

The slope of n versus V_{gs} remains constant and independent of V_{bs}, confirming again the absence of trapping (fig. 8). The difference between the values of n for $V_{bs} = 0$ and $V_{bs} \neq 0$, at given V_{gs}, is mainly due to charges in the depletion layer, and can be used to estimate the doping in the substrate near the surface.

Effect of temperature

Examples of variation of mobility and channel conductance with temperature between 156 °K and 285 °K are shown in *fig. 9*. It can be seen that for $|V_{gs}| > |V_{gm}|$ the mobility tends to increase when the temperature is decreased, as would be the case for scattering by thermal lattice vibrations (phonon scattering). The maximum mobility μ_{Hm} varies with temperature T as T^{-a}, where a is about 1.5 [13]. The slope of the n versus V_{gs} curve (not shown here) does not vary significantly with temperature, but the curve is slightly displaced, implying a variation of the threshold voltage with temperature.

For the N-type {100} channels of fig. 9 and only for those, dg_s/dV_{gs} becomes negative at large values of V_{gs} for $T < 200$ °K. This is because μ_H decreases faster than $1/n$. Such a decrease has been ascribed to quantization, which introduces changes in the average effective mass in {100} channels [14]. However more recent observations of a negative dg_s/dV_{gs} for other orientations and for P-type channels render this explanation doubtful [15].

Effect of the mobility variation on the characteristics of a MOS transistor

J. A. van Nielen and O. W. Memelink [16] have derived a theoretical relation between the drain current I_d and the drain and gate voltages V_{ds} and V_{gs}, taking into account the charges held by donors or acceptors in the depletion layer between the channel and the substrate. Trapping is assumed to be negligible.

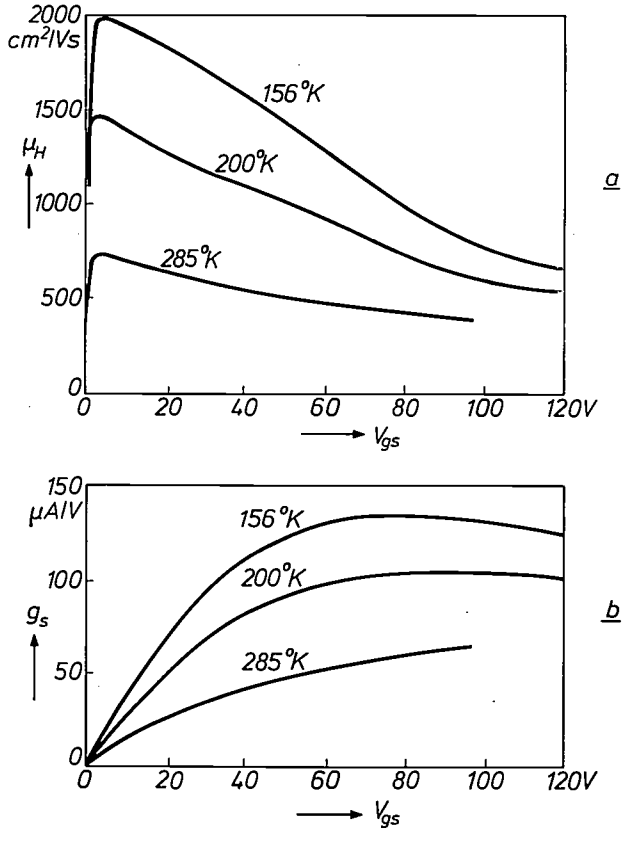


Fig. 9. Effect of temperature on *a*) carrier mobility μ_H and *b*) channel conductance g_s, both as a function of V_{gs}, for an N-type channel in a {100} plane. $V_{bs} = 0$.

The electric field in the oxide and in the depletion layer is taken normal to the surface (the "gradual-channel approximation"). This approximation is justified when, as in our case, the source-drain separation is much larger than the thickness of the depletion layer near the drain. More questionable assumptions used by these authors concern the mobility μ_c which is taken to be constant along the channel, and also independent of V_{gs}. This is not justified, since, as has been shown here, μ_c varies with V_{gs} (see also [17]) and with reverse bias V_{bs}. At any point of the channel at a potential $V(x)$ with respect to the source, the mobility is equal to the mobility of a uniform channel, with a potential $V_{gs} - V(x)$ between gate and channel, and a reverse bias $V(x)$ between channel and substrate (it is assumed that hot-electron effects can be neglected).

Fig. 10 shows the effect of the mobility variation on the I_d-V_{ds} characteristics. The solid curves represent experimental data for the large experimental MOST (*AB* in fig. 1) with a $P\{111\}$ channel. The dashed curves correspond to Van Nielen and Memelink's relation, with the mobility μ_c assumed constant along the channel and adjusted to fit the experimental curves at low values of V_{ds}. For $V_{gs} = -8.5$ V the discrepancy between the calculated and the measured current is as large as 40%. It can be seen that the results are qualitatively consistent with our previous observations:

1) For small values of V_{ds} the mobility μ_c varies with V_{gs} in the way shown by the typical curve for $V_{bs} = 0.5$ V of fig. 7 (note the maximum of μ_c at $V_{gs} \approx -8.5$ V).
2) For higher values of V_{ds} the drain current I_d falls below the theoretical curve; this corresponds to a decrease in the average value of μ_c over the channel, due to the reverse bias created at each point of the channel by the potential drop between source and drain. For very large values of V_{ds}, where I_d is close to saturation, the discrepancies between the theoretical and the experimental curves may in part be due to inaccuracies in the estimated values of the threshold voltage and the bulk doping.
3) The misfit of I_d is largest at $|V_{gs}|$-values close to and below the mobility maximum, in accordance with fig. 7.

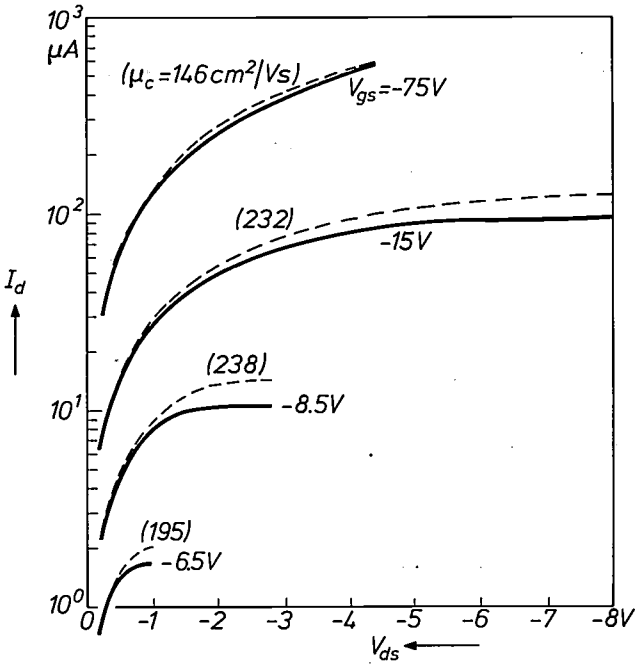


Fig. 10. Comparison of the experimental I_d-V_{ds} characteristics (solid curves) of a *P*-type {111} channel, at 285 °K, with the theoretical characteristics (dashed curves) derived under the assumption that the mobility is constant along the channel. The mobilities indicated are obtained by a best fit at small values of V_{ds}. Threshold voltage $V_{th} = -4.45$ V. Bulk doping $N_D = 3.3 \times 10^{14}$ cm^{-3}.

Quantitatively, the variation of μ_c and of the potential $V(x)$ along the channel can be derived from the experimental I_d-V_{ds} characteristics. The method is given in reference [18]. The results for $V_{ds} = -8$ V, $V_{gs} = -15$ V, for the same sample, are shown in *fig. 11*.

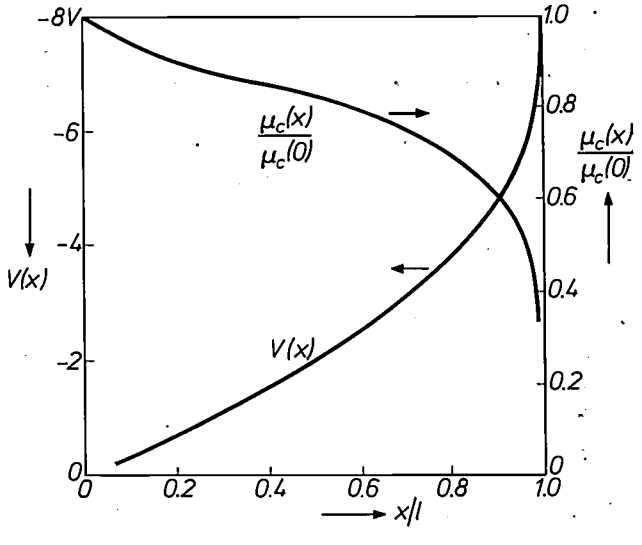


Fig. 11. The variation of the voltage $V(x)$ and the mobility $\mu_c(x)$ along the channel, as derived from the experimental I_d-V_{ds} characteristics of fig. 10. The voltage $V(x)$ and the normalized mobility $\mu_c(x)/\mu_c(0)$ are plotted as functions of the normalized distance x/l from the source (l is the source-to-drain distance), for the sample of fig. 10. $V_{ds} = -8$ V, $V_{gs} = -15$ V.

[13] J. Grosvalet, C. Jund, C. Motsch and R. Poirier, Surface Sci. 5, 49, 1966; see also the articles by Fang *et al.* under [1] and under [6].
[14] F. F. Fang and W. E. Howard, Phys. Rev. Letters 16, 797, 1966.
[15] See the article by Fang and Fowler under [1].
[16] J. A. van Nielen and O. W. Memelink, Philips Res. Repts. 22, 55, 1967.
[17] D. Frohman-Bentchkowsky, Proc. IEEE 56, 217, 1968.
[18] N. St. J. Murphy, F. Berz and I. Flinn, Solid-State Electronics 12, 775, 1969 (No. 10).

Conclusions

Summarizing our main results, we note first that the hysteresis observed in a MOST conductance as a function of gate voltage is often largely due to hysteresis in the mobility and not just to hysteresis in the carrier density. Secondly, our results for surface density of free carriers lead to the conclusion that there is no significant trapping in the region where the channel is well developed.

Our most significant observation, however, is the strong variation of surface mobility with gate voltage. The variation is of practical importance as it occurs mainly at gate voltages between threshold and some 10 V above (for oxide thickness of 0.2 to 0.3 μm), a range in which practical MOSTs often operate. The mobility is also affected by a bias between channel and substrate. The variation of the mobility should be taken into account when calculating the characteristics of a MOST. The origin of the surface scattering producing the observed variation is not well understood, and presents a challenging problem to the physicist.

Summary. The free-carrier surface density and surface mobility in large experimental MOSTs are obtained from conductivity and Hall measurements. Results are given for uniform *N*- and *P*-type channels, with various crystallographic orientations, and over a range of temperatures and reverse biases between channel and substrate. It is found that trapping of free carriers does not exceed a few per cent when the gate voltage is more than a few volts above threshold. The mobility of free carriers is very small at threshold. It increases rapidly with gate voltage and rises to a maximum of about $\frac{1}{3}$ or $\frac{1}{2}$ of its bulk value at gate voltages corresponding to free-carrier surface densities below 10^{12} carriers/cm^2. At larger gate voltages there is a slow decrease in the mobility. The mobility is also affected by a reverse bias between channel and substrate. It is shown that these mobility variations have an appreciable influence on the drain characteristics of MOSTs.

Integrated audio amplifiers with high input impedance and low noise

R. J. Nienhuis

The MOS transistor as an audio amplifier

An amplifier to be used with a voltage source of very high internal impedance like a crystal pick-up should have a high input impedance (1 MΩ or more) and introduce a minimum of noise when connected to such a voltage source. With bipolar transistors it is difficult to meet these requirements: the source impedance at which the noise factor is a minimum is no more than a few kΩ. However, the MOS transistor seems more promising in this respect. It has a very high input impedance and gives a minimum noise figure at a high internal impedance of the signal source. A drawback, however, is its low transconductance. An amplifier that can amplify the signal from a crystal pick-up sufficiently to enable it to drive a conventional output transistor directly should have a transconductance of about 50 mA/V, and a MOS transistor only has a transconductance of about 3 mA/V.

In this article we shall describe how simple integrated circuits consisting of a combination of a MOS transistor and one or two bipolar transistors can be used to give the required transconductance while preserving the desirable features of the single MOS transistor. In conclusion we shall give a brief description of a complete record-player amplifier in which an integrated circuit of this type is incorporated as a preamplifier.

Amplifier circuits with MOS transistors and bipolar transistors

A much higher transconductance than that of a single MOS transistor can be obtained by putting a MOS transistor in cascade with a bipolar transistor. The drain current of the MOS transistor then forms the base current of the bipolar transistor (*fig. 1*). This current is thus amplified in the bipolar transistor, by the current amplification factor α' [1]. However, this does not necessarily mean that the transconductance of the circuit of fig. 1 is a factor of α' greater than that of a single MOS transistor. If, for instance, we do not want the addition of the bipolar transistor to make the total current consumption larger, we must make the current in the MOS transistor α' times smaller. Now the transconductance of a MOS transistor is proportional to the square root of the current [2], so the addition of the bipolar transistor only makes the total transconductance increase by a factor of $\sqrt{\alpha'}$. Thus, if α' is 100, one can only gain a factor of 10 in this way.

A further increase in transconductance can be obtained by adding a second bipolar transistor to the circuit (*fig. 2*). This means, of course, that for a given total current consumption the current in the MOS transistor is now even smaller. It can drop to such a low value (a few microamperes) that the mobility of the charge carriers decreases [3], and the transconductance becomes even smaller than would be expected from the square-root relation with the current. The improvement in transconductance by a factor of $\sqrt{\alpha_1'\alpha_2'}$ that might at first sight be expected cannot therefore be obtained in this way. With this circuit we were able to obtain a transconductance of 40 mA/V at a current consumption of 10 mA.

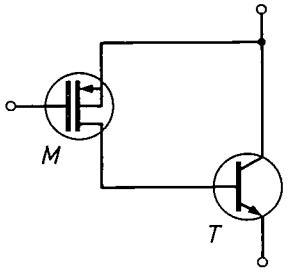


Fig. 1. Cascade circuit of a MOS transistor *M* and a bipolar transistor *T*. This circuit has a higher transconductance than a single MOS transistor.

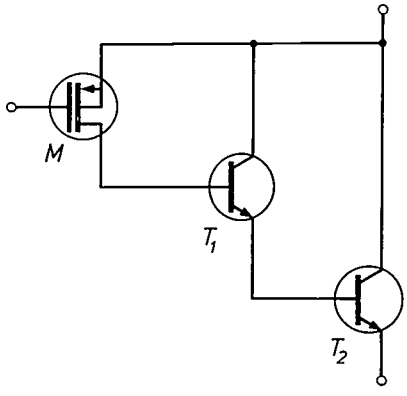


Fig. 2. An even higher transconductance can be obtained with a cascade arrangement of a MOS transistor *M* and *two* bipolar transistors, T_1 and T_2.

Ir. R. J. Nienhuis is with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven.

[1] In the fourpole-network theory of the transistor this factor is also denoted by h_{fE} or h_{12}.

[2] See equation (11) in the article by J. A. van Nielen in this issue, page 209.

[3] See the article by N. St. J. Murphy, F. Berz and I. Flinn in this issue, page 237.

Besides the addition of a second bipolar transistor, there is another way of making the transconductance larger, and this is by increasing the current in the MOS transistor. This can be done by putting a resistor between the base and emitter of the bipolar transistor (*fig. 3*). Owing to the increased current the transconductance of the MOS transistor is now greater, to such an extent that the transconductance of the circuit also increases even though a part of the current flows through the resistance and is thus lost to the bipolar transistor. Of course, the resistance chosen should not be too small. The optimum value was found to be about 800 Ω, at which, for a current consumption of 10 mA, a transconductance of 40 to 80 mA/V was reached.

Integrated versions were made of the amplifiers in fig. 2 and fig. 3. *Fig. 4* is an enlarged photograph of the monolithic circuit corresponding to the diagram of fig. 3. The crystal chip measures 0.5×0.5 mm. The location of the transistors, which enclose one another in the plane of the crystal, is indicated by the lines linked to the diagram.

A brief description will now be given of some of the characteristics of the amplifiers shown in figs. 2 and 3.

Distortion

Apart from its greater transconductance the circuit of fig. 3 has the further advantage over the circuit of fig. 2 that it distorts the signal less. In the circuit with two bipolar transistors the current in the MOS transistor is very small, as we noted earlier, and this transistor therefore introduces fairly considerable distortion. The circuit of fig. 3 is better in this respect; the distortion is smaller because of the higher bias current in the MOS transistor. With decreasing R, however, the distortion in the bipolar transistor increases. (To obtain low distortion a transistor of this type should be driven by a signal source with a high internal impedance.) There should therefore be some value of R for which the distortion is at a minimum. This is in fact the case as can be seen from *fig. 5*, which shows the relative amplitude of the second harmonic as a function of R, with the peak value of the fundamental as a parameter. Here again the d.c. current flowing in the circuit was 10 mA. It can be seen that the distortion is least when R is about 3 kΩ. Fig. 5 also shows the transconductance of the circuit; this has a maximum at $R = 800$ Ω as we saw earlier. To obtain both high transconductance and low distortion at the same time some kind of compromise has to be made. A value of 1.2 kΩ for R gives the best results.

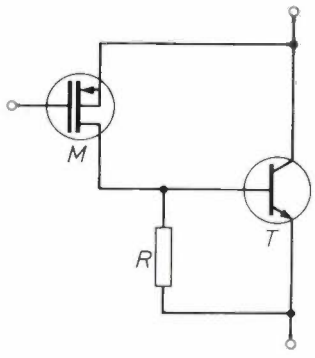


Fig. 3. Putting the resistance R into a circuit like that of fig. 1 increases the current in the MOS transistor and hence its transconductance. Provided R is not too low, the transconductance of the whole circuit is also increased.

Input and output impedance

The *input impedance* of the circuits of fig. 2 and fig. 3 is equal to that of the MOS transistor, i.e. to that of a parallel arrangement of a capacitor of about 10 pF and a resistor of 10^{12} Ω. This impedance is so high as to be of no practical significance at audio frequencies.

The *output impedance* is virtually a pure resistance which depends on the operating current of the amplifier. With the circuit that has two bipolar transistors (fig. 2) this is simply the internal impedance of the

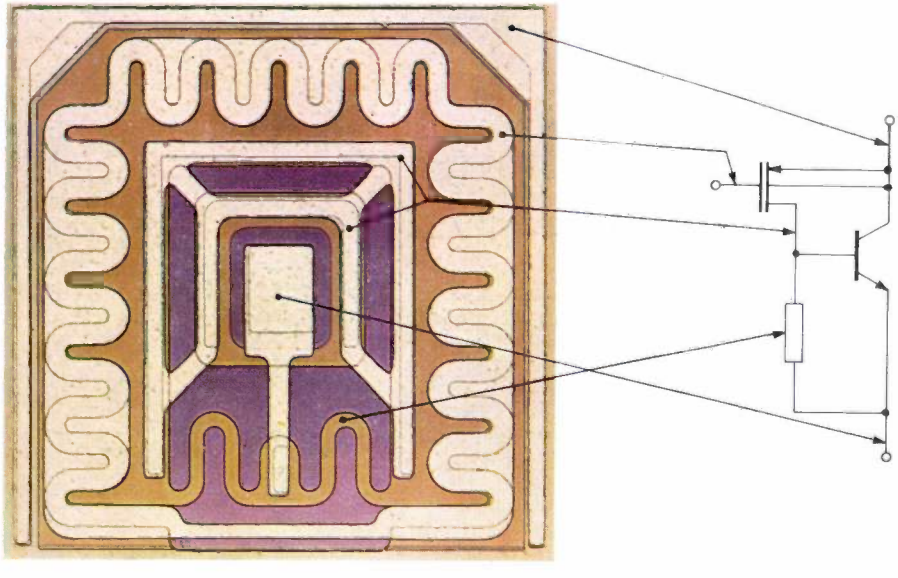


Fig. 4. Enlarged photograph of a monolithic circuit made to the circuit of fig. 3. The crystal is 0.5×0.5 mm. The location of the transistors and the resistor is indicated by the lines linked to the diagram.

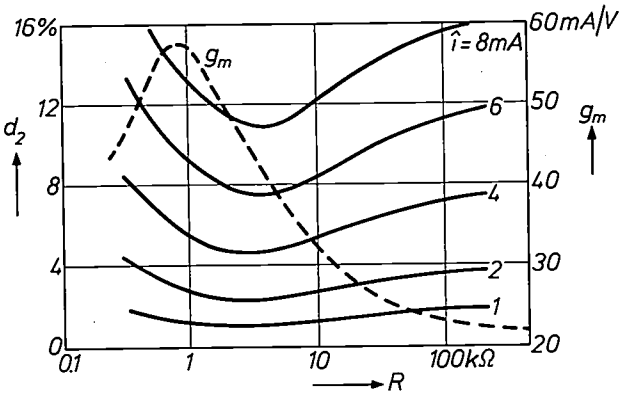


Fig. 5. The distortion in the circuit of fig. 3, with an operating point such that the d.c. current taken is 10 mA. The relative magnitude d_2 of the second harmonic appearing in the output signal for a sinusoidal input signal is shown as a function of R. The parameter $\hat{i}$ is the peak value of the alternating output current. A curve of the transconductance g_m is also shown.

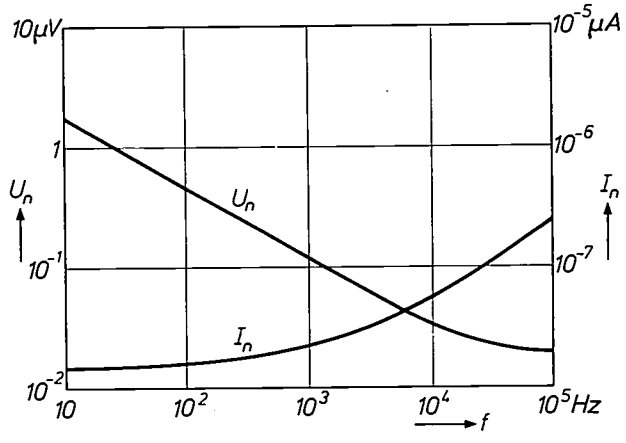


Fig. 6. Equivalent noise voltage U_n and equivalent noise current I_n (both in a frequency band of 1 Hz) as a function of frequency f for an arbitrary MOS transistor.

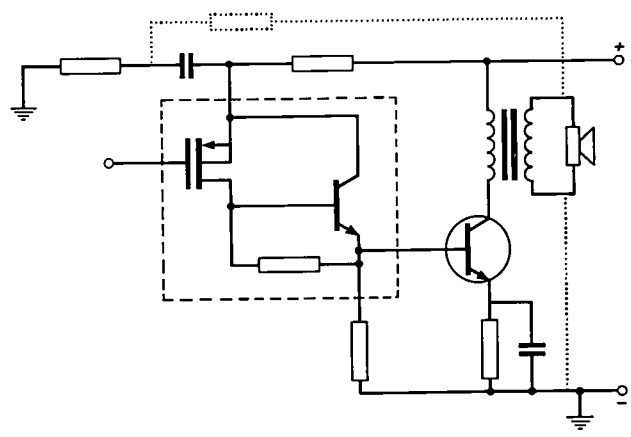


Fig. 7. Circuit of a record-player amplifier consisting of an integrated preamplifier and an output stage. The part inside the dashed line is an integrated monolithic circuit like that of fig. 4. The output power is 2 W, the distortion 4.5%. The distortion can be reduced to 3% by using negative feedback (dotted lines).

MOS transistor (at a very low current) divided by the product of the current amplification factors of the bipolar transistors. At a current of 10 mA the output resistance is about 5 kΩ. In the circuit shown in fig. 3 the output resistance is partly determined by the resistance R. The output resistance is therefore smaller; it is about 2.5 kΩ at a current of 10 mA. These values are high enough for each of these circuits to be able to drive an output stage at an acceptable distortion level.

Noise

In the two circuits discussed here the noise is entirely determined by the MOS transistor, so that in this respect the two circuits are equivalent. *Fig. 6* gives an example of the noise level of a MOS transistor, showing the equivalent noise input voltage and input current per unit bandwidth in Hz. The noise *current* is found to be so small that it plays no part even when the signal source has an internal resistance as high as that of most crystal pick-ups. To determine the total noise we need therefore only take into account the noise *voltage*. Turning to the audio frequency band (15 Hz-15 kHz), we find an r.m.s. value of 7.5 μV for the noise voltage. If the voltage output from the pick-up is 100 mV, then the signal-to-noise ratio of the MOS transistor in fig. 6 is 13 300 or 82 dB.,

A record-player amplifier with integrated preamplifier

With a preamplifier that can drive an output stage directly, like the ones discussed above, a complete record-player amplifier becomes very simple. *Fig. 7* shows the circuit of an amplifier that can deliver an output of 2 W and has an integrated preamplifier of the type shown in fig. 4. With an input signal of 100 mV the measured signal-to-noise ratio for this amplifier was 73 dB. At the same output the relative magnitude of the second harmonic is 4.5%, which is acceptable for non-professional equipment. If necessary the distortion can be reduced to 3% by introducing negative feedback, as indicated in the diagram by the dashed lines.

Summary. Because of its high input impedance a MOS transistor makes a very useful preamplifier for the signal voltage from a crystal pick-up. The transconductance of a MOS transistor is too low, however, for driving an output stage directly. The transconductance can be increased by connecting a MOS transistor in cascade with one or two bipolar transistors. Two circuits are discussed, both of which have been produced in the laboratory in integrated form. Since it is desirable to set the operating point for the MOS transistor at not too low a current, the current in one of the amplifiers was increased by putting a resistor between base and emitter of the bipolar transistor. A transconductance between 40 and 80 mA/V could be obtained in this way. Finally a circuit diagram is given for a complete record-player amplifier consisting of an integrated preamplifier and an output stage. The output power is 2 W with a distortion of 4.5% and a signal-to-noise ratio of 73 dB. The distortion can be reduced to 3% by using negative feedback.

An integrated chopper circuit with MOS transistors

B. J. M. Overgoor

When direct-coupled amplifiers are used to amplify small d.c. voltages, it is usually found that there is an output signal present even when there is no signal at the input. The output signal can then be made zero by applying to the input a particular signal referred to as the *offset voltage*. As a rule this signal is not constant, because the characteristics of transistors vary with temperature. For measuring instruments this is undesirable The output signal here should be zero when there is no signal at the input; in other words, the offset voltage must be zero, or at least very small.

The disadvantage of an offset voltage does not occur in an a.c. amplifier with capacitive input and output coupling; an amplifier of this type gives no output signal, apart from noise, when there is no input signal. This useful feature can be turned to advantage for amplifying d.c. signals if they are first "chopped" to form a.c. signals. When the a.c. voltage obtained in this way is amplified and then detected, the chopper-amplifier-detector system then acts as a d.c. amplifier. For chopping and detection identical or nearly identical switches are generally used, and these are driven by the same voltage.

The output signal of a direct-coupled d.c. amplifier contains information not only about the *magnitude* of the input signal but also about its *polarity*. Because of this information such an amplifier is suitable for use in a feedback circuit. A system consisting of a chopper, an amplifier and a detector can also be applied in this way by making use of the fact that the phase of the amplifier output signal is equal (or opposite) to the phase of the input signal. The chopper must then deliver an a.c. voltage whose phase is determined by the polarity of the input signal, and the detector must produce a d.c. voltage whose polarity depends on the phase of its input signal. This is referred to a synchronous or phase-sensitive detection; the phase of the converter control signal is the reference phase.

Of course, full benefit can only be derived from the indicated advantages of the chopper-amplifier-detector system if the d.c. voltage can be converted into an a.c. one in a simple way. The chopping process was at first effected by electromechanical switches which periodically broke the connection between the signal source and the amplifier. Later, these were superseded by electronic circuits based on devices such as photoresistors and bipolar transistors.

An objection to the use of a bipolar transistor as a switch is that it gives an offset voltage of between a few tenths of a millivolt and several millivolts. This offset voltage is difficult to compensate since it varies appreciably with temperature. When a photoresistor is used the offset voltage is much smaller, but then other difficulties are encountered: the switching frequency is limited to about 100 Hz, and chopping with a light source makes the equipment rather complicated.

In this article we describe choppers based on MOS transistors. An advantage of this type of transistor is that there is no offset voltage between source and drain. However, there are stray capacitances, although not to the same extent as in bipolar transistors, particularly between the gate and the source and drain. When a square-wave voltage is applied to the gate to switch periodically between the conducting and non-conducting state, these capacitances induce periodic voltage peaks (spikes) at both source and drain (*fig. 1*); these may even be high enough to overload the amplifier. As a result of these switching peaks a d.c. voltage appears at the output of the detector; in other words, there is again an offset voltage. This voltage is proportional to the frequency of the square-wave voltage at the gate, and also depends on the impedance of the signal source and on temperature.

A considerable improvement in this respect is obtained by using two MOS transistors instead of one, and even better results are obtained with four MOS transistors in a balanced circuit. In this article we shall describe two chopper circuits, one with two MOS transistors and one with four. The performance of the chopper with four MOS transistors has been considerably improved by making it as an integrated circuit.

The series-shunt chopper

In the chopper with two MOS transistors, one is connected in series with the input of the amplifier and the other in parallel with it (*fig. 2*). These transistors are then driven in opposite phase. The voltage peaks now appearing on the interconnected drain electrodes D_1 and D_2 occur simultaneously and are in opposite phase. They do not, however, cancel each other completely, since one transistor goes from the conducting to the non-conducting state while the other transistor

Ir. B. J. M. Overgoor is with the Philips Electronic Components and Materials Division (Elcoma), Nijmegen.

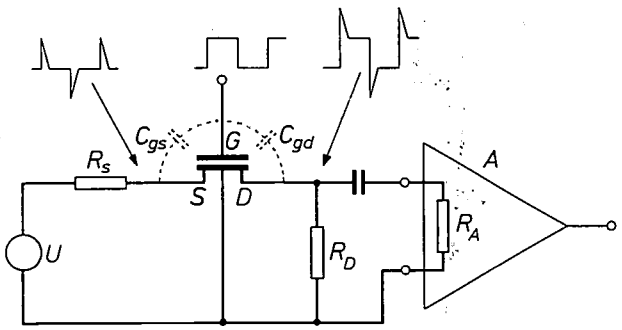


Fig. 1. A MOS transistor as a switch or "chopper" that periodically breaks the connection between the voltage source U to be measured and the amplifier A. G gate. S source. D drain. R_S internal impedance of the voltage source U. The resistance R_D and the internal impedance R_A of the amplifier together form the load. When the chopper switches, voltage peaks due to the capacitances C_{gs} and C_{gd} appear at S and D. The waveform of the voltages at G, S and D is shown in the figure.

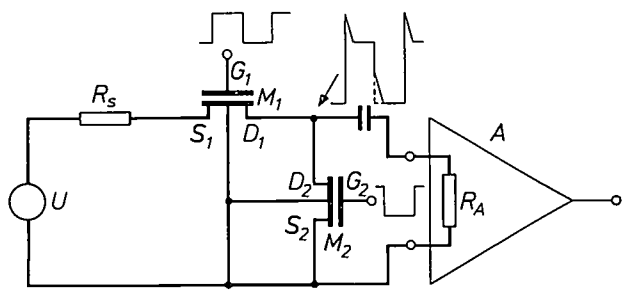


Fig. 2. Circuit of a series-shunt chopper with two MOS transistors M_1 and M_2 driven in opposite phase. The switching peaks now occurring are much smaller than in the circuit of fig. 1 and are also of the same sign.

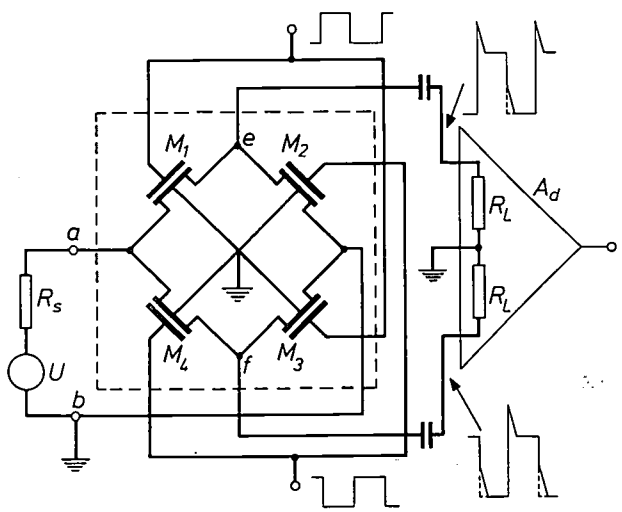


Fig. 3. Diagram of a chopper with four MOS transistors in a balanced circuit. The voltage to be measured is applied between points a and b. The a.c. voltage to be amplified, which appears between points e and f, is fed to the differential amplifier A_d. M_1 and M_3 are driven in the opposite phase to M_2 and M_4. The part of the circuit inside the dashed line is made as an integrated circuit.

does just the opposite; the symmetry is not therefore completely perfect. If the two MOS transistors have exactly the same characteristics, and the impedances in the leads to the source electrodes are identical, the resultant voltage peaks will be equal in magnitude and phase. There is then no offset voltage on synchronous detection. However, the transistors and impedances are not usually identical, and therefore two successive peaks do not have the same magnitude and phase, and an offset voltage is detected. Moreover, the peaks can still overload the amplifier.

Chopper with four MOS transistors

The disadvantages connected with the occurrence of switching peaks can be largely avoided by using a circuit consisting of two series-shunt choppers operated in opposite phase (*fig. 3*). The d.c. voltage to be measured is applied between points a and b. The a.c. voltage to be amplified, which appears between points e and f, is fed via coupling capacitors to a differential amplifier which has a high rejection factor for common-mode signals. In the same way as with the series-shunt converter discussed above, voltage peaks now appear at the two resistors R_L. If the transistors are identical, these peaks are of equal magnitude and sign and therefore are not amplified, or only very slightly, in the differential amplifier. The a.c. voltage signal is amplified, however; this has a peak-to-peak value twice as high as the d.c. voltage to be measured, and appears in full as a differential or series-mode signal at the input of the amplifier. Since the amplifier is not now driven by the switching peaks, it can be given a much greater amplification than when a single series-shunt circuit is used.

A balanced chopper like that of fig. 3 has been produced in our laboratory in the form of an integrated circuit with four MOS transistors of the P-channel enhancement type. This automatically gives four virtually identical devices. Moreover, in an integrated circuit there are no temperature differences between the transistors, which could cause undesirable thermo-electric effects that would contribute to the offset voltage.

The magnitude of the switching peaks depends of course on the amplitude of the square-wave voltage at the gates. This voltage should therefore be no higher than is necessary to switch the MOS transistors. To give a general rule, the bias and the amplitude of the square-wave voltage at the gate should have values that allow the instantaneous value of the voltage at the gate to go no further than 0.5 to 1 volt into the region of the non-conducting state, even when the spread in the threshold voltages is taken into account.

Finally, one or two other advantages of the balanced

circuit over the series-shunt chopper should be mentioned. First of all, the voltage peaks occurring at the *input* terminals are much smaller in the balanced circuit. This is because the peaks simultaneously induced at point *a* are of opposite sign, since transistors M_1 and M_4 are driven in opposite phase. The same applies to M_2 and M_3 with respect to point *b*.

During the switching of the MOS transistors it may happen that all four are in the conducting state for a short time, so that the source of the input signal is momentarily short-circuited. This can also happen with the series-shunt chopper. To avoid this periodic short-circuiting the transistors are driven with a square-wave voltage whose rise time is different from the fall time. However, this degrades the symmetry and the successive switching peaks are unequal. In the series-shunt chopper this gives an increase in the offset voltage, but the effect is much less troublesome in the balanced chopper, because the simultaneously occurring switching peaks are applied to the difference amplifier as common-mode signals.

An incidental advantage of the balanced chopper is that the signal source works into a constant load, and not a periodically varying one as with the series-shunt chopper. The input impedance of a balanced chopper with amplifier is therefore equal to that of the amplifier.

The small offset voltage that remains in the integrated choppers described here depends on the residual inequality of the four MOSTs and the internal impedance of the voltage source to be measured. The same applies to the temperature coefficient of this offset voltage, which amounts to about 1 % per °C for this signal. Furthermore, both quantities increase with the switching frequency. An idea of the magnitude of the offset voltage can be obtained from *fig. 4*, where it is shown as a function of the switching frequency f_s for various values of the signal source impedance. It can be seen that the offset voltage is proportional to $f_s^{1.3}$. The increase with f_s is greater than linear because the effect of a voltage peak has not yet disappeared when the next peak arrives; the contribution which this makes to the offset voltage increases with the rate of arrival of the peaks.

It should be noted, finally, that the measurements whose results are shown in fig. 4 were carried out using a signal source with a symmetrical arrangement of the resistance R_s. An asymmetric configuration (see fig. 3) gives a larger offset voltage.

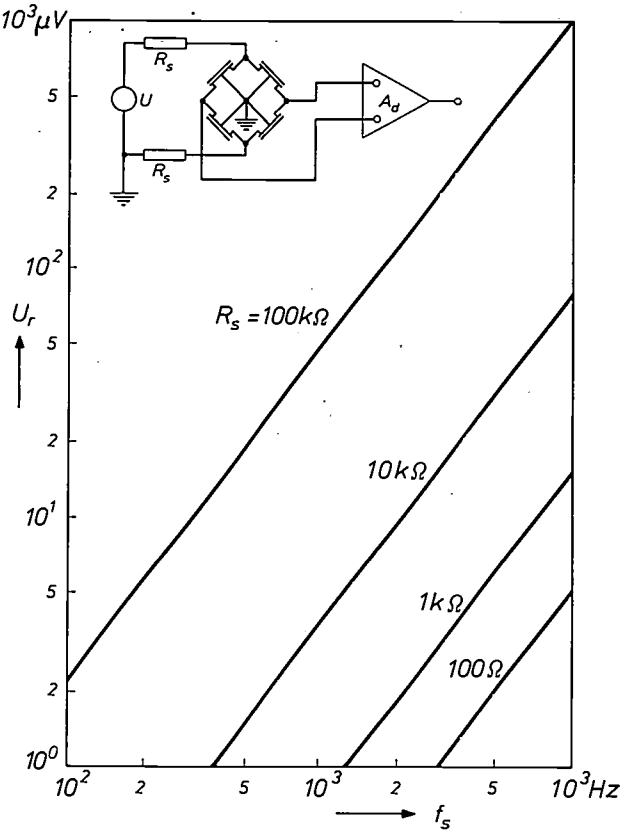


Fig. 4. The offset voltage U_r, due to the switching peaks, which occurs when a chopper like the one of fig. 3 is symmetrically connected to the voltage source, shown as a function of the switching frequency f_s. The four curves relate to different values of the internal impedance R_s of the voltage source. Thermoelectric effects are not taken into account. The threshold voltage of the four MOS transistors was between −3.0 and −3.5 V, and the upper and lower levels of the square-wave driving signal were −2.5 and −8.5 V.

Summary. A d.c./a.c. converter in the form of an integrated circuit comprising four MOS transistors has been developed for the amplification of weak d.c. signals by means of an a.c. amplifier. It has a very small residual signal, due entirely to small differences between the voltage peaks that arise through the switching of the transistors.

MOS transistors for power amplification in the HF band

R. D. Josephy

Introduction

The MOS transistor has several attractive features as a high-frequency power amplifier. One feature is that it can be used with a high supply voltage; another is its square-law characteristic. The absence of odd-order terms from this characteristic means that two frequency components lying within the passband of the high-frequency amplifier will not give rise to intermodulation products — i.e. sum or difference frequencies — falling within this passband. All of the sum and difference frequencies created by the square-law characteristic lie far outside this band and are rejected by the band-pass filters of the amplifier. In this way the MOS transistor can provide linear operation as a high-frequency amplifier.

Furthermore, in contrast to the bipolar transistor, the MOS transistor has a negative temperature coefficient of current at high current levels, and therefore a MOS transistor tends to be thermally stable even when its area is large. This leads to uniform temperature distribution over the transistor, and to freedom from thermal runaway and second breakdown, which can be serious problems in the design of bipolar power transistors.

The MOS transistor is therefore a potentially useful device for high-frequency power amplification. This is why a MOS power transistor is being developed to replace the valve in the output stages of a single-sideband transmitter operating in the 3 to 30 MHz frequency range. In this application the power at the maximum value of the envelope of the amplitude-modulated carrier — the peak envelope power — amounts to 100 W and the intermodulation-product level should be better than —30 dB. This article describes the first stages of the development, which have resulted in MOS transistors delivering an output power of 30 W with an intermodulation level below the specified value. The article starts with a discussion of design considerations connected with voltage and current limitation and the effects of frequency on performance. Experimental results are given next, and the article closes with an indication of probable directions of future advances in MOS power transistor design.

R. D. Josephy, B.A., who was with Associated Semiconductor Manufacturers Ltd. at Wembley, England, while this work was in progress, is now with The General Electric Company Ltd.

Design considerations

Power output

The output power P_o available from any transistor is proportional to the product of the maximum peak voltage swing which can be maintained across it and the peak current I_{max} which it can safely pass. For a MOS transistor the maximum voltage swing which the device can withstand is the difference between the drain-source breakdown voltage, $V_{ds\ br}$, and the drain saturation voltage at the maximum current, $V_{ds\ sat}$. Hence:

$$P_o \propto I_{max}(V_{ds\ br} - V_{ds\ sat}). \quad . \ . \ . \quad (1)$$

The saturation voltage of a MOS transistor made on high-resistivity material is approximately equal to the "effective gate voltage" V_g', i.e. the gate voltage measured from the threshold point [1]. As is apparent from *fig. 1*, this leads to saturation voltages at high currents which are an appreciable fraction of the supply voltage. It is desirable therefore to make the supply voltage as large as possible and to design a MOS power transistor with a large drain-source breakdown voltage.

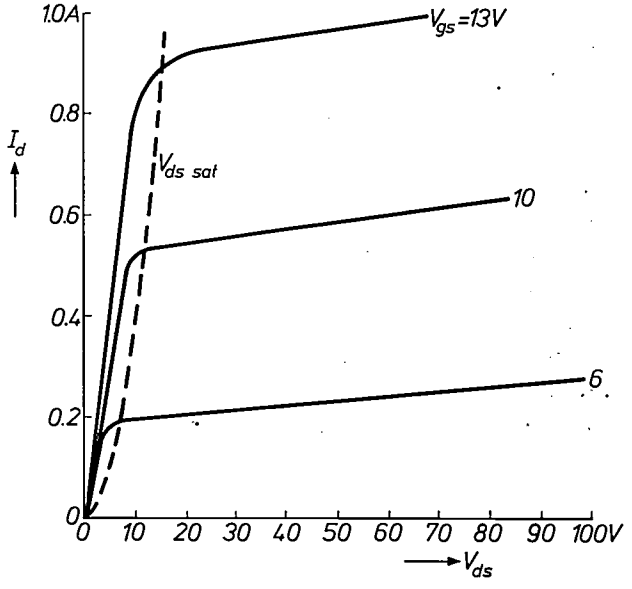


Fig. 1. Example of the I_d-V_{ds} characteristics of a MOS power transistor. The knee voltage $V_{ds\ sat}$ increases with current.

[1] See M. B. Das, Solid-State Electronics **11**, 305-322, 1968. The fundamentals and d.c. performance of MOS transistors are discussed in the article by J. A. van Nielen in this issue, page 209.

Voltage limitations on the MOS transistor

There are three types of voltage limitation for the MOS transistor: destructive breakdown of the gate oxide, avalanche breakdown of the drain junction, and punch-through between source and drain.

The first of these, destructive breakdown of the gate oxide, occurs at a field of approximately 10^7 V/cm. However, for large-area devices, a considerable margin of safety must be allowed since weak points will exist in the oxide under the gate and breakdown may take place at a lower value than expected. For a *P*-channel MOS transistor in class B operation the maximum voltage appears across the oxide at the drain when the gate voltage is at the peak of its positive swing and the drain voltage at its most negative (*fig. 2b*). Voltages of the order of 100-120 V can occur under these conditions and an oxide thickness of at least 0.2 μm is therefore required.

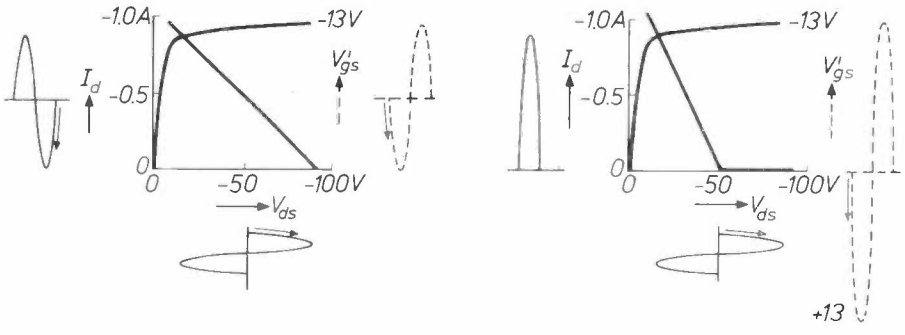


Fig. 2. Class A operation (*on the left*) and class B operation (*on the right*) of a *P*-channel MOS power transistor. In either case the drain voltage swing is limited by the drain-source breakdown voltage on one side and the saturation voltage on the other side. A large voltage difference between drain and gate occurs in class B operation at maximum drain voltage V_{ds}.

The avalanche breakdown voltage of the drain junction depends on the impurity-doping level on the substrate side of the junction: it is inversely proportional to the square root of the donor (or acceptor) atom concentration in the substrate. This breakdown voltage is also influenced by the radii of curvature of the junction [2], and by the vertical field between gate and substrate close to the drain [3]. These two factors combine to determine the field distribution in the drain depletion region and hence the junction breakdown voltage. For radii of curvature less than about 4 μm the breakdown voltage falls quite rapidly with decreasing radius, and so a junction depth of this value or higher should be used for the MOS power transistor. The dependence of drain-junction breakdown on gate voltage is shown in *fig. 3*, from which it can be seen that the breakdown voltage is lowest when the gate voltage is such as to turn the transistor off. This occurs because under these conditions the drain depletion region is made narrower at the surface by the gate field and so the field in the depletion region reaches its breakdown value at a lower drain voltage. This effect is most unfavourable when the device is used as a class B amplifier. In this mode of operation, during the cut-off half-period the drain voltage and the gate voltage swing simultaneously to their maximum values, which are of opposite sign (this is illustrated in fig. 2*b* for a *P*-channel MOS transistor). A thicker oxide may not be used to decrease the gate field because a given peak current must be reached during the other half-period and this requires a gate field of a given strength. With class A operation the situation is less critical, since the effective gate voltage is zero at maximum drain voltage (fig. 2*a*).

Punch-through breakdown takes place when the substrate resistivity is so high and the channel so short that the drain depletion region can extend to the source before avalanche breakdown occurs [4]. When the device is cut off by the gate, punch-through breakdown is similar in appearance to avalanche breakdown except that it is "softer", and its variation with gate field is in the opposite direction. When the device is conducting,

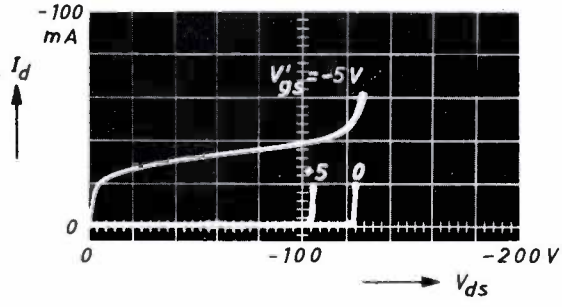
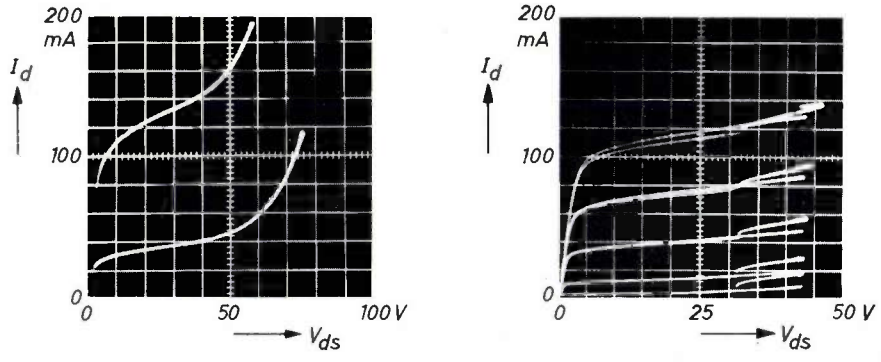


Fig. 3. Measured I_d-V_{ds} characteristics of a *P*-channel MOS power transistor showing breakdown. The breakdown voltage becomes lower when the effective gate voltage V'_{gs} turns the transistor off.

the effect of the extension of the drain depletion region to the source is to cause the drain current to depart from saturation and to vary approximately as the square of the drain voltage as it becomes space-charge limited [5]. This is shown for an *N*-channel MOS transistor in *fig. 4a*. The departure from saturation leads to a departure from the square-law I_d-V_{gs} characteristic, which causes intermodulation distortion in linear amplifiers.

Figures 3 and 4*a* relate to two MOS transistors of the same geometry and substrate doping level, but one with a *P* channel (fig. 3) and one with an *N* channel (fig. 4*a*). be used in designing a MOS transistor with a high voltage rating. For a *P*-channel MOS transistor the device whose characteristics are shown in fig. 3 has been found to be approximately optimum for drain-voltage rating. This device is made on 10 Ωcm *N*-type silicon, has a channel length of 9 μm and an oxide thickness of 0.2 μm, and is limited by avalanche breakdown. For an *N*-channel device of similar structure the voltage limitation would be much lower, as shown in fig. 4*a*.

Another type of drain-source breakdown, encountered in *N*-channel MOS transistors with large source



a b

Fig. 4. Measured I_d-V_{ds} characteristics of an *N*-channel MOS power transistor of the same geometry and bulk-doping level as the *P*-channel MOS transistor of fig. 3. *a*) Low-concentration source and drain diffusions; breakdown occurs by punch-through. *b*) High-concentration source and drain diffusions; another type of breakdown occurs introducing negative resistance into the characteristics. As soon as the drain voltage V_{ds} has reached about 43 volts the curve jumps back and starts again at a lower voltage and a somewhat higher current.

The drain-voltage limitation results from drain-junction breakdown in the *P*-channel device and from punch-through in the *N*-channel device. This difference is believed to be caused, at least partly, by impurity redistribution during thermal oxidation which, in an *N*-channel MOS transistor gives a very low impurity concentration close to the silicon surface. Hence the drain depletion region spreads considerably further towards the source than in a *P*-channel MOS transistor where the effect of redistribution is the opposite, producing an enhanced impurity concentration near the surface.

To predict accurately the conditions under which avalanche and punch-through breakdown occur in a MOS transistor, a two-dimensional analysis of the field in the drain depletion region is required, which must take account of variation of substrate doping level with distance from the surface. A satisfactory analysis of this kind has not yet been carried out [6], and so experimental results and approximate calculations must and drain surface concentrations, is shown in fig. 4*b*. This breakdown introduces a negative resistance into the characteristic when this has reached a given drain voltage; at this point the curve suddenly jumps back to a lower drain voltage value and restarts there at a somewhat higher current. The phenomenon is potentially catastrophic; it has been shown to occur locally by light emissions and by measurements of the surface

[2] S. M. Sze and G. Gibbons, Solid-State Electronics **9**, 831-845, 1966.

[3] A. S. Grove, O. Leistiko, Jr., and W. W. Hooper, IEEE Trans. **ED-14**, 157-162, 1967.

[4] See the article by J. A. van Nielen in this issue, page 209.

[5] G. F. Neumark and E. S. Rittner, Transition from pentode- to triode-like characteristics in field-effect transistors, Solid-State Electronics **10**, 299-304, 1967. The effect has been turned to practical use by P. Richman, Modulation of space-charge-limited current flow in insulated-gate field-effect tetrodes, IEEE Trans. **ED-16**, 759-766, 1969 (No. 9).

[6] Two-dimensional analyses for uniform doping level have been carried out by J. E. Schroeder and R. S. Muller, IGFET analysis through numerical solution of Poisson's equation, IEEE Trans. **ED-15**, 954-961, 1968, and by H. C. de Graaff, Gate-controlled surface breakdown in silicon *p-n* junctions, Philips Res. Repts. **25**, 21-32, 1970 (No. 1).

temperature of the chip using an infra-red microscope.

On account of these various effects *P*-channel MOS transistors are to be preferred for power devices, in spite of their lower channel mobility, until the voltage characteristics of *N*-channel MOS transistors can be improved.

Current and gain factor

In the saturated region of operation the drain current of a MOS transistor is given by:

$$I_d = \tfrac{1}{2}\beta V_{gs}'^2, \quad \ldots \quad (2)$$

where the current gain factor β is given by:

$$\beta = \mu C_{ox} w/l. \quad \ldots \quad (3)$$

Here μ is the surface mobility for holes or electrons in *P*- and *N*-channel MOS transistors respectively; C_{ox} is the gate capacitance per unit area, w is the channel width, and l is the channel length.

There are thus three variables, C_{ox}, w and l, in the device geometry which can be adjusted to obtain a large value of β, and hence give a high current. The minimum oxide thickness which can be used is limited by dielectric breakdown considerations to 0.2 μm when voltages of up to 120 V may appear across it. This leaves the channel dimensions, w and l, to determine β. To obtain the best results at the higher frequencies it is better to reduce l rather than increase w, as the time constant τ [7] of the ideal device, considered apart from its stray elements, is then decreased. Unfortunately l cannot be reduced to the technological limit in a power device because punch-through breakdown would then limit the drain-voltage rating to a low value. Given a channel length of 10 μm, which has been found to be a reasonable compromise, the following calculation yields the order of magnitude of w which will be required for a peak current of 1 A at a gate voltage of 10 V. The mobility μ is 150 cm²/Vs, and C_{ox} for a 0.2 μm thick oxide is 1.8×10^{-8} F/cm². From equations (2) and (3) it then follows that

$$w = 2\, I_d\, l/\mu C_{ox} V_{gs}'^2, \quad \ldots \quad (4)$$

hence $w \approx 7.4$ cm. Thus a very large channel width is required for large peak currents, and an interdigitated geometry is the best means of achieving this. *Fig. 5* shows an interdigitated power MOS transistor with a channel width of 4.2 cm and a chip size of 2.8×2.1 mm.

Because of their square-law I_d-V_{gs} characteristic and thermal stability it might be thought that MOS power transistors could be driven to extremely high currents. Unfortunately, however, the square-law behaviour does not persist at very high currents. When the field in the channel exceeds about 1.5 V/μm, the carrier mobility falls with increasing field and the carriers are said to be velocity limited. This causes a serious departure from the square-law characteristic. In a linear amplifier circuit this cannot be tolerated, and velocity limiting sets an upper limit to the current.

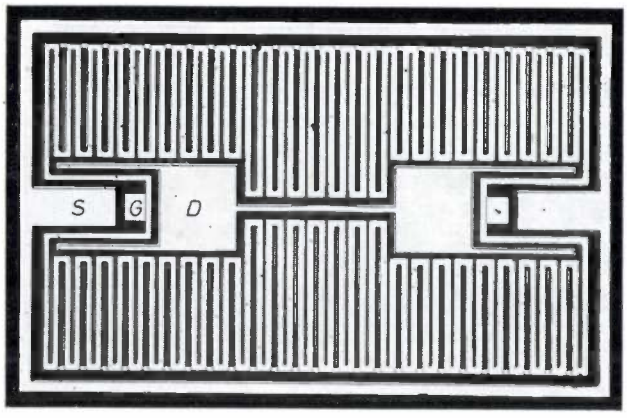


Fig. 5. MOS power transistor with interdigitated electrode structure; channel width 4.2 cm. *S* source. *G* gate. *D* drain. From the metallized source and drain areas finger-like diffusions protrude into the loops of the gate electrode. The source fingers between the gate loops are metallized with 5 μm wide aluminium strips to reduce series resistance; the drain fingers are not metallized. The gate loops are strapped together to reduce series resistance. Magnification about 30×.

So far the characteristics of the ideal device have been considered. In a real device there are stray elements which reduce the output power. The most important of these are the source and drain series resistances, R_s and R_d. These resistances arise from a combination of the resistance of the diffused source and drain fingers, the resistance of the aluminium strips on the fingers if these are used, and the aluminium-to-silicon contact resistance. Both R_s and R_d reduce the voltage appearing across the ideal device, and give a useless dissipation of power, but the source resistance also acts as a negative-feedback element, in a common-source circuit, and reduces the mutual conductance g_m. In practice R_d should be significantly less than the minimum "ON resistance" of the ideal device, which may be as low as 10 Ω. The voltage dropped across R_s should be much less than the applied gate-source voltage at the maximum current:

$$R_s I_{d\,max} \ll V_{gs'max}. \quad \ldots \quad (5)$$

Consequently R_s should be kept small compared with the ratio $V_{gs'max}/I_{d\,max}$. This ratio may typically be 5 Ω. Hence R_d should be preferably less than 1 Ω and R_s less than 0.5 Ω.

In an interdigitated structure there are two ways of reducing R_s and R_d. One is to provide aluminium strips along the diffused fingers of either the source or the drain or both. The structure shown in fig. 5 uses metallized source fingers and unmetallized drain fingers.

This is a compromise arrangement giving an acceptably small drain series resistance yet not too large an area for the device. The other approach is not to metallize either the source or the drain fingers but to make them short enough to have sufficiently low resistance. An example is shown in *fig. 6*. Such a geometry has the advantage of a higher yield of working devices because the likelihood of gate-to-source short-circuits is greatly reduced.

It is interesting to note that, in general, higher series resistances can be tolerated in MOS transistors than in bipolar transistors because of the higher impedance levels involved.

Variation in performance with frequency

For the ideal MOS transistor the time constant associated with gate capacitance and channel resistance is equal to $l^2/\mu V_{gs}'$ [7]. For a *P*-channel MOS transistor with a 10 μm channel, operated at an effective gate voltage of 10 V, this time constant is about 0.67 ns, which leads to a decrease in performance above a frequency of about 250 MHz. The stray elements may give rise to a larger time constant than this. For example the series resistances of the source and the gate form a stray time constant with the gate capacitance. It can be seen in figs. 5 and 6 that the loops of the gate are strapped together at the source and across the base of each finger. If this were not done the very large gate width necessary in a MOS power transistor could give an effective resistance of about 10 Ω in series with the gate. This, with a gate capacitance of 130 pF would give a time constant of about 1.3 ns, and the cut-off frequency would be 120 MHz.

In practice the operating frequency of a MOS power transistor may be limited by its input impedance. The equivalent circuit for the input of a MOS transistor is shown in *fig. 7*. Here $1/5g_m$ represents the distributed channel resistance, C_1 the distributed gate capacitances, and R_s and R_g are the source and gate series resistances. The overlap capacitance is neglected for simplicity. This circuit may be readily transformed into the equivalent parallel *RC* circuit, where:

$$R_{par} = \frac{1 + (R_s + R_g + 1/5g_m)^2\omega^2C_1^2}{(R_s + R_g + 1/5g_m)\omega^2C_1^2} \quad . \quad . \quad . \quad (6)$$

Thus R_{par} varies as g_m changes over the gate-voltage cycle. By differentiating R_{par} with respect to g_m, it can be shown that R_{par} has a minimum value when $R_s + R_g + 1/5g_m = 1/\omega C_1$ and that this minimum value is $2/\omega C_1$.

For the MOS power transistor shown in fig. 6, $C_1 = 130$ pF. This means that, at a frequency of 30 MHz, the parallel input resistance R_{par} will pass through a minimum value of 81 Ω whenever g_m, which oscillates between zero and 150 mA/V, passes through the value of ≈ 5 mA/V. This occurs twice in every period. This example shows that large MOS-transistor structures can have quite small effective input resistances at high frequencies, and the resulting low gain may be the most important limitation.

In a linear amplifier, the fact that the input resistance

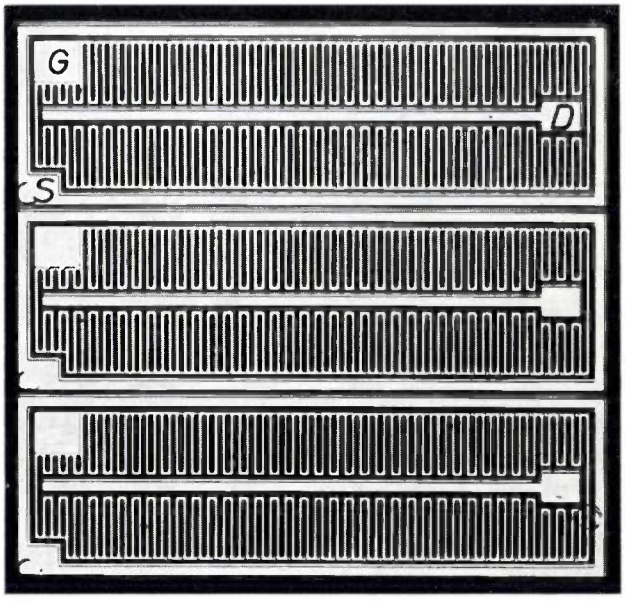


Fig. 6. MOS power transistor consisting of three identical units to be wired in parallel. *S* source. *G* gate. *D* drain. The total channel width is 12.2 cm. Both source and drain fingers are unmetallized (magnification about 30×).

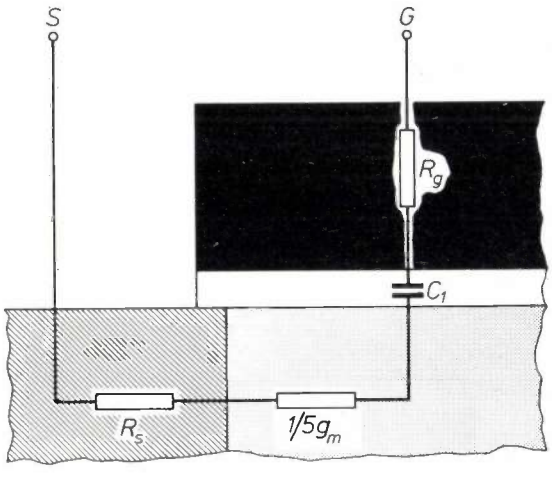


Fig. 7. Equivalent circuit for the input of a MOS transistor. R_s and R_g represent the internal source and gate resistances respectively, $1/5g_m$ the distributed channel resistance — g_m is the transconductance — and C_1 the distributed gate capacity.

[7] More about the time constant τ can be found in the articles by P. A. H. Hart and F. M. Klaassen and by R. J. Nienhuis in this issue, pages 216 and 259.

changes during the input voltage swing results in distortion of the input signal, particularly if the gate capacitance and frequency are large, so that $2/\omega C_1$ represents a very small value.

Thermal behaviour

The effect of temperature on the drain current of a MOS transistor is determined by two factors: a drift in threshold voltage and a mobility fall with increasing temperature [8]. These changes are usually in opposite directions and so a biasing point of zero temperature coefficient exists. However, at high currents the falling mobility predominates, and an increase in temperature tends to produce a fall in current. Experimental results confirming this are presented in the next section.

This behaviour has extremely important consequences for practical devices. A large MOS transistor is thermally stable for fluctuations of current either over the surface of the device or with time. The stability over the surface ensures that the distribution of current over a device will be good without the introduction of stabilizing resistances, which is often necessary in bipolar power transistors. The stability of a MOS transistor for time changes of current means that thermal runaway does not occur. Consequently, a MOS transistor is potentially a rugged device. The phenomenon of second breakdown [9] would for the same reason not be expected in the MOS power transistor and indeed it has not been observed except for the special case considered in the section on voltage limitations [10].

Experimental results

MOS power transistors have been fabricated using the geometry of fig. 5. As this was designed before the information about differences in voltage behaviour between *N*- and *P*-channel MOS transistors was available, it was intended for use as an *N*-channel MOS transistor and had a channel length of 16 μm. The *N*-channel devices made with this geometry did not produce the expected power because of their low voltage limitation. Because of this, *P*-channel devices were made. The availability of the photomasks led to the decision to maintain the same geometry although the optimum channel length in this case would be ≈ 10 μm.

The power output to be expected from this device in a linear amplifier can be simply calculated and compared with experiment. For linear-circuit operation, the device must be operating substantially within the square-law part of its I_d-V_{gs} characteristic, and so the load line must not cross the saturation "knee" of the I_d-V_{ds} characteristic (see fig. 2 on p. 252). For a single device, operating in class B, the peak envelope power P_{pe} which can be transferred to the load is given by:

$$P_{pe} = \tfrac{1}{2}(I_{d\,max}/\sqrt{2})(V_{d\,max}/\sqrt{2}), \quad \ldots \quad (7)$$

where $V_{d\,max}$ is the maximum peak voltage to which the drain swings about the supply voltage. With the help of eqs. (1), (2) and (3) this can be reduced to:

$$P_{pe} = \mu C_{ox} w V_{gs}'^2 (V_{ds\,br} - V_{ds\,sat})/16\,l. \quad \ldots \quad (8)$$

For the power transistor under consideration $\mu \approx 150$ cm²/Vs, $C_{ox} = 1.8 \times 10^{-8}$ F/cm², $w = 4.2$ cm and $l = 16$ μm. The minimum value of the drain breakdown voltage $V_{ds\,br}$ was 90 V but the device was only driven to 85 V in this experiment. The maximum gate voltage was 14 V, and $V_{ds\,sat}$ was 15 V. Hence the peak envelope power to be expected is 6.1 W.

Pairs of these devices were mounted on the same header and wired in parallel, and assessed in the frequency range 3 to 30 MHz [11]. A two-tone test signal was used to measure the level of intermodulation products, a test of the circuit linearity. A peak envelope power of 11.5 W was obtained with an intermodulation-product level of −30 dB. Hence each chip was giving 5.75 W, or 95% of the expected power.

The yields of working devices obtained from this design indicated that a future device with a considerably wider channel could still maintain a practicable yield. In addition the experiments on breakdown voltage referred to above suggested that the highest power output would be achieved from a *P*-channel MOS transistor with a channel length of about 10 μm. The structure of fig. 6 was therefore fabricated, with a channel length of 9 μm and width of 12.2 cm, and assessed under similar conditions to those described above. Insertion of the values into equation (8) yields an expected output power of 31.5 W. The measured peak envelope output power from these devices was in fact 30 W.

One drawback of these devices was that in order to avoid the difficulties of neutralization, which would be needed because of their rather large feedback capacitances, it was found necessary to use an untuned input circuit. In addition, to avoid distortion of the input signal, the gate had to be damped with a 50 Ω resistor, which resulted in a reduced gain for the amplifier.

Measurements of the variation of drain current with temperature of a device of the type shown in fig. 5 have been made and are shown in *fig. 8*. The temperature of the transistor was controlled by an oil bath, and the appropriate gate voltage was applied in the form of pulses with a repetition frequency such that the current through the transistor did not appreciably alter its surface temperature. These curves show quite clearly that the drain current falls with increasing temperature at high currents.

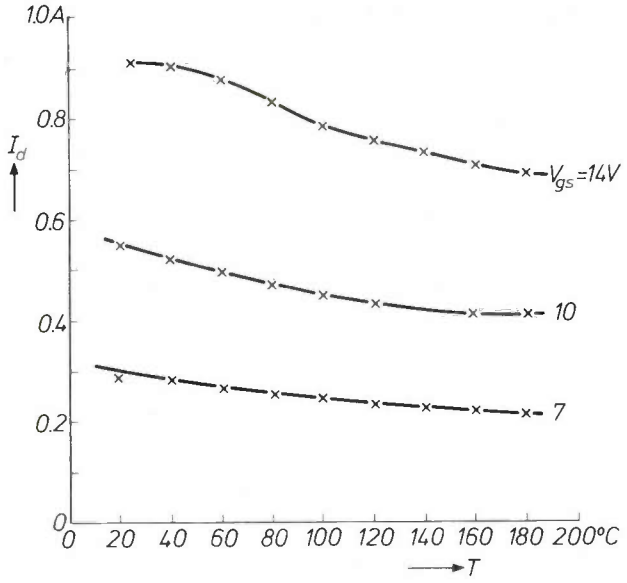


Fig. 8. The drain current I_d for the MOS power transistor of fig. 5, plotted against the temperature T of the transistor. The curves show a negative temperature coefficient.

The distribution of temperature over the surface of an interdigitated MOS power transistor has been measured using an infra-red microscope. Typical results, in the form of isotherms, are shown in *fig. 9*.

Methods of improving MOS power transistor performance

It has already been stated that the feedback capacitance of a MOS power transistor may be relatively large, and that the drain breakdown voltage depends on the gate voltage. Both these difficulties can be overcome by using an offset-gate structure, in which the gate does not overlap the drain but stops a few microns

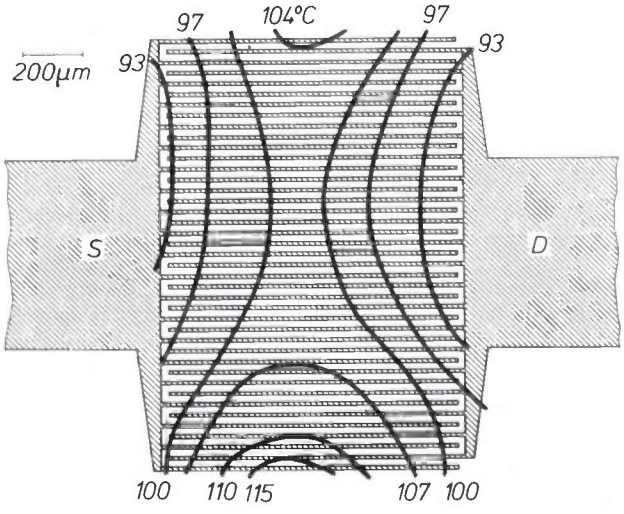


Fig. 9. The temperature distribution across the surface of a MOS power transistor is nearly uniform. *S* source. *D* drain.

short of it (*fig. 10*). Devices of this kind have been made and the drain breakdown voltage (on 8 Ωcm material) was increased from 90 V to 125 V, and was independent of gate voltage, while the feedback capacitance was 3.5 pF, compared to 50 pF for the full-gate device. In general, however, such a device will suffer from having a higher saturation voltage than a full-gate type, because of the series resistance introduced into the channel. To avoid this effect, the device must be a depletion type and must have a threshold voltage

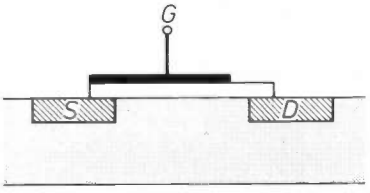


Fig. 10. Schematic cross-section of a MOS transistor provided with an offset gate for reducing the feedback capacity. *S* source *G* gate. *D* drain.

such that the conductivity of the uncontrolled part of the channel is greater than that of the controlled part at all times. In practice this means a threshold voltage of around 15 V on the depletion side. It is very difficult to achieve a stable threshold voltage of this value with conventional MOS technology.

For producing such a "heavy depletion" device two lines of attack are available. One is to use a dielectric which gives a stable turn-on voltage of the required value. This has been done successfully with *N*-channel MOS transistors using a double layer of silicon dioxide and silicon nitride, but these devices suffer from the low breakdown voltage discussed in the section on voltage limitations, and for *P*-channel transistors where a large positive threshold voltage is required the problem is much more difficult. The other approach is to produce a permanent built-in channel, so that the device no longer relies on inversion of the surface layer, but simply depletes or enhances a thin skin of silicon of opposite type from the bulk. Such a layer can be

[8] H. C. de Graaff and J. A. van Nielen, Electronics Letters **3**, 195-196, 1967.

[9] Second breakdown is a particular difficulty with bipolar power transistors. It is found if the collector current vs. voltage characteristic is continued beyond the point of first or Zener breakdown and shows up as a sudden jump of the curve to a lower voltage value — not unlike the breakdown phenomena shown by fig. 4*b*. Some authors believe that it is associated with local thermal instability. See H. A. Schafft, Second breakdown — a comprehensive review, Proc. IEEE **55**, 1272-1288, 1967.

[10] It has also been reported by T. Asakawa and N. Tsubouchi, Second breakdown in MOS transistors, IEEE Trans. **ED-13**, 811-812, 1966.

[11] The high-frequency measurements were carried out by J. Ling of the Mullard Central Application Laboratories.

produced by epitaxy, diffusion, or ion implantation. The device characteristics are critically dependent on the resistivity and thickness of the layer [12].

It is also possible to overcome the feedback and voltage problems by using a tetrode structure with two independent gates. This would also remove the difficulty of providing a "heavy depletion" device. MOS tetrodes have been made successfully for small-signal operation [13], but for a power device, simplicity of geometry and economy of area are so important that the offset-gate structure seems to be more promising at the present time.

Summary. The MOS transistor has several attractive features as a high-frequency power amplifier. An important one is its negative temperature coefficient, which gives a nearly uniform temperature distribution and freedom from thermal runaway and second breakdown. A high output power requires a high current-carrying capacity and a high drain-junction breakdown voltage. *P*-channel MOS transistors have a higher breakdown voltage than *N*-channel MOS transistors; a detailed theory is not available. A high current requires a thin oxide layer and a short, wide channel; high-frequency performance requires a short channel. Oxide thickness and channel length are limited by breakdown to 0.2 μm and 10 μm respectively. Experimental MOS power transistors with a channel width of 4.2 cm and 12.2 cm have been made by giving drain and source an interdigitated structure. In the HF band (3-30 MHz) the measured peak envelope power was 5.75 W with the 4.2 cm channel width and 30 W at 12.2 cm, i.e. nearly equal to the calculated value. Where a small feedback capacitance is required the offset-gate approach is the most promising one for MOS power transistors.

[12] An example of such a device is discussed in the article by J. A. van Nielen, M. J. J. Theunissen and J. A. Appels in this issue, page 271.

[13] T. Okumura, The MOS tetrode, Philips tech. Rev. 30, 134-141, 1969 (No. 5).
R. J. Nienhuis, A MOS tetrode for the UHF band with a channel 1.5 μm long; this issue, page 259.

A MOS tetrode for the UHF band with a channel 1.5 μm long

R. J. Nienhuis

In the MOS transistor the gate metallization partly overlaps the diffused region of the drain, owing to unavoidable variations of dimension. This gives rise to a capacitance between gate and drain which is generally of the order of 1 pF. Because of this capacitance there is negative feedback from the drain to the gate, which increases with the frequency of the signal. This limits the application of MOS transistors to frequencies below about 100 MHz.

In the MOS tetrode [1] the feedback capacitance is much smaller. This device, which consists of a series configuration of two MOS transistors, can therefore be used for higher frequencies. What will the upper limit of frequency be for an optimal design? An analysis of the characteristics of the MOS tetrode indicates that for good high-frequency performance the channel for the first transistor should be as short as possible, but this introduces manufacturing problems. However, by using a special process we have succeeded in making well defined channels with a length of no more than 1.5 microns, and the result is a MOS tetrode which can still give a gain of about 5 dB at 1000 MHz. This shows that the MOS tetrode could be used in the UHF band [2], for example in the tuner of a television set. In this application, as will be shown later, the MOS tetrode also provides an attractive means of gain control.

Some characteristics of the MOS tetrode

The MOS tetrode is a cascode arrangement of two MOS transistors (also called MOS triodes). In this arrangement the two triodes are connected in series, the first with its source earthed and the second with its gate earthed for a.c. voltages (see *fig. 1a*, which gives a diagram of a MOS tetrode with an *N*-type channel). The cascode circuit is made as a single unit, one diffused region serving at the same time as the drain for the first triode and as the source for the second. The diffused region is called the "island" and has no contacts. The signal voltage v_{isl} for the island is approximately equal to the output voltage v_d at the drain of the tetrode divided by the voltage amplification factor μ_2 of the second triode. This means that the feedback through the feedback capacitance C_{fb1} to the first gate is a factor of μ_2 smaller than in a single transistor. The effective feedback capacitance of the complete tetrode is therefore μ_2 times smaller than for a single transistor. In a simplified equivalent circuit for the MOS tetrode (fig. 1*b*) it is therefore usually permissible to leave it out altogether. These features are illustrated in *fig. 2a*, in which the measured values of C_{fb} for a type BFS 28 MOS tetrode are shown as a function of the d.c. voltage V_{ds} at the drain. At a drain voltage V_{ds} of less than about 5 V the second triode of the tetrode is not saturated and operates as a simple series resistance. The C_{fb} at the drain of the first triode

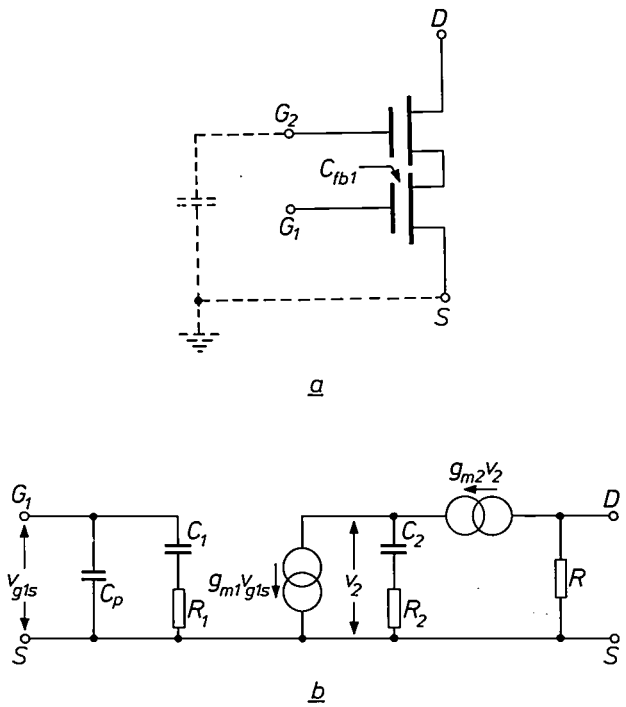


Fig. 1. *a*) Schematic diagram of a MOS tetrode. *S* source. G_1 gate of the first triode. G_2 gate of second triode; this is earthed to a.c. voltages. *D* drain. C_{fb1} feedback capacitance of first triode. *b*) Simplified equivalent circuit of the MOS tetrode. C_p stray input capacitance. C_1 capacitance between first gate and the channel. R_1 part of the channel resistance of the first triode. g_{m1}, g_{m2} transconductances of the first and second triodes. C_2 capacitance between the second gate and the channel. R_2 part of the channel resistance of the second triode. R output resistance of the tetrode.

Ir. R. J. Nienhuis is with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven.

[1] T. Okumura, The MOS tetrode, Philips tech. Rev. **30**, 134-141, 1969 (No. 5).

[2] In accordance with international usage we take the VHF (very high frequency) band to include the frequencies from 30 to 300 MHz, and the UHF (ultra high frequency) band to include the frequencies from 300 to 3000 MHz.

can be measured through this series resistance, and it is found to be about 0.8 pF. If the drain voltage V_{ds} is far enough above 5 V, then the second triode is saturated and has the voltage amplification factor μ_2. The tetrode is now in its characteristic mode of operation and the value of C_{fb} measured at the drain is μ_2 times smaller, in this case 40 times smaller, i.e. 0.02 pF.

The capacitance between the drain of the tetrode and the second gate does not contribute to the feedback as the second gate is earthed to a.c. voltages.

The output resistance R of a MOS tetrode is about μ_2 times greater than that of the single MOS transistor. The drain voltage V_{ds} in the MOS transistor has some effect on the saturation current $I_{d\,sat}$ because the drain voltage affects the length of the conducting channel and hence the transconductance [3]. This also applies to the first triode of the MOS tetrode. Since, however, the voltage variations on the island are only $1/\mu_2$ of those at the drain, the effect referred to is also reduced by a factor of $1/\mu_2$. As a result the output resistance becomes μ_2 times higher, as can be seen from the measured values given in fig. 2*b*. This shows the output resistance R as a function of the drain voltage V_{ds} of the tetrode. As soon as this voltage is high enough to give the tetrode action, the output resistance increases by a factor of about 40.

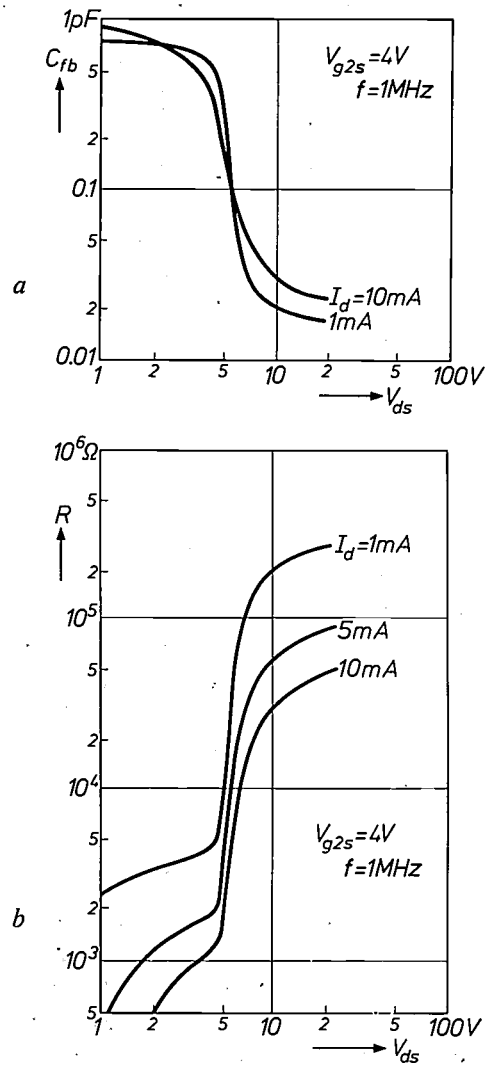


Fig. 2. *a*) Feedback capacitance C_{fb} and *b*) output resistance R of the MOS tetrode BFS 28 as a function of drain voltage V_{ds}. When V_{ds} goes higher than about 5 V the second triode is saturated. The tetrode then enters its characteristic mode of operation and the feedback capacitance C_{fb} measured at the drain becomes a factor of 40 smaller and the output resistance R a factor of 40 higher.

An additional advantage of the MOS tetrode, besides its small feedback capacitance and high internal impedance, is that the gain can be controlled by means of the voltages V_{g2s} at the second gate, since V_{g2s} can be used to bias the first triode to the saturation limit. This limit is reached when the potential of the island $V_{isl\,s}$ is equal to $V_{g1s} - V_{th}$. If both triodes of the tetrode have the same characteristics, then $V_{g2s} - V_{isl\,s} - V_{th} = V_{g1s} - V_{th}$, or $V_{isl\,s} = V_{g2s} - V_{g1s}$ [1], and therefore the saturation limit is reached when $V_{g2s} = 2\,V_{g1s} - V_{th}$. If the first triode is saturated but close to this limit, then V_{g2s} already has an appreciable effect on the current through this triode and hence on its transconductance. The transconductance of the tetrode as a whole is equal to that of the first triode [4] and is therefore affected by V_{g2s} in the same way. This is illustrated by *fig. 3*, which shows that the slope of the I_d-V_{g1s} curves becomes less steep as V_{g2s} decreases. This facility for gain control is used in radio and television receivers for the automatic gain control, which matches the receiver sensitivity to the strength of the incoming signal. If a decoupled resistor R_s is incorporated in the supply lead of the MOS tetrode, as shown in the inset of fig. 3, the d.c. operating point of the MOS tetrode is automatically shifted when a strong signal is received so that not only does the gain decrease but the maximum input voltage that can be amplified without noticeable distortion is increased. This helps to improve the linearity of the MOS tetrode. Because of this, and the absence of odd terms in the square-law characteristic of the MOS transistor and tetrode, there is less cross-modulation than with a bipolar transistor or thermionic valve (*fig. 4*).

Short channel required for high frequencies

When we seek to make the most of the good high-frequency characteristics of the MOS tetrode by means of an optimum design, we find that the design of the first transistor of the tetrode is of particular importance, since this determines the input characteristics and transconductance of the tetrode. The high-frequency behaviour is strongly affected by stray capacitances and

inductances. These cannot be accurately calculated, and all that one can do is to try and keep the stray elements small by keeping the dimensions small. The situation is different for the actual amplification mechanism of the MOS transistor; in this case the relation between characteristics and channel dimensions can be calculated. Let us look at fig. 1*b*, in which most of the stray elements have been omitted. The first transistor is represented by a current source which delivers a current of magnitude g_{m1} times the input voltage v_{g1s}. To find some indication of the usefulness of the MOS transistor at high frequencies we consider the frequency at which the input current i_1 of the first transistor (neglecting the stray capacitance C_p) is equal to its output current $g_{m1}\, v_{g1s}$. Since at this frequency the reactance of C_1 is still a few times greater than R_1, we may write as an approximation $\omega C_1 = g_{m1}$ or $\omega C_1/g_{m1} = 1$.

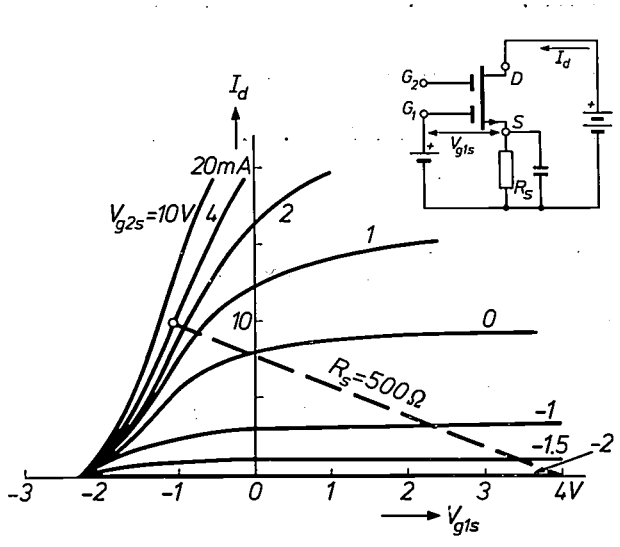


Fig. 3. The I_d-V_{g1s} characteristics of a MOS tetrode with V_{g2s} as parameter. When V_{g2s} decreases the slope of the curve decreases and hence also the amplification of the tetrode. This effect is used for automatic gain control. If a decoupled resistor R_s (*see inset*) is included in the supply lead, the d.c. voltage V_{g1s} between G_1 and S becomes dependent on the current operating point. When a large input signal is received, V_{g2s} becomes lower; the current I_d then decreases and consequently the voltage drop across R_s also decreases. As a result the bias V_{g1s} increases; the operating point I_d, V_{g1s} shifts along the sloping dashed line to the right and the maximum signal excursion (i.e. that can be applied without completely cutting off the MOS transistor) increases. This enables the larger input signal for which the gain was reduced to be handled without excessive distortion.

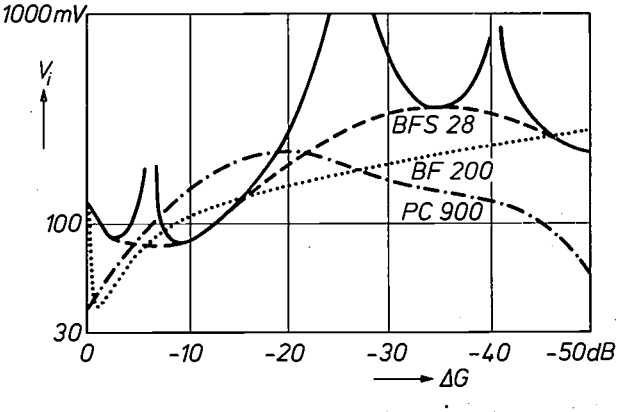


Fig. 4. The magnitude V_1 which an unwanted signal at the input of a receiver may reach before the cross-modulation reaches 1% depends on the operating point of the amplifying device (MOS tetrode, transistor or thermionic valve) in the input circuit of the receiver. This operating point is determined by the automatic gain control, which gives a greater reduction in the gain G the stronger the desired input signal. The results of measurements on the BFS 28 MOS tetrode are better than those obtained with the BF 200 bipolar transistor or the PC 900 triode valve, which are widely used in input stages. The measurements on the PC 900 were made at a supply voltage of 280 V.

We have already encountered the time constant $\tau = C_1/g_{m1}$ in a previous article in this issue [5]. The smaller the value of τ, the better the high-frequency characteristics of the MOS transistor. In order to look into the relation of τ to the channel dimensions, we write:

$$\tau = \frac{C_1}{g_{m1}} = \frac{C_{ox}\, wl}{\sqrt{2\mu C_{ox} I_d w/l}} = \sqrt{\frac{C_{ox} w l^3}{2\mu I_d}} \quad \ldots \quad (1)$$

Here C_{ox} is the capacitance per unit area between the gate and the channel, and l and w are the length and width of the channel. The expression for g_{m1} has been given in a previous article in this issue [3]; it follows from equations (7) and (11) in that article. We see that for a given value of the drain current I_d the time constant τ of the first transistor is proportional to $l^{3/2}$. For a small value of τ the first transistor must therefore have a short channel.

Another factor of importance besides τ is the maximum available power gain G_m, i.e. the power gain available with ideal matching. This is a theoretical quantity that describes the performance of an active linear four-terminal network (or two-port) as an amplifier. It is used in practice to give an indication of the quality of an amplifying device, represented as a two-port [5]. In the case where the feedback capacitance C_{fb} can be neglected, the maximum available power gain is given by the simple expression:

$$G_m = \frac{|Y_{21}|^2}{4\ \mathrm{Re}\ Y_{11}\ \mathrm{Re}\ Y_{22}} \quad \ldots \quad (2)$$

Here Y_{21} is the transfer admittance of the two-port, i.e. the output current divided by the input voltage, and Y_{11} and Y_{22} are the input and output admittance, respectively. In designing the tetrode the denominator of (2) should be kept as small as possible. There is much to be gained from this, particularly by paying attention to Re Y_{11}.

[3] See the article by J. A. van Nielen in this issue, page 209.
[4] See equation (7) of [1].
[5] P. A. H. Hart and F. M. Klaassen, The MOS transistor as a small-signal amplifier; this issue, page 216.

From the equivalent circuit of fig. 1*b* it can be shown that:

$$Y_{11} = j\omega C_p + \frac{\omega^2 C_1^2 R_1 + j\omega C_1}{1 + \omega^2 C_1^2 R_1^2}.$$

At the frequencies at which the MOS tetrode is used, the impedance $1/\omega C_1$ is still several times greater than R_1, so that the denominator of the fractional term does not differ much from unity; we may therefore simplify to:

$$Y_{11} \approx \omega^2 C_1^2 R_1 + j\omega(C_1 + C_p). \quad . \quad . \quad (3)$$

For these frequencies R_1 can be approximated by $R_1 = 0.2/g_{m1}$. This gives Re $Y_{11} = 0.2\omega^2 C_1^2/g_{m1}$. Inserting the expressions given in (1) for C_1 and g_{m1} we come to the conclusion that:

$$\text{Re } Y_{11} \propto w^{3/2}/^{5/2}. \quad . \quad . \quad . \quad . \quad . \quad (4)$$

This again demonstrates how important it is to have the shortest possible channel in the first transistor.

Process for making short channels

The length of the channel of a MOS transistor is equal to the distance between source and drain diffused areas. When a MOS transistor is made by diffusing into the silicon substrate through a hole etched in the oxide layer, the diffusion also spreads out laterally to a distance approximately equal to the diffusion depth. This has the result that the distance between the diffusions is a little smaller than the distance between the etched holes. Also, any variation in the spacing of the two etched holes is matched by a corresponding variation in the spacing of the two diffusions. Such variations always arise, owing to inaccuracies in the etching process, but if a very short channel is to be produced between the diffusions the effect of these variations becomes disproportionately large. We therefore favour a process in which the channel length is not determined by a distance between two contact diffusions, but by the dimensions of a single etched hole. A special process of this kind has been developed, which also has the advantage that the oxide layer formed is thin directly above the channel but thick elsewhere, so that the overlap of the gate electrodes does not give excessive stray capacitances to the source, island and drain. The successive stages in the process are illustrated in *fig. 5* for the case of a MOS tetrode with an *N*-type channel [6]. The special feature of this process is that in step *4* a phosphorus-doped oxide layer is applied from which *N*-type diffusions grow in step *6* with a thickness of only a few tenths of a micron and with accurately defined boundaries. The two boundaries are sufficiently accurately defined to enable a very short channel to be left between them.

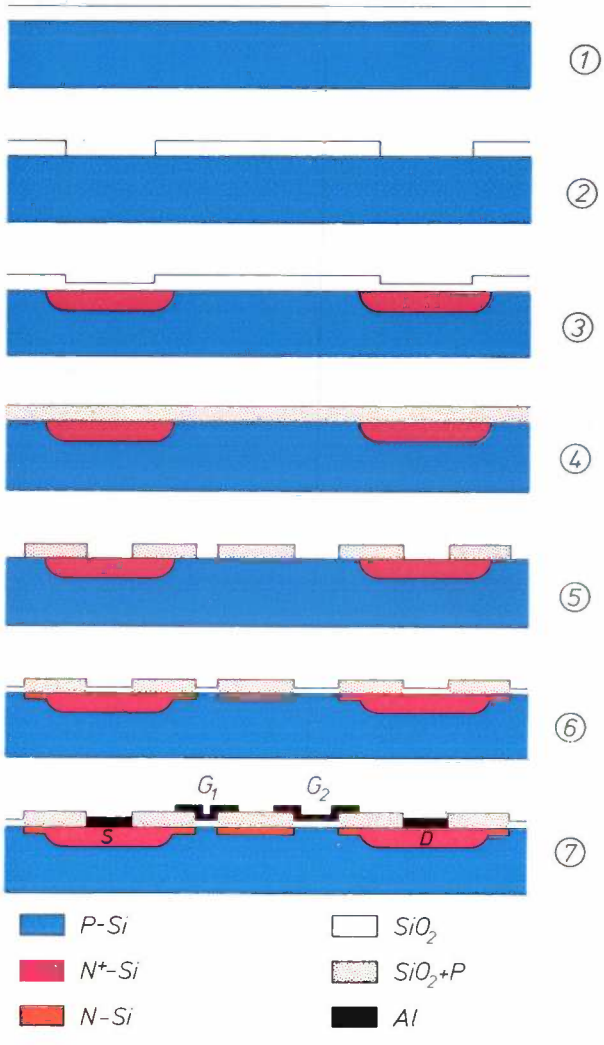
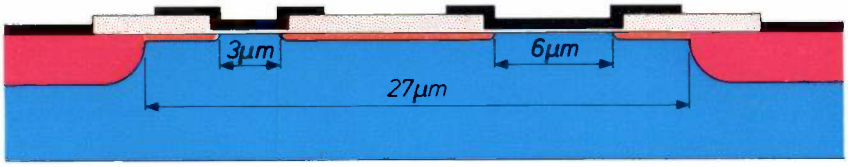


Fig. 5. Stages in the production of a MOS tetrode with a very short channel.
1) Oxidation of the *P*-type silicon substrate.
2) Holes are etched in the oxide layer for source and drain.
3) Source and drain are produced by N^+ diffusions to a depth of about 2.5 μm.
4) All oxide is removed and another oxide layer, doped with phosphorus is applied. Phosphorus is a donor and produces *N*-type silicon when it diffuses into the silicon substrate.
5) Holes are etched in the new oxide layer for source and drain contacts and for the two gates.
6) An *N*-type diffusion only a few tenths of a micron deep is made from the oxide. A thin oxide layer forms in the holes.
7) Aluminium contacts are deposited.

We were able in this way to make channels with a length of no more than 1.5 μm. This would seem to be the lower limit, because at smaller distances the danger of punch-through [3] between island and source would be too great with the maximum drain voltage of 20 V which MOS tetrodes must be able to handle. This is why the channel of the second transistor of the tetrode, whose length has less effect on the high-frequency characteristics, is always made longer; the tetrode can then take a higher drain voltage. Putting the gate electrodes on thick oxide away from the channel area, which is done to avoid excessive stray capacitances, has



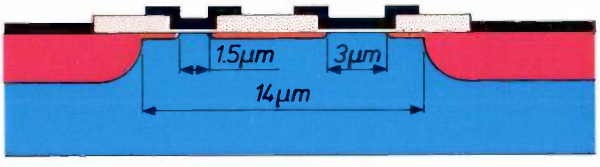


Fig. 6. Cross-section of two MOS tetrodes, drawn to scale, made by the process illustrated in fig. 5. (The thickness of the substrate is not to scale.) Type BFS 28 (*a*) is for the VHF band, the other type (*b*), with a channel only 1.5 μm long, is for the UHF band.

the incidental advantage that the location of the metallization is not too critical.

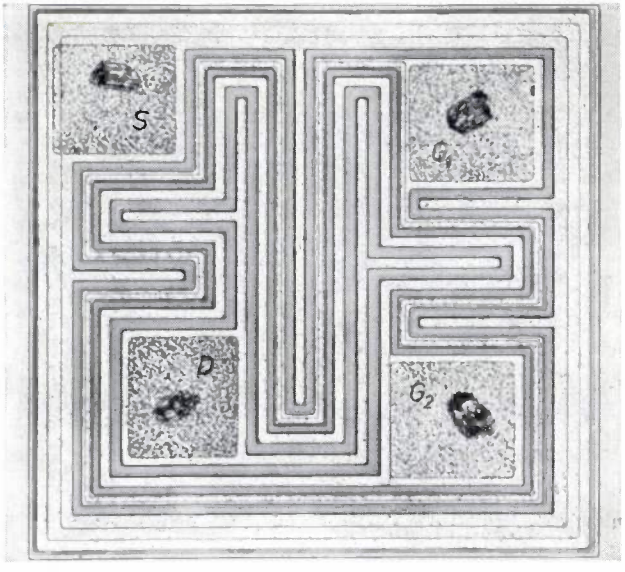
Two types of MOS tetrode have been made by the process described, one with a channel of 3 μm and intended for the VHF band, and the other with a 1.5 μm channel, for the UHF band. The first type is now in production under the type designation BFS 28. *Fig. 6* shows cross-sections of both types drawn to scale, except for the thickness of the substrate. *Fig. 7* shows photomicrographs of both tetrodes produced on a surface 0.5×0.5 mm. The configuration is such that a maximum channel width w is obtained on the area not affected by the four contact areas at the corners. The photographs show clearly the difference in channel length between the first and second transistors of the tetrode and the overlap of the gate metallization [7].

Comparison of the characteristics of the two types

The dimensions of the channel in the first triode of both types are:

BFS 28: $l = 3$ μm, $w = 2.4$ mm,
UHF tetrode: $l = 1.5$ μm, $w = 3.7$ mm.

If we use these dimensions to calculate the ratio of the time constants τ of the two tetrodes, assuming the same operating point and otherwise identical parameters, we find from equation (1) that τ for the BFS 28 is 2.16 times greater than for the UHF type. This result is not easily verified by measurements, because of the marked and sometimes dominant influence of all kinds of stray effects. At the high frequencies we are concerned with there are also stray effects in the metal encapsulation



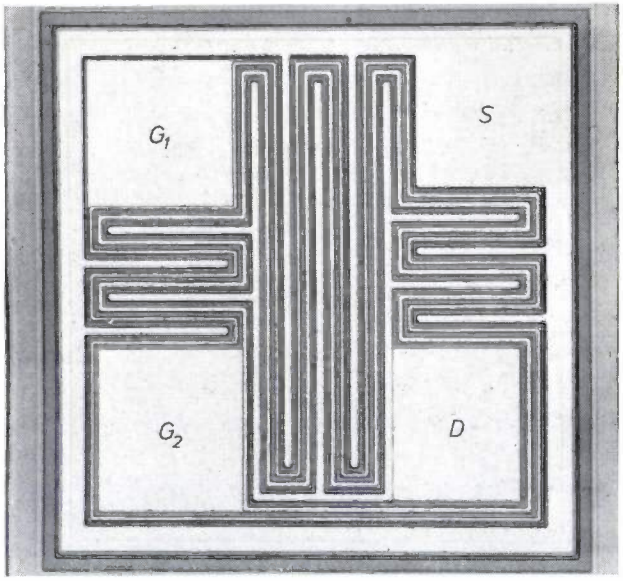


Fig. 7. *Above:* The VHF tetrode type BFS 28 (channel length $l = 3$ μm, channel width $w = 2.4$ mm). *Below:* The UHF tetrode ($l = 1.5$ μm, $w = 3.7$ mm). The light surfaces are the aluminium tracks and terminal areas; the corresponding electrodes are indicated on the terminal areas. Both MOS tetrodes are produced on a surface 0.5×0.5 mm. The indentations that can be seen in the terminal areas of the BFS 28 were made by test probes during checking procedures.

[6] Another process is ion implantation, in which the shallow doped regions are produced by bombarding the substrate with fast donor or acceptor ions. The already applied gate metal acts as a mask in this process and ensures sharp definition of the boundary of the channel. See the article by J. M. Shannon in this issue, page 267.

[7] The photomasks for the UHF tetrode could only be made by pushing the step-and-repeat processing camera to the limits of its performance. See F. T. Klostermann, Philips tech. Rev. **30**, 57, 1969 (No. 3), where a detail of the UHF tetrode is shown on page 69.

(TO-72). This is illustrated in *fig. 8*, where the complex transfer admittance Y_{21} measured for both types is plotted with the frequency as parameter. The remarkable increase in $|Y_{21}|$ with frequency is probably due to stray effects in the encapsulation, which make it difficult to measure the characteristics of the device itself. The UHF tetrode gives less phase shift than the BFS 28; this indicates that it does have better high-frequency characteristics.

In equation (1) we see that for the same drain current I_d the ratio of the transconductances of two MOS transistors is equal to the ratio of the values of $w^{1/2}l^{-1/2}$. Calculating this ratio for the two types, we find that the transconductance of the UHF tetrode should be 1.75 times greater than that of the BFS 28. To a first approximation the quantity Y_{21} is equal to the transconductance g_{m1} of the first transistor of the tetrode, but its value is also affected by stray effects in the MOS tetrode. Nevertheless, we find that fig. 8 shows a ratio of about 1.75.

The ratio of the values of Re Y_{11} can also be calculated from the dimensions (see equation 4). We find that Re Y_{11} for the BFS 28 should be 2.94 times the value for the UHF tetrode. Measurements of this quantity as a function of frequency are given in *fig. 9*, and do in fact show approximately this ratio, particularly at lower frequencies. The effect of stray elements is more easily seen here; the chief stray element is the capacitance C_p between gate and source (see fig. 1*b*) and this does not contribute to Re Y_{11}. In this case, therefore, calculations and measurements seem to agree. The curve of Re Y_{11} as a function of frequency corresponds approximately to equation (3).

Fig. 9 also shows the results of measurements of the output conductance Re Y_{22}. The variation of this quantity with frequency differs from one tetrode to another. There is as yet no generally accepted theory that can explain these differences.

Part of the total output admittance Y_{22} is given by the series arrangement of the depletion-layer capacitance C_d around the drain and the substrate resistance R_b. If we call this contribution Y_{22}', we may write:

$$Y_{22}' = \frac{\omega^2 C_d^2 R_b + j\omega C_d}{1 + (\omega C_d R_b)^2}. \quad \ldots \quad (5)$$

At those frequencies where $(\omega C_d R_b)^2$ is very much less than 1, Re Y_{22}' is proportional to ω^2. Experimentally, the frequency dependence found for Re Y_{22} invariably corresponds to an exponent of less than 2, which indicates that other mechanisms come into play. For example, the voltage across R_b, which is due to the output signal at the drain, also modulates the channel and hence the output signal.

If, finally, we want to calculate the maximum available gain G_m of the two tetrodes from the measured values of $|Y_{21}|$, Re Y_{11} and Re Y_{22}, using equation (2), we encounter the difficulty that the measurements of Y_{21} are no more than provisional, as appears from the anomalous form of the curves in fig. 8. Assuming that $|Y_{21}|$ is constant, at 13 mA/V for the BFS 28 and 20 mA/V for the UHF tetrode, the calculations give the gain values shown in *fig. 10* for three different fre-

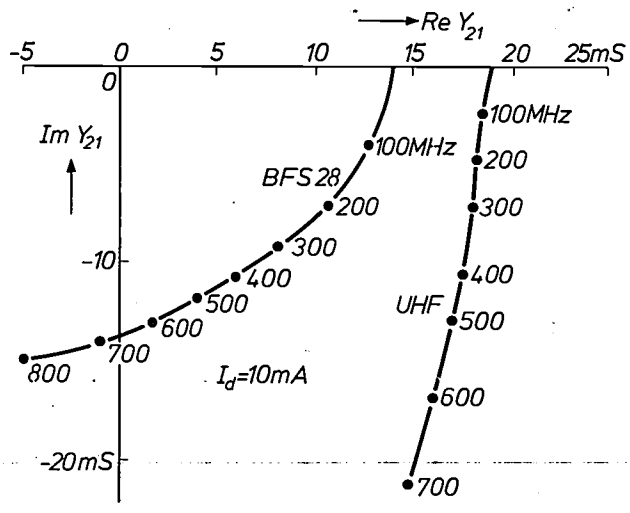


Fig. 8. The transfer admittance Y_{21} of the MOS tetrodes is a complex quantity with a phase angle that increases with frequency. The increase in $|Y_{21}|$ at high frequencies is probably due to stray elements that arise when the standard metal encapsulation of the MOS tetrodes is used. The unit S (siemens) is equivalent to A/V.

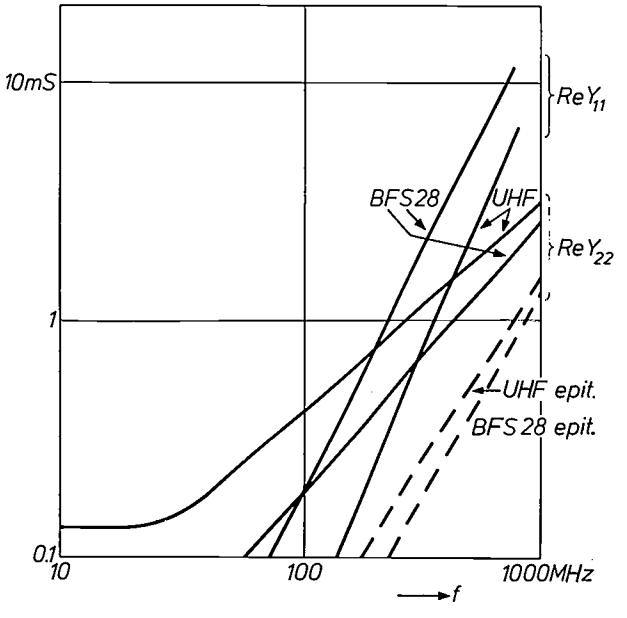


Fig. 9. Re Y_{11} of both MOS tetrodes increases approximately as the square of the frequency; the value of Re Y_{11} for the UHF tetrode is two to three times lower than for the BFS 28. The variation of Re Y_{22} with frequency is approximately linear for both types. A better approximation to a square law is obtained with epitaxial versions of both types on a low-resistance substrate; in these versions, moreover, Re Y_{22} is smaller (dashed lines).

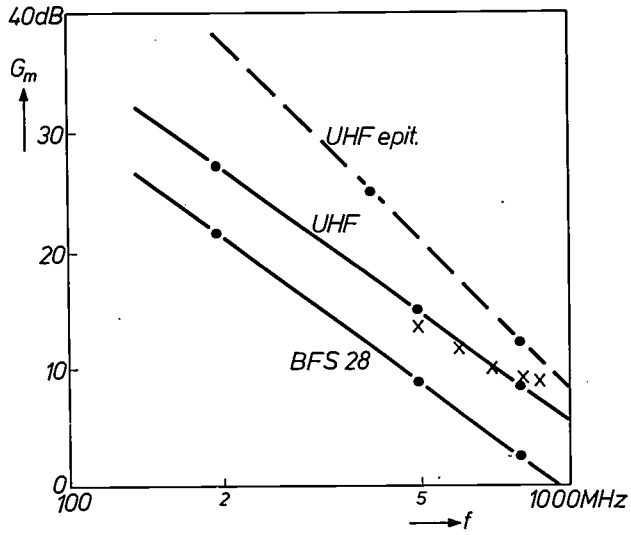


Fig. 10. The maximum available power gain G_m of the MOS tetrodes calculated for three frequencies, assuming Y_{21} to have a constant value of 13 mA/V for the BFS 28 and 20 mA/V for the UHF tetrode. These approximate to the values from fig. 8. Direct measurements of the gain of the UHF tetrode (*crosses*), also adjusted to 10 mA, show good agreement with the calculations. The dashed line gives the calculated gain for the epitaxial type of UHF tetrode on a low-resistance substrate.

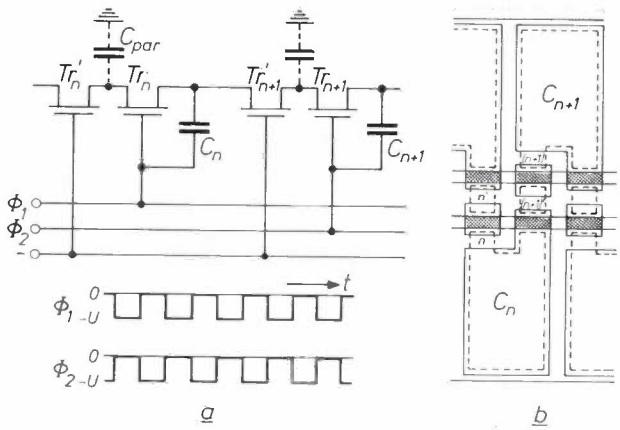
quencies. The values calculated for the UHF type are about 6 dB higher than those for the BFS 28. Measurements made for a UHF tetrode incorporated in a UHF tuner are indicated by crosses in fig. 10; it can be seen that the calculation gives a fairly good description of the characteristics of the tetrode in spite of the reservation noted earlier, and that this type can be used as an amplifying device up to frequencies in the region of 1 GHz.

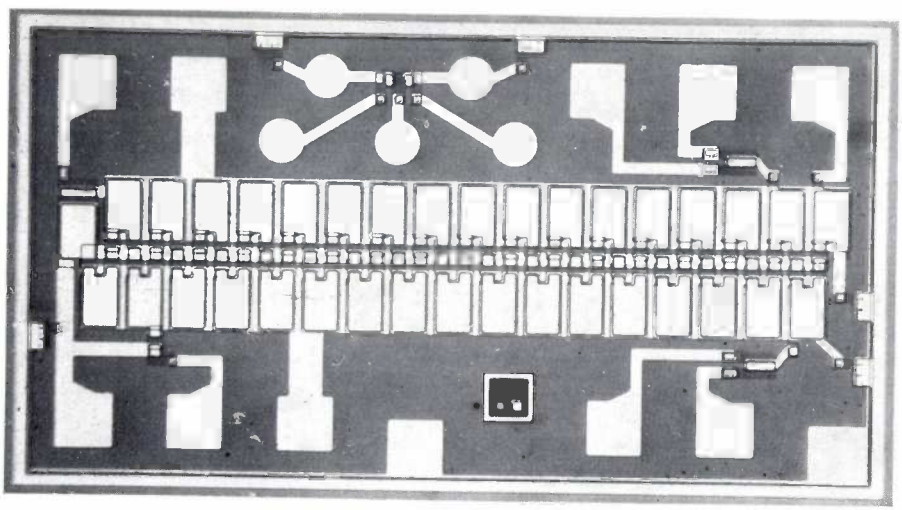
We have referred above to the effect of the substrate resistance R_b on Re Y_{22}. It is desirable to keep R_b as small as possible, both in order to minimize Re Y_{22} (see equation 5) and to minimize feedback to the channel via the substrate. In order to obtain a lower value of R_b, experimental versions of both tetrodes were made in a shallow P-type layer grown epitaxially on a heavily doped P^+ substrate. These epitaxial types are found to have a lower output conductance Re Y_{22} (fig. 9), whose frequency variation moreover approximates more closely to an ω^2 curve. The input conductance Re Y_{11} in the epitaxial types is about 20% higher than in those on homogeneous material, and the transconductance has approximately the same value. A calculation of the power gain G_m of the epitaxial-type UHF tetrode is also given in fig. 10. This shows that even higher gains are obtainable in this way.

Summary. A MOS tetrode is a cascode circuit of two MOS transistors. It has a smaller feedback capacitance than a single MOS transistor and can therefore be used as an amplifier up to frequencies in the UHF band (above 300 MHz). Calculations show that the length of the channel in the first transistor of the tetrode must be made as small as possible to achieve maximum gain at high frequencies. Short channels of well-defined length are obtained by producing shallow diffused regions with limited sideways spread in the silicon substrate from a doped oxide layer. Two tetrodes have been made by this process, one with a channel length of 3 μm (type BFS 28) and the other with a channel length of 1.5 μm. The latter type, available only in a laboratory version, has about 6 dB more gain than the BFS 28, and still gives about 5 dB of gain at 1 GHz.

Integrated bucket-brigade delay line using MOS tetrodes

An advantage of the bucket-brigade delay line [1] using MOS transistors, as compared with the bipolar version, is that there is no attenuation of the signal on account of gate currents in the transistors. The MOS version can therefore have a large number of stages in cascade without having to put amplifying stages between them. With such large numbers, however, another source of distortion becomes significant. This is a small feedback of the drain voltage to the source in the individual MOS transistors that form the stages of a MOS bucket-brigade delay line; this feedback increases as the internal resistance of the MOS transistor diminishes. As a result the transferred signal sample affects the residual voltage of the source, and hence the reference voltage for the next signal sample. The effect is that a residue of the signal sample is added to the next one, causing attenuation at high signal frequencies. The residue per stage in a MOS delay line is about 1/1000 of the signal sample, which limits the number of stages to between 100 and 200. For many applications this is not sufficient; a delay of 100 ms at a bandwidth of 5 kHz, as used for artificial reverberation, requires no fewer than 2000 stages.





The feedback can be reduced by adding to each stage a second MOS transistor biased so that it always operates in saturation (Tr_n' and Tr_{n+1}' in fig. *a*; all the transistors are of the *P*-channel type). These extra transistors together with the original ones form tetrodes; since the internal resistance of such a tetrode is much greater than that of a single transistor [2], smaller residual values may now be expected. However, there is a limitation to this improvement because of the presence of a parasitic capacitance C_{par}, which retains a part of the charge when a signal is transferred. Nevertheless, if this capacitance is kept low, the residue can be reduced by a factor of between 10 and 20 compared with the residue in the original circuit, making it possible to connect several thousand stages in cascade.

The photograph shows a bucket-brigade delay line of this type, with 32 stages and a chip size of 0.95 × 1.65 mm (maximum shift frequency 100 kHz). A detailed sketch of two successive stages (fig. *b*) shows the storage capacitors and the channels of the different transistors [3]. The aluminium strips that form the storage capacitors and the gates of the extra transistors can be seen in the photograph as horizontal white tracks; the boundaries of the various regions in these tracks show up as thin dark lines (because of the differences in oxide thickness). The chip also contains a few transistors for use as input and output circuits.

F. L. J. Sangster

[1] F. L. J. Sangster, The "bucket-brigade delay line", a shift register for analogue signals, Philips tech. Rev. **31**, 97-110, 1970 (No. 4).

[2] See the article by R. J. Nienhuis in this issue, page 259.

[3] This drawing follows the same conventions as fig. 9*b* on page 283 of this issue.

F. L. J. Sangster is with Philips Research Laboratories, Eindhoven.

Ion-implanted high-frequency MOS transistors

J. M. Shannon

The conventional method of doping a semiconductor is to diffuse the dopant into the semiconductor lattice at high temperature. In recent years, however, considerable interest has arisen in a new doping method which uses energetic ions of the dopant. The required dopant ions are accelerated in an electric field before impinging on the semiconductor target. This ion-implantation method is attractive because the concentration of dopant atoms in an implanted layer can be controlled accurately down to the lowest doping levels and doping can be carried out at low temperatures with device metallization already in place. These two features of ion implantation enable MOS transistors to be made with a better high-frequency performance than those manufactured by conventional diffusion methods.

For high-frequency performance, a MOS transistor must have a narrow source-drain separation to give a high cut-off frequency, and the parasitic capacitance between the drain and the substrate, which shunts the output terminals, must be small to minimize loss of power gain. Furthermore, if the full gain of the device is to be used without a neutralizing stage, the device must have a small feedback capacitance C_{fb}. The drain capacitance of a MOS transistor depends upon its area and the width of the drain-to-substrate depletion layer, while the feedback capacitance is determined by the gate-drain overlap. Gates on MOS transistors made conventionally by diffusion have to be defined photolithographically over the gap between the source and drain diffusions and an overlap of typically 2-3 μm occurs owing to the tolerances in the photolithographic process.

A cross-section through a *P*-channel MOS transistor having a low feedback capacitance is shown in *fig. 1*. The transistor is made by defining a metal gate between two widely spaced P^+ diffused contact regions in silicon and then bombarding the device with acceptor ions having sufficient energy to penetrate the gate oxide and implant P^+ regions below the oxide. The ions do not however have enough energy to penetrate the metal gate and consequently the P^+ contact regions are extended up to a position directly below the gate. In this way the gate is automatically registered over the source-drain gap. A microsection through

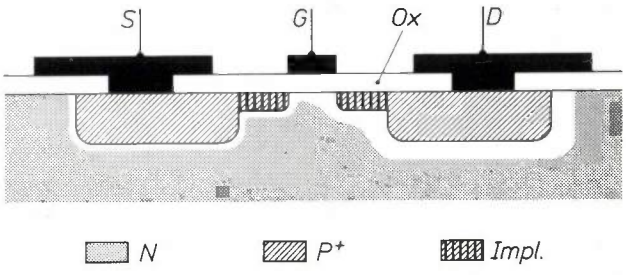


Fig. 1. Schematic cross-section of an autoregistered MOS transistor on an *N*-type silicon substrate. The hatched contact regions of the source and drain are produced by diffusion, and the cross-hatched parts by implantation. *Ox* is the oxide layer, and *S*, *D*, *G* are the metallic contacts to source, drain and gate.

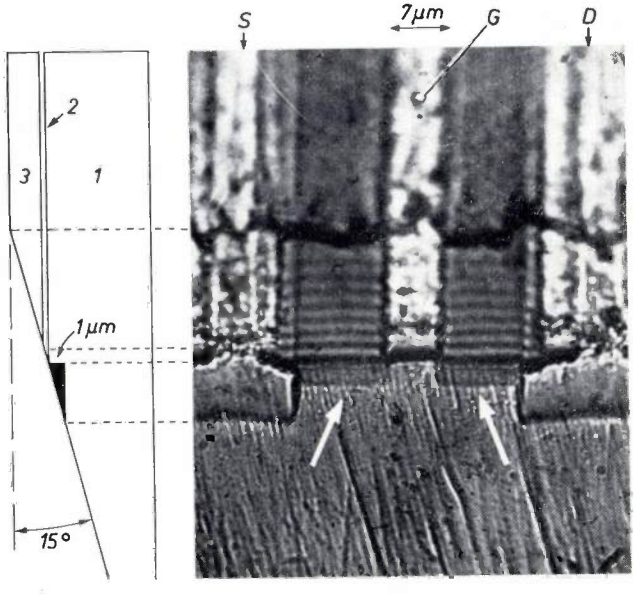


Fig. 2. A 15° microsection through an "autoregistration" MOS transistor (see schematic cross-section at the left) made by implanting 6×10^{15} boron ions/cm² at 60 keV through an 0.12 μm oxide. The implanted P^+ regions (white arrows) extend the diffused contact regions and automatically register the gate over the source-drain gap. In the cross-section *1* is silicon, *2* aluminium and *3* a protective top-layer.

a *P*-channel MOS transistor made using this "autoregistration" procedure is shown in *fig. 2*. Following implantation the device has to be annealed to remove radiation damage and to encourage implanted atoms into substitutional lattice sites where they are electrically active.

P-channel and *N*-channel devices have been made using boron and phosphorus ions respectively [1].

J. M. Shannon, B.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] The boron and phosphorus implants were done by the United Kingdom Atomic Energy Authority, Harwell, England, under the supervision of J. H. Freeman and J. Stephen.

With the exception of nitrogen these are the lightest acceptor and donor impurities in silicon and thus require the smallest energies to penetrate the gate oxide and give a reasonable junction depth. The distribution of atoms implanted into a substrate is approximately gaussian. The mean range of boron ions lies just below the silicon/silicon-dioxide interface when implanting at 35 keV through a 0.12 μm oxide, while phosphorus, being about three times as heavy as boron, requires an energy of 100 keV to give a similar profile.

The sheet resistance of implanted layers decreases with increasing annealing temperature as damage anneals out and more of the implanted atoms become active. When aluminium is used as the gate metal, 500 °C is the highest annealing temperature which can be used before diffusion of aluminium into the gate oxide becomes significant. The sheet resistances of layers implanted at room temperature through oxide used for "autoregistration" purposes are typically 2 kΩ/□ and 0.6 kΩ/□ for boron and phosphorus layers respectively after annealing at 500 °C. The junctions are formed approximately 0.3 μm below the silicon-dioxide interface. The implanted regions add resistance in series with the channel of the device, but the implanted regions only need be a few microns wide and consequently the additional resistance has a negligible effect on the mutual conductance of the device.

There is no evidence that implantation through the oxide surrounding the gate during the autoregistration stage increases the number of gate shorts, and stability tests at high temperatures on MOS transistors stabilized with phosphorus glass indicate that the stability of the gate oxide of autoregistered transistors is not significantly different from the stability of oxides on conventionally made devices.

The low feedback capacitance of autoregistered devices (gate-drain overlaps are typically 0.25 μm) enables them to be used at high frequencies without neutralization. For example *N*-channel autoregistered MOS transistors on 1.5 Ωcm material with 3 μm channel lengths have been made [2] with a calculated maximum stable power gain G_{ms} of 8 dB at 1 GHz. As the maximum frequency of oscillation f_{max} of these devices was 800 MHz, all the available power gain could be used without a neutralizing stage [3].

One way to increase the frequency capability even further is to reduce the capacitance of the drain depletion layer, which shunts the output of the transistor when it is operated in the common source-substrate mode.

An *N*-channel MOS transistor structure designed to reduce the parasitic drain capacitance while maintaining a low output conductance is shown in *fig. 3*. The bulk of the drain junction area lies in a high-resistivity epitaxial layer while the channel is located within a more highly doped implanted layer. The drain depletion layer will be wide in the high-resistivity epitaxial layer and thus the drain capacitance will be small. As well as extending down into the substrate, the drain depletion layer also extends sideways towards the source, and if the drain voltage is large enough the depletion layer will extend the whole way across and the current between source and drain will cease to be controlled by the gate voltage [4]. This punch-through condition will occur at low drain voltages when using narrow source-drain separations and high-resistivity substrates. When locating the channel in an implanted layer which is deeper than the junction formed by the autoregistration stage, the width of the depletion region in the channel region will be much narrower and the punch-through voltage is increased to a value determined by the separation of the diffused contact regions.

The device is made on an epitaxial layer so that the drain-capacitance charging current finds a low-resistance path through the substrate to the source connection, thereby minimizing losses [5].

One of the terms that contribute to the output conductance Re Y_{22} of the device is related to the incremental change in channel length with drain voltage [4]. With an implanted layer the depletion layer width, and hence the pinched-off channel length, will vary slowly with drain voltage and the device will have a low output conductance as well as a low drain capacitance.

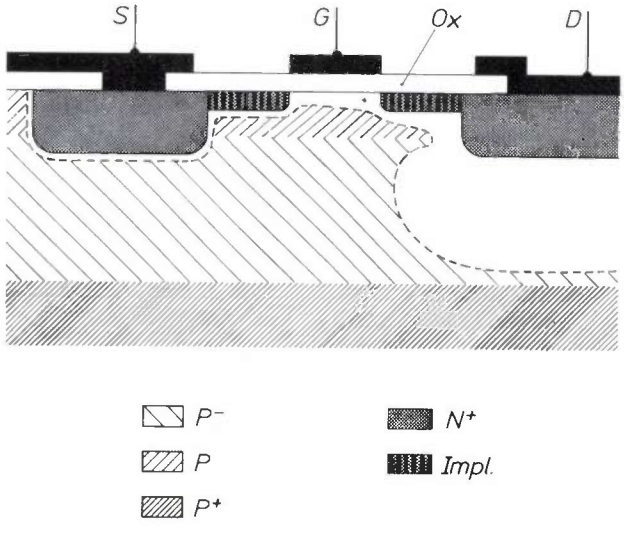


Fig. 3. An autoregistered MOS transistor with an implanted *P* layer. The channel is situated in the implanted layer while the major part of the drain depletion region is in high-resistivity P^--material epitaxially grown on a P^+ substrate giving a low drain capacitance.

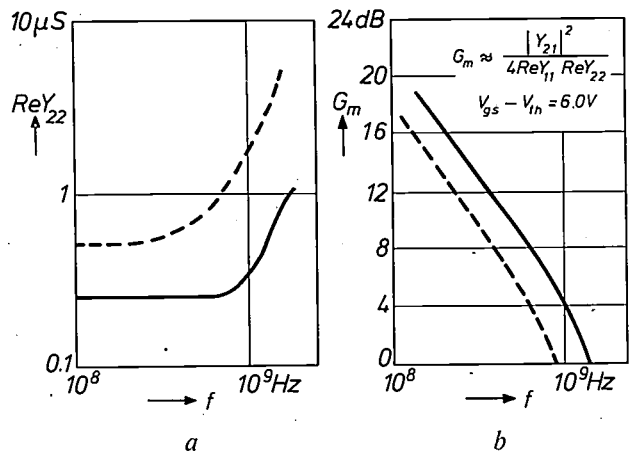


Fig. 4. *a*) The output conductance Re Y_{22} (expressed in siemens; 1 S = 1 A/V) and *b*) the maximum available power gain G_m of a MOS transistor with an implanted layer (fig. 3) plotted as a function of frequency *f*. The layer was made by bombarding the substrate with 7.5×10^{12} boron ions/cm² at 100 keV. For comparison curves have been drawn (*dashed lines*) for a similar transistor made without an implanted layer, on a 2 Ωcm (111) substrate. Both transistors have 3 μm channel lengths.

The output conductance of an autoregistered *N*-channel device incorporating a layer made by implanting 7.5×10^{12} boron ions/cm² at 100 keV into a 15 Ωcm epitaxial *P*-layer before growing the gate oxide is compared with that of a simple autoregistered structure in *fig. 4a*. The latter structure was made in 2 Ωcm material. The lower output conductance at low frequencies of the former structure is due to a higher doping level in the channel. At high frequencies the output conductance increases due to the shunting effect of the drain capacitance across the output terminals. This effect has been considerably reduced with the implanted layer structure.

The maximum available power gain G_m of a device at a given frequency depends on the biasing conditions. In fig. 4*b* the power gains of the two devices are plotted for the same gate voltages above threshold; the formula shown in the figure refers to MOS transistors in which the feedback capacitance C_{fb} is so small as to be negligible. Although the mutual conductance Y_{21} of the implanted-layer structure is low (see below) due to the highly doped channel region [5] [6], the low-frequency gain is higher than that of the simple autoregistered structure owing to a much smaller output conductance Re Y_{22}. At high frequencies the low parasitic drain capacitance of the implanted-layer structure delays the 12 dB/octave fall-off and gives f_{max} at 1.4 GHz.

Ion implantation introduces a high density of defects, hence additional scattering occurs in the channel of MOS transistors made on these layers. The effective channel mobility of a MOS transistor can be estimated from the combined substrate- and gate-controlled mutual conductance, which is equal to βV_{gs} and independent of substrate doping [7]. The quantity β is proportional to the channel mobility and contains the dimensions of the MOS transistor [4]. The mobility can be obtained from it if the dimensions are substituted. The gate-controlled and gate-plus-substrate-controlled mutual conductance are plotted as a function of the effective gate voltage $V_{gs} - V_{th}$ in *fig. 5* for devices made on the same slice, one with an ion-implanted layer. In the measurements the effective length of the channel was kept constant by adjusting the value of V_d so that $V_{gs} - V_d$ was a constant. The slopes of these curves at low effective gate voltages give effective channel mobilities of 340 cm²/Vs for the non-implanted and 275 cm²/Vs for the implanted device. The difference between these values is close to that expected from additional impurity scattering in the more highly doped implanted layer and suggests that any radiation damage in the layer remaining after growth of the gate oxide does not have a major effect on carrier mobility.

As fig. 5 indicates the mutual conductance of these devices does not increase linearly with $V_{gs} - V_{th}$ as predicted by simple theory, but tends to saturate. This effect is particularly noticeable with the more lightly

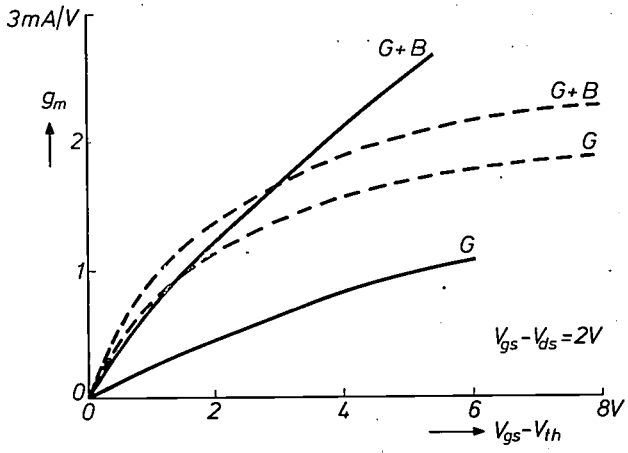


Fig. 5. The gate-controlled (*curves G*) and gate-plus-substrate-controlled (*curves G + B*) mutual conductance g_m of MOS transistors with 3 μm channel lengths plotted against the effective gate voltage $V_g - V_{th}$. The solid curves refer to a device made on an implanted layer in a 10 Ωcm substrate ($V_{th} = +0.7$ V), the dashed curves refer to a device made without an implanted layer on the same substrate ($V_{th} = -3.9$ V).

[2] J. M. Shannon, J. Stephen and J. H. Freeman, Electronics **42**, No. 3, 96, 3 Feb. 1969.
[3] A discussion of a c. behaviour of MOS transistors, including neutralizing and the definition of the different types of power gain, can be found in the article by P. A. H. Hart and F. M. Klaassen in this issue, page 216.
[4] See the article by J. A. van Nielen in this issue, page 209.
[5] See J. M. Shannon, Electronics Letters, **5**, 181, 1969 (No. 9), and the article by R. J. Nienhuis in this issue, page 259.
[6] J. A. van Nielen and O. W. Memelink, Philips Res. Repts. **22**, 55, 1967.
[7] M. B. Das, Solid-State Electronics **11**, 305, 1968.

doped substrate. The pinch-off potential of the device made on the lightly doped substrate is approximately three times that of the device on the more highly doped implanted layer and there are much higher fields parallel to the channel. It seems probable that the carrier velocity in these short-channel devices where such high fields are present tends to saturate in a manner similar to that observed in bulk material under high fields; the carrier velocity ceases to increase linearly with electric field but tends to a limit. This means that for the device without an implanted layer the power gain under biasing conditions shown in fig. 4*b* is close to the maximum obtainable.

This work is published by permission of the United Kingdom Ministry of Defence (Navy Department).

Summary. MOS transistors have been made with a low feedback capacitance using ion implantation to implant the source and drain regions with the metal gate in position acting as a mask. All the available power gain from these autoregistered *P*-channel and *N*-channel devices made using boron and phosphorus ions respectively could be used without needing neutralization. The high-frequency performance of a MOS transistor has been improved further by making the drain junction area in high-resistivity material thus reducing the parasitic drain capacitance. The channel region of the device is located in an implanted layer. Devices with implanted layers have been made with a maximum frequency of oscillation of 1.4 GHz.

MOS transistors in thin monocrystalline silicon layers

J. A. van Nielen, M. J. J. Theunissen and J. A. Appels

MOS transistors and MOST circuits are usually made by the planar technique [1], starting from wafers of monocrystalline silicon with a thickness of at least 100 microns. For some time now, interesting results have been obtained with silicon "wafers" which are much thinner, e.g. 1 to 2 microns. To make them easier to handle these layers are usually mounted on an insulating substrate.

What are the advantages of such a thin silicon layer, and what can it be used for? First of all, in making diffusion zones it is possible to ensure that they extend through the whole thickness of the layer. Since the *P-N* junctions are now only perpendicular to the thin layer, the capacitance of the diffusion zones is considerably reduced. Secondly, it is not difficult to etch away the whole thickness of parts of the silicon layer. For example, the various circuit elements can be isolated from one another in separate islands. Since the interconnections (the "wiring") between the islands are then no longer carried by the silicon, this also results in a reduction in the wiring capacitance. Thirdly, by using thin silicon layers it becomes possible to make a type of MOS transistor that has no *P-N* junction at all. We shall discuss this type of transistor in more detail later.

One of the methods that has been used for some time for making MOS transistors and MOST devices in a thin silicon layer is based on the combination of silicon and sapphire [2]; good results have also been obtained with silicon and certain spinels [3]. Sapphire, i.e. crystalline aluminium oxide, is an excellent insulator, and in a certain orientation the arrangement of the atoms in the surface resembles that of silicon so closely that it is possible to grow monocrystalline silicon epitaxially on this surface from a gaseous atmosphere (heteroepitaxy). In this way a layer of silicon a few microns thick can be obtained which is firmly joined to a layer of sapphire perhaps 200 microns thick. The product is easy to handle and is reasonably resistant to the high temperatures that have to be used for diffusing the required dopants (about 1150 °C).

This method is not however ideal in all respects: the silicon grown on sapphire and on the other materials mentioned does not entirely meet the quality requirements usually demanded of the silicon used as starting material in the planar technique. It contains more crystal defects, and therefore the mobility of the charge carriers and in particular the life time of the minority carriers are less than they are "normally". There are more crystal defects mainly because the sapphire or spinel lattice does not completely match the silicon lattice.

At Philips Research Laboratories in Eindhoven a method has been developed by means of which thin silicon films of high quality can be made by first growing an epitaxial layer a few microns thick on a silicon substrate and afterwards removing the substrate. This is done by means of an exceptionally neat electrochemical etching process, developed by H. J. A. van Dijk [4], which makes it possible to remove the silicon selectively, depending on the degree of doping. As expected the crystal quality of the epitaxial layer is in no way affected by the etching.

What we have said above has been confined to MOS transistors and MOST devices, but it will be evident that the possibility of improving the insulation between the components of a circuit and of reducing the wiring capacitance is of considerable significance for solid-state circuits.

The etching process and further operations

If we put a silicon anode in a dilute solution of HF, the silicon will be selectively etched away depending on a number of factors: the most important ones are the conduction type of the silicon, the concentration of the dopant and the magnitude of the applied voltage. This selective etching action can be used for removing epitaxial layers from the substrate on which they were grown. This is possible if the substrate is more strongly doped than the epitaxially grown silicon. Under appropriate conditions the etching process then stops "automatically" at the boundary of the epitaxially grown part. The best results up to now have been obtained with the combinations N^+-N and N^+-P. The combinations P^+-N and P^+-P have also

Ir. J. A. van Nielen, Ir. M. J. J. Theunissen and J. A. Appels are with Philips Research Laboratories, Eindhoven.

[1] See A. Schmitz, Philips tech. Rev. 27, 192, 1966.
[2] See J. D. Filby and S. Nielsen, Single-crystal films of silicon on insulators, Brit. J. appl. Phys. **18**, 1357-1382, 1967.
[3] See G. W. Cullen, G. E. Gottlieb, C. C. Wang and K. H. Zaininger, J. Electrochem. Soc. **116**, 1444, 1969 (No. 10).
[4] See H. J. A. van Dijk and J. de Jonge, J. Electrochem. Soc. **117**, 553, 1970 (No. 4), and M. J. J. Theunissen, J. A. Appels and W. H. C. G. Verkuylen, J. Electrochem. Soc. **117**, 959, 1970 (No. 7).

been investigated, but have proved to be less suitable because the resultant surface is somewhat uneven.

The electrochemical etching is followed by ordinary chemical etching. This is necessary because in the narrow boundary region between the substrate and the epitaxial region the doping concentration makes a diffuse transition instead of forming a sharp junction. Owing to the effect of the doping concentration on the electrochemical etching process it is not possible to preserve a homogeneous silicon layer by using this method alone.

Thus, with the combination N^+-P a layer of N-type silicon (which is extremely thin) is left behind on the surface of the P-type region; this layer is removed in the chemical etching stage. Chemical etching also has is evaporated on to the oxide film. This substrate is tough enough to stand up to the various operations that still have to be carried out, in particular the electrochemical etching and the diffusion at high temperature.

A second method is illustrated in *fig. 2*. Here the application of the insulating substrate is the last step. First the normal processing is carried out to make the transistor or circuit diffusion zones in the epitaxial layer. The surface is then temporarily protected by a coating of wax and a glass plate. The electrochemical and chemical etching processes come next. The wax is then dissolved away and the very thin substrate is attached to an insulating substrate, such as a ceramic plate, by means of a suitable polymer.

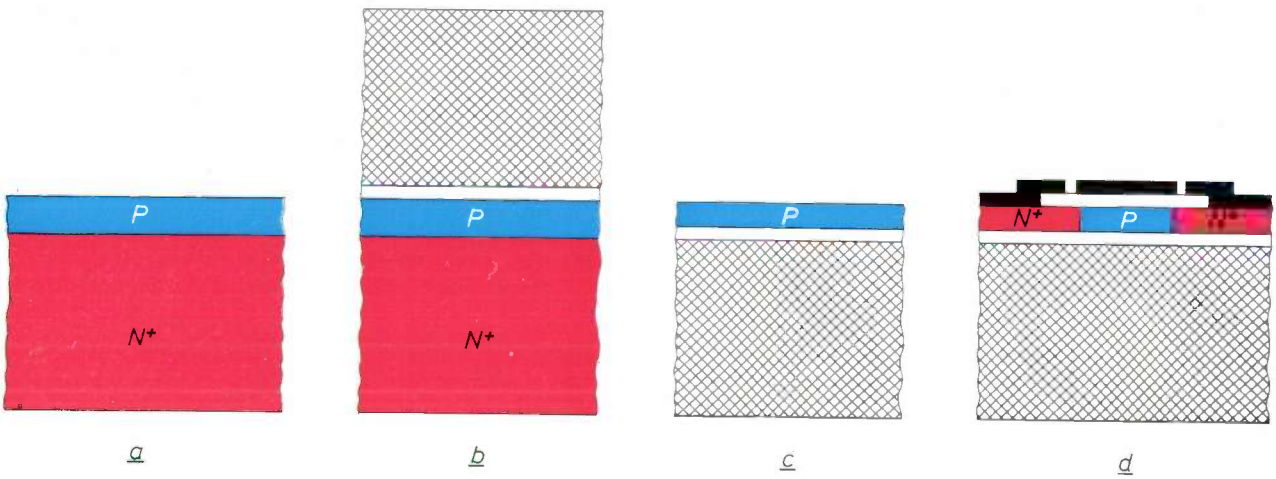


Fig. 1. One of the methods of making MOS transistors in thin silicon layers. Instead of N^+ and P-type silicon N^+ and N-type silicon can also be used.
a) A thin P-type silicon layer is grown epitaxially on a single-crystal wafer of N^+ silicon, 200 μm thick.
b) The epitaxial layer is oxidized and polycrystalline silicon is evaporated on to it. The oxide film and the polycrystalline silicon (cross-hatching) together form the insulating substrate.
c) The N^+ layer is etched away electrochemically. This is done selectively: the etching stops at the boundary of the epitaxial region. Chemical etching completes the operation. (Note that the wafer is upside down.)
d) MOS transistors are produced in the silicon layer by the usual masking and diffusion techniques.

to be used if it is necessary to etch down to a zone in which strongly doped diffusion zones are located (fig. 2).

The silicon layers can be etched to a thickness of 0.5 μm. It should be noted in this connection that the minimum layer thickness is not determined by the etching process but primarily by irregularities in the thickness and doping of the epitaxial layer.

There are various ways in which an insulating substrate can be applied. *Fig. 1* shows the method in which the insulating layer is applied before etching. An oxide film of about 1 μm is grown on the epitaxial layer, and a 200 μm thick film of polycrystalline silicon

Examples of application

To illustrate the potential of the new method, we shall first consider the reduction in drain capacitance that can be obtained, and then discuss a MOS transistor without P-N junctions that has very good high-frequency performance.

Reduction of the drain capacitance

By making a MOS transistor in a thin layer it can be made in such a way that there is no P-N junction below the drain (fig. 1d). Only the side walls of the drain then contribute to the P-N junction capacitance. *Fig. 3* gives a comparison of the drain capacitance of a

P-channel transistor before and after etching. The figure shows that the capacitance is reduced by a factor of 30, corresponding to the reduction in junction area that is obtained after making the slice thinner (by the second method).

The drain capacitance of a conventional MOS transistor is proportional to the area of the drain region, and the transconductance is proportional to the circumference of the drain. For this reason such transistors are usually made in the form of a ribbon, which is "folded" to save space. The drain capacitance of the thin-layer MOS transistor described above is not proportional to the area but to the circumference. This means that the transistor can be given any shape, for example circular. An advantage of such a simple shape is that it simplifies the alignment of the masks used in the fabrication of the transistor. Moreover the series resistance of the drain is negligibly small.

An important point is that the various useful features are not obtained at the expense of others [6]. For example, measurements before and after etching have shown that the etching causes no noticeable change in the mobility of the charge carriers in the inversion channel or in the drain leakage current.

A MOS transistor for the UHF band

Our second example is a MOS transistor with no *P-N* junction which has both a low drain capacitance and a low feedback capacitance, and can be used in the UHF band.

If source and drain diffusion of the *same* conduction type are produced in a substrate of normal thickness, giving for example an N^+-N-N^+ combination, an applied voltage between source and drain sets up a high current which cannot be controlled by the gate electrode. The situation is different if the thickness of the substrate is reduced to 2 microns or so. The depletion zone caused by a negative gate voltage can then cover a considerable part of the N layer. As long as no inversion channel appears at the surface, the depletion layer can be made to expand by increasing the gate voltage, and thus the current can be controlled. In this way a depletion-type MOS transistor with no *P-N* junction has been obtained. Given appropriate doping and an N layer of appropriate thickness, the depletion zone can be made to cover the whole thickness of the layer before an inversion channel appears.

As soon as the depletion zone reaches the underside of the N layer (*fig. 4*), the current is "pinched off" and

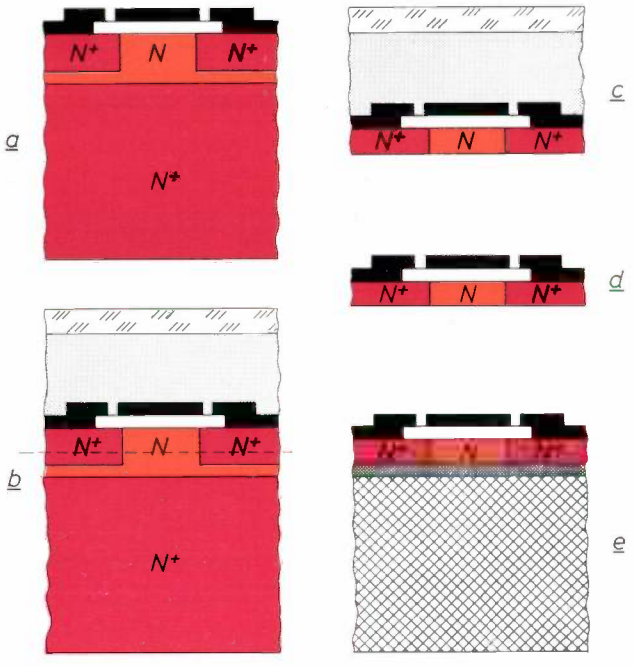


Fig. 2. An alternative to the method illustrated in fig. 1.
a) After the epitaxial growth of N-type silicon on N^+ silicon the operative parts of the transistor are made. The example given here relates to a MOS transistor without a *P-N* junction (an N^+-N-N^+ transistor, see text).
b) The operative part of the transistor is temporarily protected with wax (shaded region) and a glass plate.
c) The substrate (N^+) is etched away electrochemically. Silicon is etched away chemically down to the dotted line at (*b*). Earlier etching (during stage *a*) permits the separation in this stage of the individual transistors.
d) Glass and wax are removed.
e) The transistor is fixed to a ceramic plate (cross-hatching) by means of a polymer (darkly shaded region).

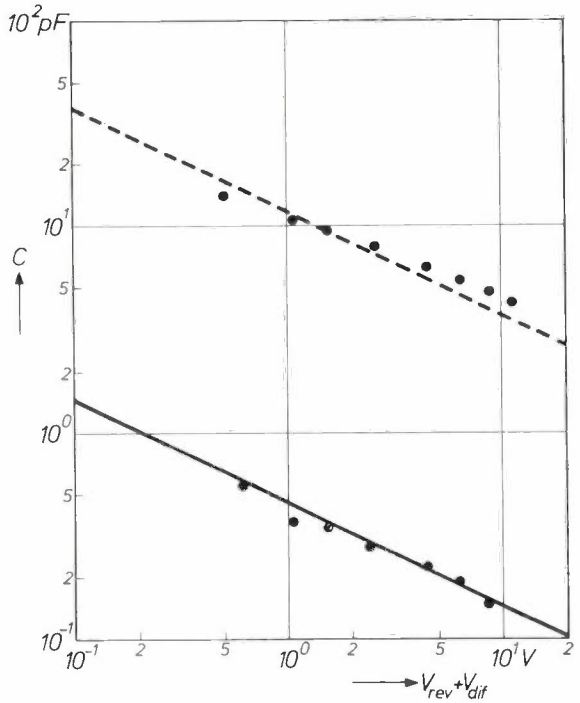


Fig. 3. Drain capacitance C_d of a 3 μm thick N^+-P-N^+ MOS transistor as a function of the reverse bias V_{rev} plus a constant diffusion voltage V_{dif} of 0.6 V [5]. The drain is circular with a diameter of 360 μm. For comparison the drain capacitance before etching is shown (dashed line).

[5] See L. Heijne, Philips tech. Rev. **25**, 120, 1963/64, in particular page 131.
[6] We also found that this was so when bipolar transistors were made: there was no difference before and after etching apart from a higher collector series resistance.

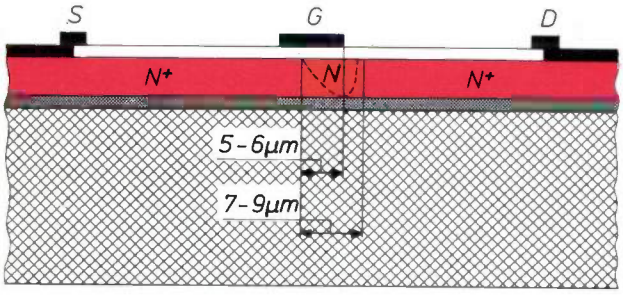


Fig. 4. Schematic cross-section of the MOS transistor with no *P-N* junction (N^+-*N*-N^+) and with offset gate. *S* source. *G* gate. *D* drain. The dashed line indicates the depletion zone.

current saturation occurs, just as in a conventional MOS transistor. Since the potential between gate and channel is greatest at the drain, the pinch-off point lies on the drain side. An important difference from the conventional MOS transistor is that the current conduction does not take place solely in a thin (inversion) layer at the surface, but over the whole layer (less the depletion zone, of course). Consequently the mobility of the charge carriers is greater than in the conventional MOS transistor. It will also be evident that the type of transistor described here has a low drain capacitance.

In the case of a P^+-*P*-P^+ transistor the doping and the thickness of the *P* layer can be chosen in such a way that all free holes between source and drain are driven out by the positive charge present in the oxide layer [7]. At zero gate voltage there is then no current conduction; no current can flow until a negative gate voltage is applied. In this way an *enhancement type* MOS transistor with no *P-N* junction has been produced. In combination with conventional *N*-channel MOS transistors of the enhancement type, these transistors can be used for making complementary MOST circuits.

Fig. 4 shows a schematic cross-section of our transistor. The drain is a disc with a diameter of 600 μm. The gate and source are arranged round it in the form of rings. The distance between source and drain is about 8 μm. To achieve a low feedback capacitance this transistor has what is termed an offset gate, that is to say the gate metal does not extend up to the drain, leaving uncovered about 2.5 μm of *N*-type silicon channel adjoining the N^+ drain region. The resistance of this *N* region is so low that the high-frequency characteristics of this transistor are not adversely affected. The region may be regarded more or less as an extension of the drain. Nevertheless, the feedback between drain and gate is small because the depletion layer (dashed line) extends a little way into this *N* region at saturation.

Fig. 5 is a photograph of the transistor, which is about 2 μm thick, lying on a photomask. The photograph shows that the silicon layer is so thin that the pattern of the photomask can be seen through the transistor. The transistor can be seen again in *fig. 6*, this time on a standard mount.

At a drain current of 10 mA and a drain voltage of 10 V the transconductance of the transistor is 6 mA/V. The feedback capacitance is no more than 0.15 pF, and the capacitance between source and drain is about 0.5 pF. The cut-off frequency at which the available gain is 1 is 1400 MHz (*fig. 7*).

The high cut-off frequency is due both to the low drain and feedback capacitances and to the high

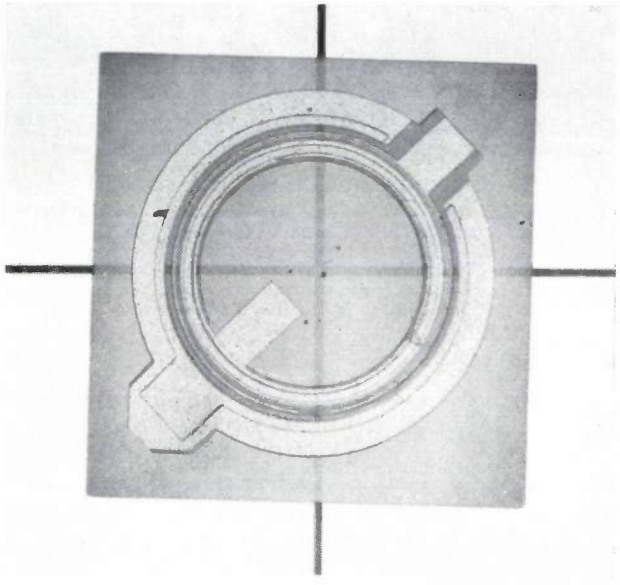


Fig. 5. Photograph of the transistor in fig. 4 *without* substrate, lying on a photomask. The silicon layer is so thin that the pattern of the photomask can be seen through the transistor (the cross is part of the pattern). The silicon chip is 1 mm square.

Fig. 6. The transistor in fig. 5 on a standard TO 18 mount.

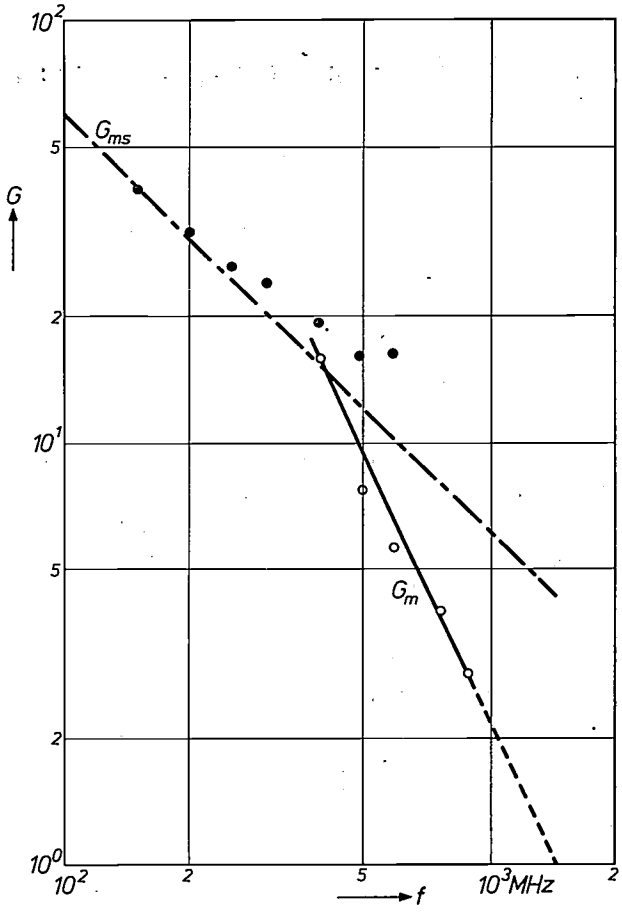


Fig. 7. Maximum available gain G_m (circles) of the MOS transistor described in the text, with no *P-N* junction, and with offset gate. The transistor is fixed to a standard mount TO 18. The cut-off frequency at which the maximum available gain is equal to 1 is obtained by extrapolation of the values measured between 400 and 900 MHz. Measurement of the *Y*-parameters, at frequencies up to 600 MHz, shows that the transistor is unstable as an amplifier at frequencies below 400 MHz, which implies that the gain G_m is not defined below 400 MHz. The maximum stable gain G_{ms}, calculated from the *Y*-parameters, is therefore given for this frequency range. The values found (black dots) lie above chain-dotted line of slope −1 since no correction was made in this calculation for the apparent increase of Y_{21}, due to the inductance of the contact wires and the pins of the mount.

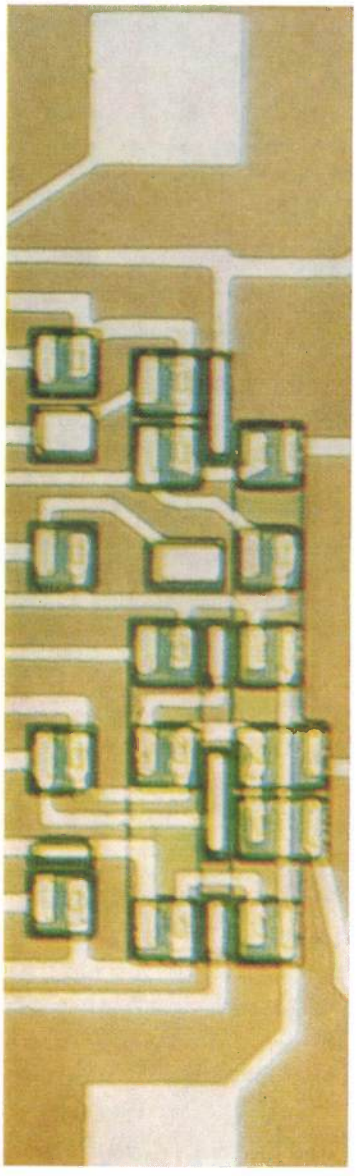
mobility of the charge carriers. For comparison a conventional MOS transistor of the depletion type was made with the same masks. From measurements of the *Y*-parameters of both types of MOS transistor it can be shown that a conventional MOS transistor with identical transconductance, feedback and input impedance, would have a cut-off frequency of only 500 MHz, because of its higher output capacitance.

As discussed in other articles in this issue, a cut-off frequency of 1400 MHz can also be obtained by means of ion implantation [8] or by using a tetrode structure with short channels [9]. The cut-off frequency of the transistor discussed here can be further increased by making the channel shorter and reducing the thickness of the oxide film, which in the present case was 0.2 μm. A thickness of 0.1 μm is quite common nowadays, and it should be possible to halve the length of the channel.

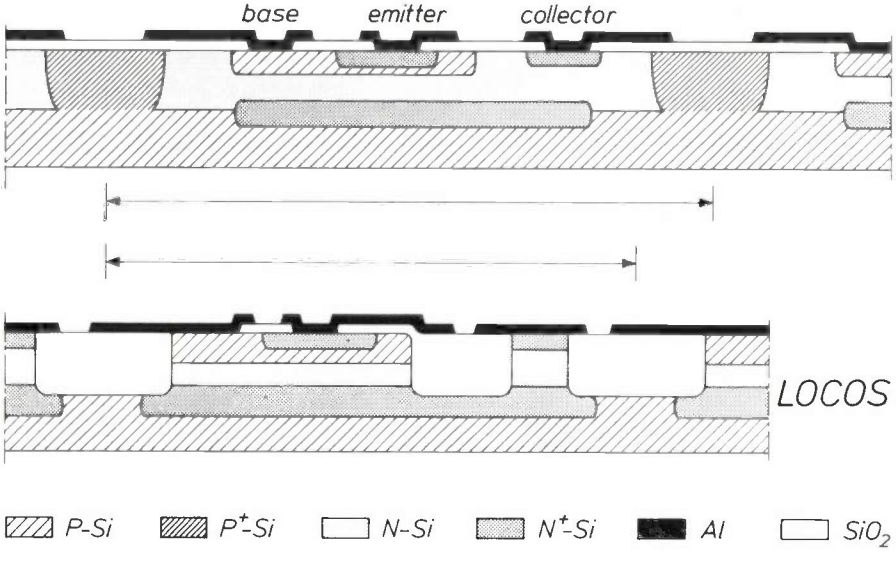
[7] See the articles by J. A. Appels, H. Kalter and E. Kooi and by J. A. van Nielen in this issue, pages 225 and 209.
[8] See the article by J. M. Shannon in this issue, page 267.
[9] See the article by R. J. Nienhuis in this issue, page 259.

Summary. Thin silicon layers of high quality can be made by first growing an epitaxial layer of silicon a few microns thick on a silicon substrate, and then removing the substrate by selective electrochemical etching. Two of the various possible applications of the method are discussed: a MOS transistor 3 microns thick which has a drain capacitance about 30 times lower than that of a conventional MOS transistor, and a MOS transistor 2 microns thick which has no *P-N* junction and has an offset gate (channel about 8 μm long, of which about 2.5 μm is not covered by the gate metal; feedback capacitance 0.15 pF, cut-off frequency 1400 MHz).

LOCOS technology



One of the articles in this issue [1] gives a description of a method for making integrated circuits, developed at Philips Research Laboratories, in which thick oxide layers are sunk into the silicon chip to avoid the necessity for large steps in the metallization. This is the LOCOS technique. It is of importance not only for MOS transistors, but also for integrated circuits with bipolar transistors [2]. Comparison of the two drawings, showing the cross-section of an N^+PN transistor in an integrated circuit made by the conventional procedure and one by the LOCOS technique, indicates that with the LOCOS technique the P diffusion isolating the transistor from its neighbours has been replaced by a thick strip of oxide. Unlike the P diffusion, this oxide strip borders directly



on the P layer of the transistor; this arrangement saves a good deal of space on the chip. Another important advantage is the reduction of the parasitic capacities which is obtained with the LOCOS technique. In the example shown here the connection between the collector and the appropriate contact zone is made by a buried layer that passes underneath the thick oxide strip. This enables the base zone and the collector contact zone to be defined by a single photomask and — because they are flanked by oxide — the zones arrive automatically at the right place (autoregistration), which is an important advantage as well. The colour photograph shows a view from above of a part of an experimental integrated circuit, made with the LOCOS technique. Each of the green squares is a bipolar transistor. The metal connections (yellow-white) and the bonding pads are all situated on thick oxide (grey-brown).

[1] J. A. Appels, H. Kalter and E. Kooi, Some problems of MOS technology; this issue, page 225.

[2] 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 (No. 2).
J. A. Appels and M. M. Paffen, Local oxidation of silicon; new technological aspects, *ibid.*, **26**, 157-165, 1971 (No. 3).
E. Kooi *et al.*, LOCOS devices, *ibid.*, **26**, 166-180, 1971 (No. 3).

Digital integrated circuits with MOS transistors

L. M. van der Steen

The simple structure and low dissipation of MOS transistors allow them to be formed in large numbers per unit surface area on a crystal chip. Moreover, since a MOST has a very high input impedance and allows current to flow in both directions, some MOS circuits require fewer components than the corresponding bipolar circuits. These features make the MOS transistor eminently suitable for use in integrated circuits. In this article we shall describe some examples of digital integrated circuits of this type, made entirely from *P*-channel enhancement MOSTs.

Little need be said here about the MOS transistor; the *N*-channel type has been dealt with in detail elsewhere in this issue [1], and the *P*-channel type (see *fig. 1*) differs from this only in polarity. The I_d-V_{ds} family of characteristics and the graph of $I_{d\,sat}$-V_{gs} have the same shape as the graphs in fig. 2 of [1], except that currents and voltages are now negative. For application in digital circuits the *P*-channel MOST may be regarded as a switch that passes current when V_{gs} is sufficiently negative ($|V_{gs}| > |V_{th}|$, the threshold voltage V_{th} being typically −3 to −4 volts) and does not pass current when V_{gs} is between 0 volts and V_{th}.

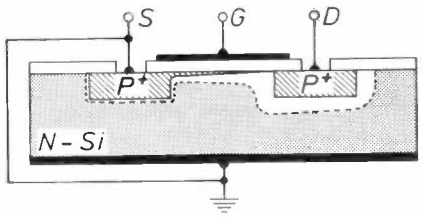


Fig. 1. *P*-channel MOS transistor on a substrate of *N*-type silicon. The source *S* and the drain *D* are P^+-type regions diffused in the silicon. The gate *G* is of aluminium (shown black) and is isolated from the substrate by a layer of silicon dioxide (white). Substrate and source are earthed; a negative voltage is applied to the drain *D*. If a sufficiently negative voltage is applied to the gate (lower than the threshold voltage V_{th}, which is also negative), a thin inverted *P*-type layer is formed beneath the gate, enabling current to flow between *S* and *D*. Below the source, drain and the channel there is a depletion layer, which acts as an insulator since it contains hardly any free charge carriers.

The manufacture of integrated MOS circuits

The voltages applied to a MOST are always arranged so that the drain and inversion layer are reverse-biased with respect to the substrate and therefore isolated from it by a depletion layer (see fig. 1). The source may be connected to the substrate; if it is not, the voltage on the source is always such that this electrode also is surrounded by a depletion layer. The natural isolation provided by this depletion layer presents considerable advantages in the manufacture of integrated circuits made up from MOS devices, since there is now no need to make an isolated island of the appropriate material by means of an epitaxially grown silicon layer and an isolation diffusion for each transistor, as there is with integrated circuits using bipolar transistors [2]. The number of photoetching processes needed for manufacture is thus reduced from six for bipolar-transistor circuits to four for MOS circuits.

The situations after the various photoetching processes [3] are shown schematically in *fig. 2*. The manu-

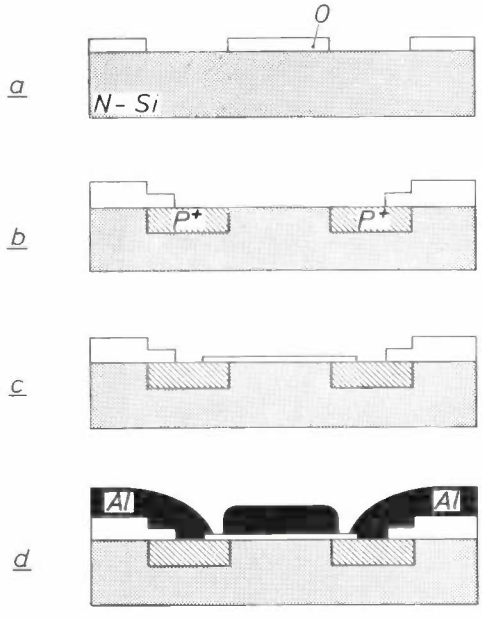


Fig. 2. Some stages in the manufacture of a MOS transistor. Windows for the source and drain are etched in an oxide layer (*O*) on the substrate (*a*). After diffusion of these electrodes, a window for the oxide layer below the gate is etched in the oxide film formed during the diffusion process (*b*). After the oxide layer has been formed, the windows are etched for the contacts with the source and drain (*c*). Finally an aluminium layer is deposited, from which the electrodes are formed by etching away surplus aluminium (*d*).

Drs. L. M. van der Steen is with the Philips Electronic Components and Materials Division (Elcoma), at Nijmegen.

[1] J. A. van Nielen, Operation and d.c. behaviour of MOS transistors; this issue, page 209.

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

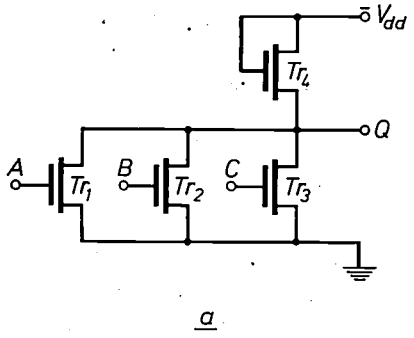
[3] For more details see the article by J. A. Appels, H. Kalter and E. Kooi in this issue, page 225.

facture of a *P*-channel MOS circuit starts with a single-crystal wafer of homogeneously doped *N*-type silicon (of diameter say 50 mm and thickness 250 μm), on which a silicon-dioxide film is formed in a hot oxygen-rich atmosphere. A photoetching process is used to make windows in this oxide film at the places where the source and drain are to be located (fig. 2*a*). The wafer is then heated in a boron-containing atmosphere, so that the boron diffuses through the windows into the *N*-type silicon and causes regions of *P*-type silicon to form. During the diffusion process a new layer of silicon dioxide forms on the wafer, and windows are made in this layer in a second etching process (fig. 2*b*). By heating in an oxygen-rich atmosphere a thin layer of silicon dioxide (0.1 μm) is then formed in these windows to give isolation between the substrate and the gate electrodes. In a third etching process windows are made in this oxide layer above the source and drain (fig. 2*c*). A layer of aluminium is then deposited, from which the various electrodes and connections between the components of the circuit are formed by removing the surplus aluminium in a fourth and last etching process.

Since in this technology there is no isolation diffusion, which takes up a good deal of space, a very large number of components can be formed on a silicon chip. And since the dissipation of MOS transistors is very low, such a high packing density is a practical proposition for many circuits. In a dynamic shift register, for example, which is discussed later on in this article, the local packing densities are as high as 6×10^4 components per cm^2, and the average packing density for the complete monolithic circuit is 2×10^4 cm^{-2}. A monocrystalline silicon chip measuring a few mm^2 can now accommodate complete MOS circuits of 300 to 1500 components; in the near future monolithic circuits with several thousand components will be possible. At the present, the average packing density that can be achieved in integrated circuits made with bipolar transistors is generally much smaller: about a few thousand per cm^2.

MOS logic circuits

Simple logic circuits such as NOR gates or NAND gates are generally used as the "building blocks" for digital circuits, since all digital circuits, from simple bistable circuits to storage elements, shift registers, etc., can be made up from such basic elements [4]. A gate circuit is very easily made with MOS transistors; *fig. 3a* shows such a circuit, which has four MOS transistors Tr_1 to Tr_4. Briefly, the circuit operates as follows. The drain and gate of Tr_4 are both connected to the negative supply voltage, so that Tr_4 always operates in the saturation region. However, current can only flow through



a

A	B	C	Q
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	0

b

Fig. 3. *a*) The NOR gate, made as an integrated MOS circuit. The supply voltage V_{dd} is negative. The voltages at the inputs *A*, *B* and *C* can be either zero or negative. The voltage at the output *Q* is at the zero level when one or more of the inputs are negative, since there is then current in the circuit. When negative logic is used (the negative voltage is then related to the state 1 and the zero level to state 0), the relation between the output level and the input levels is given by $Q = \overline{A + B + C}$, which is a NOR function. *b*) The truth table of the circuit with negative logic.

Tr_4 provided one or more of the transistors Tr_1 to Tr_3 are passing current, i.e. provided the voltage at one or more of the inputs *A*, *B* and *C* is sufficiently negative. The circuit is designed in such a way that in this situation the voltage at the output *Q* is at the "zero level", i.e. between 0 volts and V_{th}. If the voltage at each of the three inputs is at the zero level, no current flows and the output voltage is negative. The transistors Tr_1, Tr_2 and Tr_3 in this circuit thus behave as switches, and so are called *switching transistors*; whereas Tr_4 acts as a resistor and is therefore called a *load transistor*.

The voltage levels can be related in two ways to the states 0 and 1 used in the mathematical treatment of logical operations by Boolean algebra. In *positive logic* state 1 denotes the high level and state 0 the low, and *vice versa* in *negative logic*. It is the practice to use negative logic for logic circuits built up from *P*-channel MOS transistors: $A = 1$ here therefore means that the voltage at point *A* is negative, and $A = 0$ means that this voltage is at the zero level. In the circuit given in fig. 3*a* we then have: $Q = 1$ if $A = 0$, $B = 0$ and $C = 0$; $Q = 0$ if one or more of the inputs *A*, *B* and *C* are 1. Thus, *Q* is *not* 1 if *A or B or C* is 1; this NOT-OR relation is represented by the NOR function $Q = \overline{A + B + C}$, so that the circuit of fig. 3*a*, using negative logic, is a NOR gate. Fig. 3*b* gives the truth table for this case. In the transition from negative to positive logic AND and OR gates change their name, a NOR gate then becoming a NAND gate. This can be seen from the table in fig. 3*b* by interchanging ones and zeros; *Q* is then zero (i.e. *not* 1) only if *A and B*

[4] See for example page 48 in: E. J. van Barneveld, Digital circuit blocks, Philips tech. Rev. **28**, 44-56, 1967, and page 20 in: C. Slofstra, The use of digital circuit blocks in industrial equipment, Philips tech. Rev. **29**, 19-33, 1968.

and C are 1, so that Q is then indeed equal to $\overline{A.B.C.}$

The switching transistors, which are in parallel in fig. 3*a*, can also be arranged in series; the result is then a gate circuit which, with negative logic, acts as a NAND gate. If the output voltage in a current-conducting circuit is to remain within the zero level, the total resistance of the switching transistors must not exceed a particular value. This requirement presents no difficulties if the switching transistors are in parallel; if they are in series, however, an increasing number of inputs requires the use of larger and larger switching transistors, and the circuit then takes up a lot of space on the crystal chip. The circuit with parallel transistors is therefore preferred.

Inverter circuit

To describe the behaviour of the gate circuit during the transition between the levels, we can take the simplest situation, for which $B = 0$ and $C = 0$. A can be 0 or 1, and Q is then 1 or 0 as the case may be; Tr_1 and Tr_4 then constitute an inverter circuit. We shall now analyse this circuit, referring to *fig. 4*, and from this analysis we shall derive the requirements that have to be met in the design of the transistors.

Provided that there is no saturation the current in a MOS transistor is related to the applied voltage V_{ds} and the gate voltage V_{gs} by:

$$I_d = \beta(V_{gs} - V_{th} - \tfrac{1}{2}V_{ds})V_{ds}. \quad . \ . \ . \quad (1)$$

Here the threshold voltage V_{th} is the voltage between gate and source at which inversion starts, i.e. at which current begins to flow. If $|V_{ds}| \geqq |V_{gs} - V_{th}|$, saturation occurs and the current is given by:

$$I_{d\,sat} = \tfrac{1}{2}\beta(V_{gs} - V_{th})^2. \quad . \ . \ . \ . \quad (2)$$

In these equations $\beta = \mu C_{ox} w/l$, where μ is the mobility of the holes in the channel, C_{ox} the oxide capacitance per unit surface area (i.e. of the capacitor formed by gate, oxide and substrate), w is the width and l the length of the channel. (See equations (6) and (9) of the article mentioned in [1].)

For transistor Tr_1 in fig. 4 we have $V_{gs} = V_i$ and $V_{ds} = V_o$, so that:

$$I_{d1} = \beta_1(V_i - V_{th1} - \tfrac{1}{2}V_o)V_o \quad . \ . \ . \quad (3)$$

if $|V_i - V_{th1}| > |V_o|$ (no saturation), and

$$I_{d1\,sat} = \tfrac{1}{2}\beta_1(V_i - V_{th1})^2 \quad . \ . \ . \ . \quad (4)$$

if $|V_i - V_{th1}| \leqq |V_o|$ (saturation).

Since the gate of Tr_4 is connected to the drain, so that $|V_{ds}| > |V_{gs} - V_{th4}|$ at all times, this transistor will always operate in the saturation region. In fact Tr_4 acts here only as a resistor; a transistor is used at this position because a diffused resistor of high enough value would take up too much space on the crystal wafer. For Tr_4 we have $V_{gs} = V_{ds} = V_{dd} - V_o$ (see fig. 4), so that:

$$I_{d4\,sat} = \tfrac{1}{2}\beta_4(V_{dd} - V_o - V_{th4})^2. \quad . \ . \quad (5)$$

Fig. 5 shows the I_d-V_o characteristics of Tr_1; these have V_i as parameter and at $V_i - V_{th1} = V_o$ they all meet the horizontal part of the curve where the current is saturated. The figure also shows the curved load line which is obtained because Tr_4 is used as a load resistance and is derived from equation (5). Since at all times

$$I_{d1} = I_{d4}, \quad . \ . \ . \ . \ . \ . \quad (6)$$

the points where the load line intersects the I_d-V_o characteristics of Tr_1 are the operating points of the

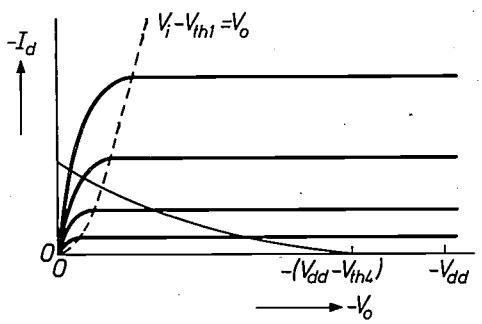


Fig. 5. I_d-V_o characteristics of the transistor Tr_1 from the inverter circuit of fig. 4, showing the curved load line which results when Tr_4 is used as a load resistor. In the characteristics of Tr_1 the parameter is V_i: when V_i is more negative the current is higher. The dashed line $V_i - V_{th1} = V_o$ connects the points where the current through Tr_1 goes into saturation.

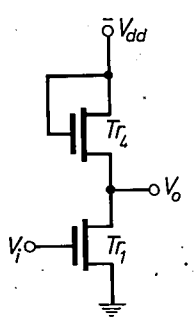


Fig. 4. MOS inverter circuit derived from the NOR gate of fig. 3*a* with $B = C = 0$. V_i and V_o are the input and output voltages. When V_i is at the negative level ($V_i = 1$), V_o is at the zero level ($V_o = 0$), and *vice versa*.

circuit at different values of V_i; these operating points set out in a V_o-V_i diagram form the V_o-V_i characteristic of the inverter circuit (*fig. 6*).

In the characteristic given in fig. 6 three regions can be distinguished. In region 1, V_i is at the zero level, i.e. $0 < |V_i| < |V_{th1}|$. Tr_1 is now not conducting, so that $I_{d1} = I_{d4} = 0$, and from equation (5) it follows that $V_o = V_{dd} - V_{th4}$. This is therefore the "1 level" of the output voltage. In fig. 6 this situation

corresponds to the horizontal part of the characteristic; in fig. 5 the operating point lies on the V_o-axis.

In region II V_i falls below the threshold voltage, so that Tr_1 starts to conduct. At the same time, however, $|V_i - V_{th1}| < |V_o|$, and therefore in fig. 5 the operating point lies to the right of the line $V_i - V_{th1} = V_o$, so that Tr_1 operates in saturation. In fig. 6 a linear transition now arises between the two voltage levels; the region is bounded here by the lines $V_i = V_{th1}$ and $V_i - V_{th1} = V_o$.

In region III V_i is even more negative, and here $|V_i - V_{th1}| > |V_o|$. The current through the circuit has further increased, and Tr_1 is now no longer in saturation. In this situation $V_i = 1$ and hence $V_o = 0$. To ensure that this is so, the voltage divider formed by Tr_1 and Tr_4 must be so arranged that V_o is between 0 and V_{th1} when the circuit is passing current (as a rule V_o is put at approximately 0.1 V_{dd}). Region III lies to the right of the line $V_o = V_i - V_{th1}$ in fig. 6 and to the left of it in fig. 5.

We shall now take a closer look at the situation in the transition region II, in which both transistors operate in saturation. From equations (4), (5) and (6) it then follows that:

$$\tfrac{1}{2}\beta_1(V_i - V_{th1})^2 = \tfrac{1}{2}\beta_4(V_{dd} - V_o - V_{th4})^2 \qquad (7)$$

and this gives:

$$V_o = -(\beta_1/\beta_4)^{1/2}(V_i - V_{th1}) + (V_{dd} - V_{th4}). \qquad (8)$$

In this region the V_o-V_i characteristic is therefore a straight line with a slope of $\alpha = -(\beta_1/\beta_4)^{1/2}$. *Fig. 7* shows a set of these characteristics for various values of α and for the same values of V_{dd}, V_{th1} and V_{th4}. In designing a circuit the choice of α is partly determined by the requirement noted above that in region III the output voltage should be at the zero level. This means that the slope must exceed a particular limiting value; in practice it is usual to take $\beta_1/\beta_4 \geqq 20$.

Since the transistors are parts of a monolithic circuit, and are thus formed during the same operation, they have the same thickness of oxide layer below the gates and the same mobility for the holes in the channels. In β_1 and β_4 the factors μ and C_{ox} are therefore identical, so that

$$\frac{\beta_1}{\beta_4} = \frac{w_1/l_1}{w_4/l_4}.$$

The ratio β_1/β_4 is therefore determined entirely by the w/l aspect ratios of the transistors. If, for example, the switching transistor has the dimensions $w = l = 10$ μm (so that $w_1/l_1 = 1$), and if β_1/β_4 should have a value of 20 to give the right slope, then w_4/l_4 for the load transistor should have a value of 1/20. Thus we could have $w = 10$ μm and $l \doteq 200$ μm.

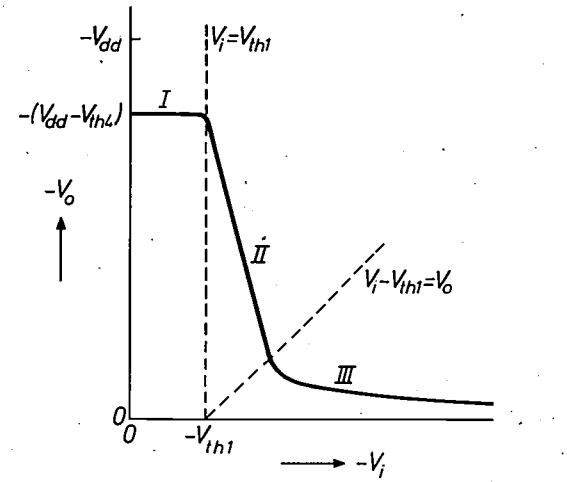


Fig. 6. V_o-V_i characteristic of the inverter circuit in fig. 4, derived from fig. 5 by plotting the points where the load line intersects the I_d-V_o characteristics of Tr_1. In region I, $V_i = 0$ and $V_o = 1$; in region III, $V_i = 1$ and $V_o = 0$. Region II is the transition between these situations. The curves $V_i = V_{th1}$ and $V_i - V_{th1} = V_o$ form the boundaries between the regions. In I and II the curve is virtually a straight line.

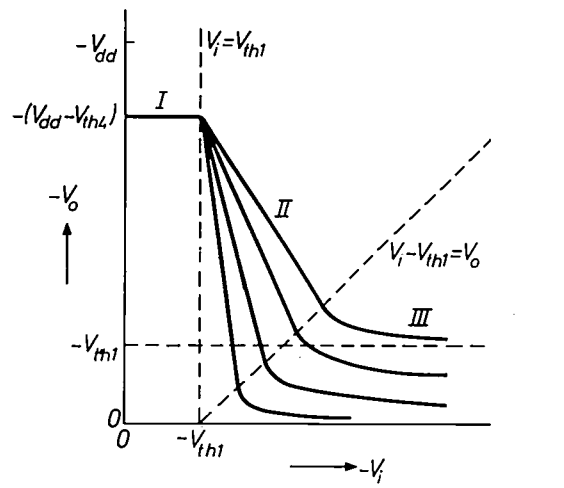


Fig. 7. Some V_o-V_i characteristics for different values of β_1/β_4. In the transition region II the slope of the curves is proportional to $(\beta_1/\beta_4)^{1/2}$. The output voltage in region III is within the zero level only for the two steepest curves in this region.

The characteristics of figs. 6 and 7 hold for the NOR circuit of fig. 3*a* when only one switching transistor is conducting. If several switching transistors are activated simultaneously, V_o adjusts itself to a still lower value; the NOR circuit will operate correctly in any situation if each of the switching transistors has the correct value of β to suit the load transistor.

If the load transistor were to be replaced by a diffused resistor, a widely used integrated-circuit element, the surface area of this resistor would be much larger than that of the load transistor. The resistance between source and drain of a MOS switching transistor of minimum dimensions ($w = l = 10$ μm) is several tens of kΩ. This means that a load resistance of several hundred

kΩ should be used to give the correct voltage division. A diffused resistor of this value would be about 10 mm long and 10 μm wide, whereas the corresponding load transistor with the same width is only 0.2 mm long. Although the load transistor has the advantage of being small it has the disadvantage that the full supply voltage can never be obtained at the output of the circuit, since the threshold voltage of the load transistor is always lost. Another disadvantage is that the response of the circuit with the load transistor is slower than when a diffused resistor is used. However, the most important feature is the extra space gained on the chip.

The maximum switching frequency

The maximum permissible switching frequency of a NOR circuit is determined by the speed at which the output voltage switches over, and hence by the rate at which the capacitor connected to this output is charged up. Of course, we must consider the worst case here, i.e. the transition from 0 to 1, at which the switching transistors stop conducting and the output capacitor is therefore only charged up by the load transistor.

In a logic circuit the load on the output of a NOR will usually be several inputs of other NOR circuits. Here we shall take the simplest case where the only load on the NOR is a single input of another NOR (e.g. input A in fig. 3*a*), so that the load is determined mainly by the capacitance C_1 of this input. Then the charging current $I_{d4\,sat}$ is equal to $\frac{1}{2}\beta_4(V_{dd} - V_o - V_{th4})^2$, and is therefore also equal to $C_1 dV_o/dt$. From the differential equation obtained by equating these two expressions it can be shown that the RC time constant is proportional to C_1/β_4, and hence to $\alpha^2 l_1^2/\mu$.

Since the value of α is already fixed, we must make l_1 small to be able to reach a high switching frequency. The channels in the switching transistors must therefore be as short as possible and μ must be as large as possible.

A much more unfavourable situation is found when the output of a NOR gate is not connected to another NOR circuit but to a point outside the integrated circuit. The load capacitance is then usually determined by the printed wiring of the board on which the integrated circuit is mounted. This wiring capacitance may easily be 1000 times larger than the input capacitance of a NOR (a few tens of pF against a few tens of fF; 1 fF $= 10^{-15}$ F), and consequently the switching frequency for the logic circuit as a whole would be 1000 times lower. Measures therefore have to be taken to keep the RC product as small as possible. This is done by using larger load transistors for the final NOR circuit connected to this high capacitance, which make the current 100 times greater than in the earlier circuits. The resistance of the load transistor in this circuit is then only 1/100 of the original value, so that the RC product at the output is now not 1000 times larger but only 10 times. The switching transistors in the last stage must of course be made larger in the same way, and this in turn gives a 100 times greater capacitive load for the penultimate stage. To maintain a reasonable RC time constant the current through this stage must therefore be increased as well, say by a factor of 10, so that larger transistors are also necessary here. The output stages of a circuit thus take up a relatively large amount of space on the chip; in fact it may happen that no transistors at all of minimum dimensions can be used in a small circuit.

MOS shift registers

It can be seen from the above that MOS transistors are most suitable for use in integrated circuits with a small number of outputs. This is the case, for example, in stores with a fixed information content (read-only stores), in "scratch-pad" stores and in shift registers. A shift register consists of a sequence of bistable circuits, or cells whose information content, i.e. the state of the individual cells, moves up one place when a shift pulse or a combination of pulses is applied. Thus, information supplied to the input of the first cell appears, after a number of shift pulses, at the output of the last cell of the register. Shift registers are widely used as storage elements in the arithmetic unit of a computer.

A distinction is made between *static* and *dynamic* shift registers. In static shift registers the information content of the register is stored for an unlimited time after a shift pulse. In dynamic shift registers the information content is soon lost and therefore has to be repeatedly replenished. The shift pulses must consequently be repeated at a particular minimum frequency, and this means that the information in a dynamic shift register can never be kept in the same place. Another disadvantage of dynamic shift registers is that they require a more complex combination of shift pulses than a static type. On the other hand, a dynamic shift register is a great deal faster than a static one, and its dissipation is also much lower. Moreover, in integrated-circuit form a dynamic shift register can have a greater packing density than a static shift register.

The static shift register

Fig. 8a gives the diagram of a static shift-register cell built with MOS transistors. The cell consists of a bistable circuit (or flip-flop) formed by two inverters Tr_1-Tr_2 and Tr_3-Tr_4 coupled to form a loop by means of Tr_6 and Tr_7. The capacitors C_1 and C_2, which play an essential part in the operation of the cell, are formed by the capacitances between the gates of Tr_1 and Tr_3 and the substrate, and by some stray capacitances. The

output voltage V_o represents the information content of the cell. States 0 and 1 correspond, as in the inverter circuit, to the zero level (V_o between 0 volts and V_{th1}) and the negative level ($V_o = V_{dd} - V_{th4}$). The information in the preceding cell of the register is presented as a voltage V_i at the input of the cell. The gate signals Φ_1, Φ_2 and Φ_3, consisting of shift-pulse trains (fig. 8*b*) are also fed to the circuit; the pulses Φ_2 and Φ_3 differ only in the steepness of their leading edges. Whenever a combination of shift pulses arrives, the information content of the previous cell is taken over; in other words, the voltage V_o assumes the level of V_i.

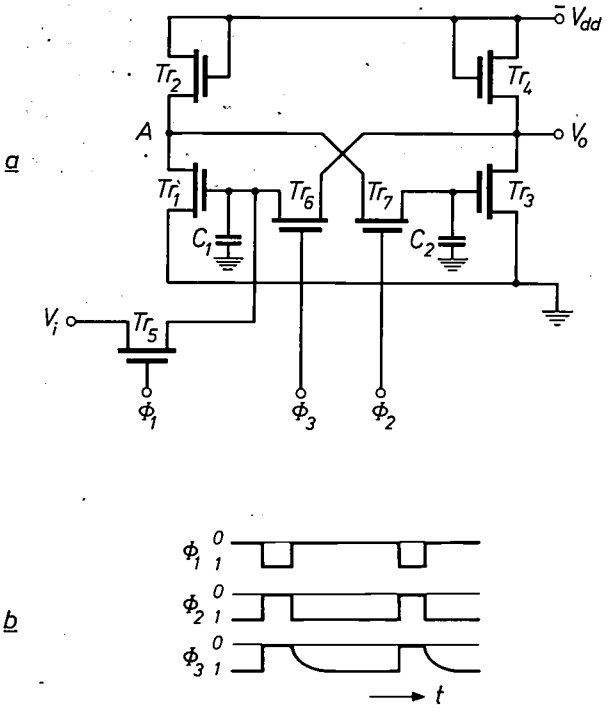


Fig. 8. *a*) Diagram of a cell of a static shift register. Tr_1-Tr_2 and Tr_3-Tr_4 form two inverter circuits, which are coupled to form a closed loop by means of Tr_6 and Tr_7. The level of V_o (zero or negative) is regarded as the "state" of the cell; V_i indicates the state of the preceding cell. *b*) The control (gate) signals Φ_1, Φ_2 and Φ_3 form a train of shift pulses. The information at the input is transferred to the output by each group of three pulses.

Fig. 9a gives a photograph of a 64-bit static shift register built up from cells of the type shown in fig. 8*a*. An idea of the way in which the various components are joined together to form a circuit can be obtained from fig. 9*b*, which illustrates a single cell from the register; this is the cell that can be seen at the upper left of the photograph.

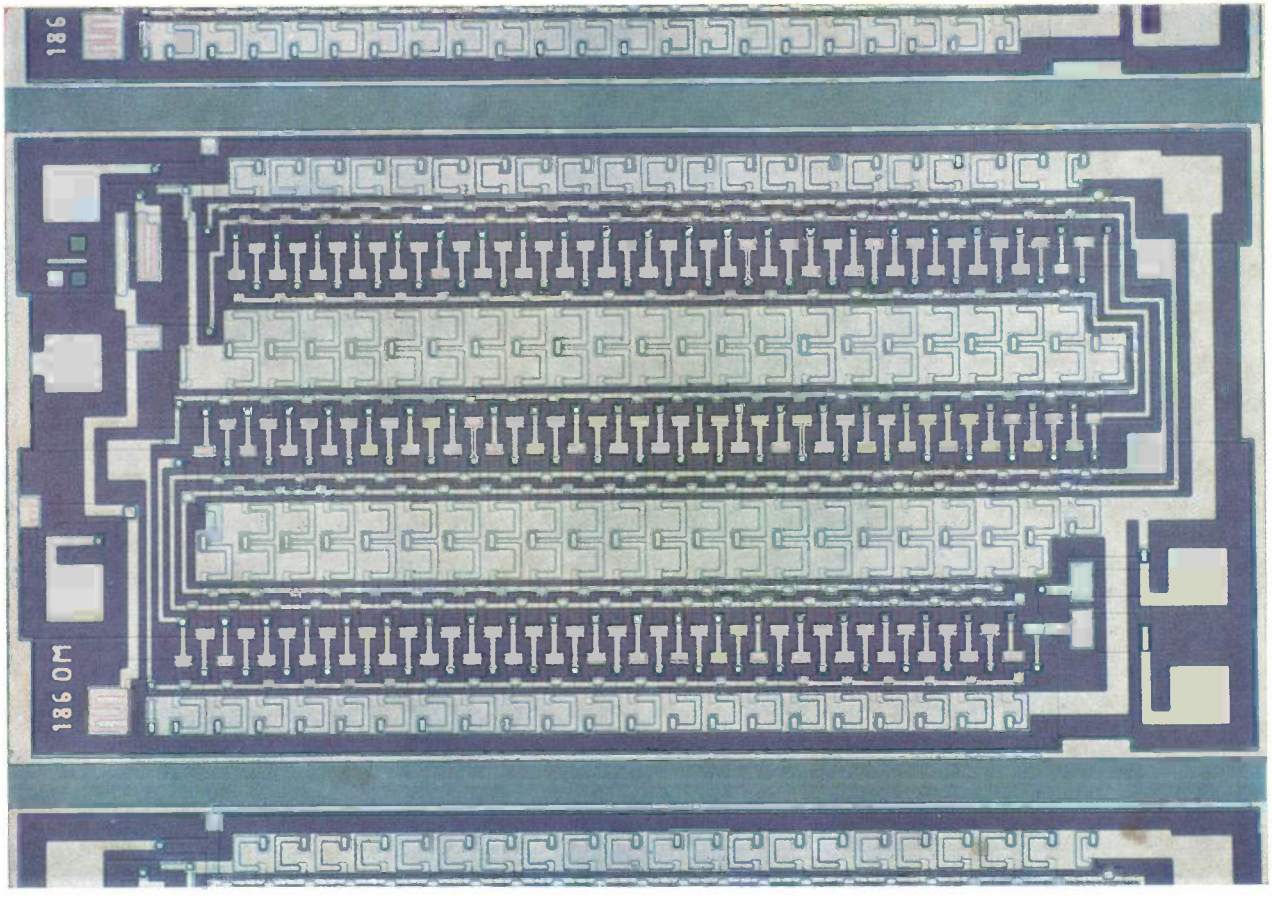
The cell works as follows. In the situation where Φ_1 is at zero level (see fig. 8*a*) and Φ_2 and Φ_3 are negative, Tr_5 passes no current while Tr_6 and Tr_7 are both conducting. Assuming that the output voltage V_o is at the zero level, then the gate of Tr_1, through Tr_6, is also at this zero voltage (i.e. C_1 is not charged). Tr_1 then passes no current, so that point A is negative, and this negative voltage appears via Tr_7 at the gate of Tr_3 (i.e. C_2 is negatively charged); Tr_3 passes current, so that V_o is held at the zero level. The two inverters are thus cross-coupled via Tr_6 and Tr_7, so that the output voltage remains unaltered.

At the instant when the shift pulse Φ_1 arrives the sequence of events is as follows. Φ_2 and Φ_3 go to the zero level, so that Tr_6 and Tr_7 no longer conduct and the cross-coupling is broken. The output voltage, however, is maintained because the capacitor C_2 is negatively charged and Tr_3 therefore continues to pass current. Since Φ_1 has become negative, Tr_5 starts to conduct. Assuming that the content of the preceding cell is 1, then C_1 will become negatively charged via Tr_5, so that Tr_1 also starts to conduct. If Φ_1 returns to the zero level, Tr_5 becomes non-conductive again, but the negative voltage on C_1 remains and Tr_1 continues to conduct. At the same time Φ_2 and Φ_3 go negative again, so that Tr_6 and Tr_7 again start to conduct. Since the leading edges of these pulses are not equally steep, Tr_7 starts to conduct somewhat earlier, so that C_2 discharges first through this transistor and Tr_1; as a result, Tr_3 stops conducting and V_o becomes negative. Meanwhile Tr_6 also starts to conduct and the cross-coupling is restored, but now with a negative output voltage, with C_1 negatively charged and C_2 discharged. The information can now be stored again for an unlimited time.

The principle of this cell is thus fairly simple. During the shifting process the cross-coupling is broken; the old information is held at the output by the capacitor C_2, which keeps the inverter Tr_3-Tr_4 in its former state. The new information is meanwhile applied to C_1, causing Tr_1-Tr_2 to assume the new state. Next, C_2 is raised to the new voltage (via Tr_7), so that Tr_3-Tr_4 also assume the new state, and the cross-coupling is then restored through Tr_6.

During the transfer of information a capacitive storage is used twice. The capacitor C_1 need only retain the charge during the leading edge of Φ_3, that is to say until the cross-coupling is restored, while C_2 must hold the charge during the time that Φ_2 and Φ_3 are at the zero level. Both capacitors can only discharge through the leakage current of the *P-N* junctions of Tr_6 and Tr_7. At room temperature the time constant of this discharge can have a value of a few tens of milliseconds; if we compare this with the conventional pulse duration of 1 to 20 μs, we see that the storage action of the capacitors is amply sufficient.

The maximum permissible frequency of the shift pulses depends on the speed at which the circuit changes state. This speed is determined by the time



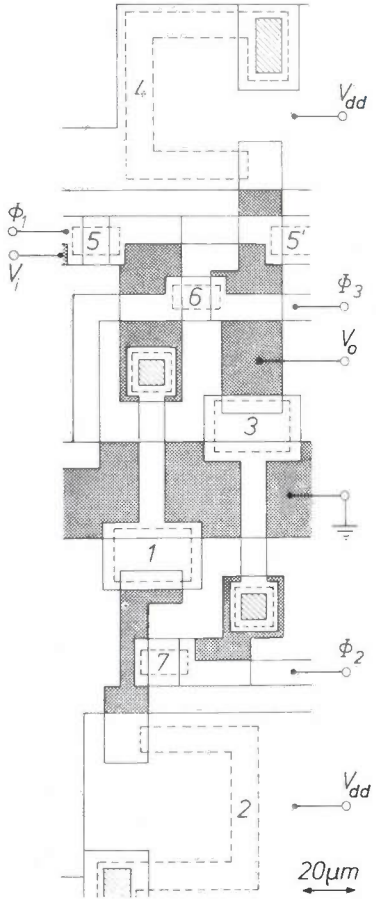


Fig. 9. *a*) A static shift register (64 bits) using the cell shown in fig. 8*a*. The circuit contains 458 components and measures 1.15 by 2.3 mm. The blue-grey strips are the scribed lines along which the crystal wafer can be broken to separate the circuits from each other. In the blue regions the upper layer consists of oxide, in the white regions it consists of aluminium.
b) Sketch of a single cell from the shift register. The areas covered with oxide are shaded; the shading is light where there is only substrate material under the oxide, and dark where there is a diffused region below the oxide. The areas where aluminium has been deposited by evaporation are left white, whatever is beneath them. The solid lines in the aluminium indicate boundaries between diffused regions and substrate material; the dashed lines indicate regions with a thin oxide film (these are mostly the isolation layers under the gate electrodes). These boundaries can also be seen on the photograph, since there is a difference in height here. The contact windows are shown cross-hatched; these are the locations where the aluminium layer makes contact with an underlying diffused region of source or drain. The various transistors in the diagram are indicated by their number, placed in the channel.

A number of these cells are placed side by side, and the various signals are applied to the strips of aluminium, which are clearly visible in the photograph. The sources of Tr_1 and Tr_3 are earthed by connecting the central diffused layer with the earthed substrate at a point which lies outside this diagram.

needed to charge C_1 or C_2. In the transition from state 1 to state 0, described above, C_1 must be charged via transistor Tr_4 of the previous circuit. During the transition from 1 to 0, C_2 must be charged via Tr_2 and Tr_7. Since Tr_2 and Tr_4 are load transistors, they have a fairly high resistance; however, when they are combined with switching transistors of minimal dimensions (C_1 and C_2 are then about 50 fF), the RC time constant for charging the capacitors can still be as small as 50 ns. In the situation illustrated, i.e. the transfer of information between two successive cells in the shift register, this gives a maximum shift-pulse repetition frequency of 1 MHz. The speed of the register as a whole is of course adversely affected by the output stage in the same way as described for the NOR gate; in practice the maximum frequency would be about 250 kHz.

The dynamic shift register

Fig. 10a shows the diagram of a dynamic shift register cell built with MOS transistors. The level of the output voltage V_o again indicates the state of the cell, and V_i the state of the preceding cell. C_1 and C_2 are the capacitances of the gates of Tr_1 and Tr_4, C_1' is the capacitance C_1 of the next cell. There are four shift signals Φ_1, Φ_2, Φ_3 and Φ_4 (fig. 10*b*), consisting of negative pulse trains; after each group of four shift pulses the input voltage is taken over by the output. In fig. 10*a* it can be seen that both Φ_1 and Φ_3 are fed to two points of the circuit. *Fig. 11a* gives a photograph of a part of a dynamic shift register in which this cell is used; fig. 11*b* again shows a diagram of one of the cells from this register.

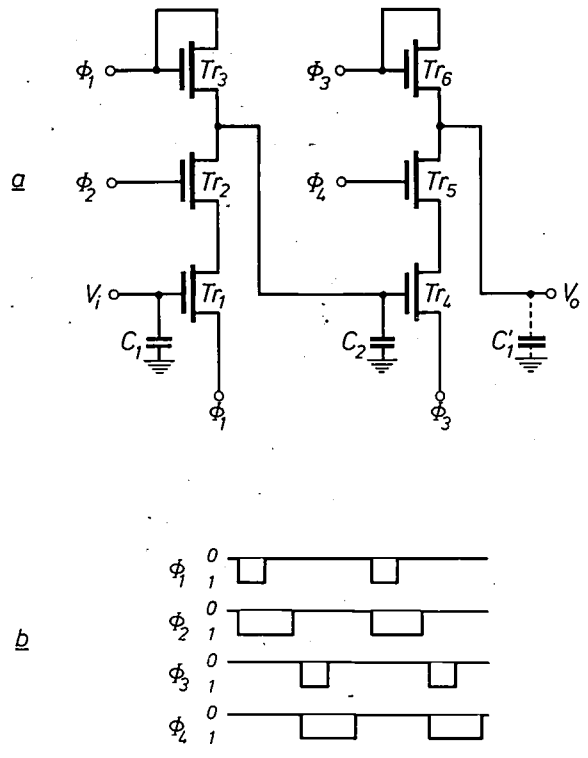


Fig. 10. *a*) Diagram of a cell from a dynamic four-phase shift register. The output voltage V_o indicates the state of the cell, and V_i the state of the preceding cell. *b*) The gate signals Φ_1, Φ_2, Φ_3 and Φ_4 consist of negative pulse trains; the information at the input is transferred to the output by each group of four of these shift pulses.

To explain the operation of the cell we start from the situation where V_i is negative; Tr_1 is then conducting. At the moment when the first shift pulses arrive, Φ_1 and Φ_2 go negative, so that Tr_2 and Tr_3 also start to conduct. From the source supplying the voltage Φ_1 the capacitor C_2 will now be charged to a negative voltage via Tr_3 and the series arrangement of Tr_1 and Tr_2. The voltage Φ_1 then returns to zero, but Φ_2 remains negative. As a result, Tr_3 stops conducting but Tr_1 and Tr_2 continue to pass current, and since the gate of Tr_1 is now at the zero level, C_2 discharges through these transistors. When Φ_2 also returns to zero, C_2 is thus nearly discharged. Now when the shift pulses of Φ_3 and Φ_4 arrive, Tr_5 and Tr_6 start to conduct but Tr_4 remains non-conducting because C_2 is discharged. C_3 is now charged (through Tr_6 only), but when Φ_3 returns to zero it can no longer discharge. Thus, when Φ_4 also goes back to zero, a negative voltage remains at the output.

If the input voltage is at the zero level, C_2 cannot discharge because Tr_1 does not go into conduction. Consequently C_2 remains negative, which then enables C_3 to discharge, and the output voltage goes to zero. We see that in both cases the voltage that was at the input is transferred toward the output after four shift pulses. Since four phases can be distinguished in this process, a circuit of this type is known as a *four-phase* shift register.

The dynamic shift register of fig. 10*a* is based on the storage action of the capacitors C_1 and C_2. These capacitors, however, discharge as a result of the leakage currents in the *P-N* junctions of the corresponding transistors, and since the dynamic circuit has no cross-coupling as in the static type, the information would be gradually lost. To prevent this, the loss of charge is regularly compensated by the signals Φ_1 and Φ_3, and this means that the whole shift-pulse pattern must be repeated at a particular minimum frequency. The register therefore continues to transfer the information. The minimum frequency at which the shift pulses have to be supplied depends on the magnitude of the capacitances and leakage currents. Since leakage currents of *P-N* junctions are involved the repetition frequency is temperature dependent; at room temperature a frequency of 40 or 50 Hz may be high enough, but for every 10 degrees increase in temperature the minimum frequency required increases by a factor of 2.

For the static shift registers MOS transistors of different dimensions are needed (switching and load transistors) in order to obtain the desired voltage division in the inverter circuits. This is not necessary for the dynamic registers, and switching transistors of minimum dimensions can therefore be used. The capacitances C_1 and C_2 in fig. 10*a* have about the same value as those in the static shift register in fig. 8*a*, but during the transfer process they are charged up through transistors whose resistance is a tenth of that of the load transistors in the static register. Because of this the maximum permissible shift frequency for the dynamic shift register is an order of magnitude higher than for the static type; it is about 10 MHz. Another consequence of only using switching transistors is that the component density on the chip is higher in dynamic registers.

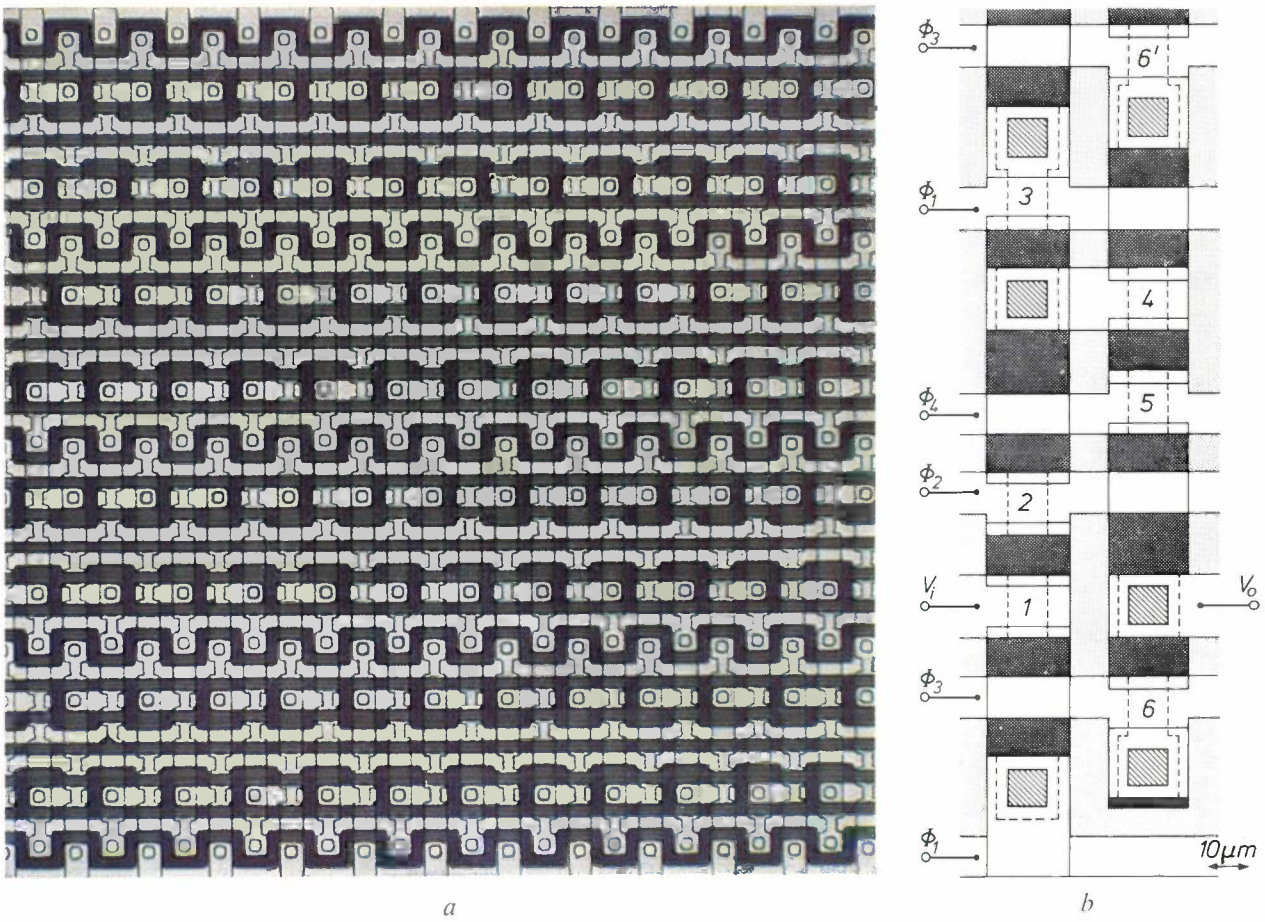


Fig. 11. *a*) Part of a dynamic shift register built up with cells of the type shown in fig. 10*a*. *b*) Sketch of one cell of the shift register. The various areas are indicated in the same way as in fig. 9*b*.

The cells are located in a row on the crystal wafer and the various voltages are applied to the aluminium strips. The photograph shows four such rows of cells one above the other, the strips for the voltages Φ_1 and Φ_3 being connected to adjacent rows; the transistor Tr_6 in the sketch therefore belongs to a cell in a following row.

The dynamic register also dissipates less heat. In the static shift register there is always current flowing through one of the two inverter circuits, and the dissipation per cell is 2 mW. In the dynamic shift register no current can flow direct to earth; energy is only dissipated when the capacitors are charged. The dissipation is therefore substantially lower than in the static case, and moreover it is proportional to the frequency of the shift pulses. The dissipation per cell is 50 μW/MHz; even at a pulse frequency of 5 MHz the dissipation per cell is therefore no more than 250 μW, which is a great deal better than the 2 mW per cell of the static register.

As we noted earlier, there is a penalty to be paid for these advantages: the information cannot be kept in the same place and a more complex combination of shift pulses is required.

Summary. The source and drain in MOS transistors are isolated by a depletion layer from the substrate, so that no isolation diffusions are required. MOS transistors can therefore be packed very densely on a crystal wafer. Since MOSTs also dissipate little heat, they are very suitable for use in integrated circuits. After a brief description of the technology, some examples of such circuits are given. The first is a logic circuit: the NOR gate, in which the MOS transistors are used as switches and also as resistors: the MOST is smaller than a diffused resistor of the same resistance. The other examples, a static and a dynamic shift register, make use of the specific advantages of the MOST: its very high input impedance and the fact that the current through a MOST can flow in both directions. A cell of the static shift register consists of a bistable circuit; the cell of the dynamic shift register is based on the storage action of a capacitor. In the dynamic shift register there is no cross-coupling to prevent the information from being lost, and the voltage on the capacitor therefore has to be continuously replenished by the shift pulses (4 pulses per cycle). This means that the information in the dynamic register must be kept moving through at a specific minimum frequency. Compared with the static type the dynamic register has the two advantages of a maximum pulse frequency that is 10 times higher (10 MHz against 1 MHz), which is possible because only MOSTs of minimum dimensions are used, and a lower dissipation (50 μW/MHz per cell as against 2 mW per cell).

A MOS transistor store with discretionary wiring

A. F. Beer

Introduction

There has been in the past few years an increasing interest in the possibility of using integrated circuits for computer storage. This interest has arisen for a number of reasons. Since integrated circuits are used for the logic operations in data-processing machines, the introduction of I.C.s into storage would reduce the present problems associated with driving ferrite-core stores from the relatively low signal levels available from I.C.s. Furthermore, integrated circuits offer potential advantages in speed, size and also cost. However, although integrated circuits can be very cheap, the costs associated with their assembly into large systems is high so that in order to make an economic semiconductor storage system it is necessary to incorporate as many storage elements (bits) as possible into one package. Such an approach is generally termed large-scale integration.

Mullard Research Laboratories have chosen a particular technique to solve some of the problems of large-scale integration in the development of a fully integrated 32-word, 32-bit-per-word store, which includes independent read and write selection matrices and has a worst-case read-write cycle of about 450 ns. The design and technology for making this store is described in this article.

The MOS transistor was chosen as the active device in the bit circuits since its simple processing should lead to high yield and high packing density. However, even with MOS transistors it is unlikely that the yield will be high enough to make a 1024-bit static store on one silicon chip without some faults[*]. Therefore it was decided to incorporate on one silicon chip more bit circuits than are finally required, test them and devise an interconnection pattern which joins good circuits on the chip and by-passes faulty ones. This technique is known as discretionary wiring. For the store described in this article an initial array of 40×40 bits is provided. The interconnection pattern joins 1024 of these appropriately to form the 32-word, 32-bit store.

When the first draft for this article was written, several small storage units had been completed and had worked successfully, but on each of the 1024-bit chips some bits failed to work. Investigation of these units showed that the failures were due to faulty contacts between the interconnection lines and the appropriate bits. The difficulty has since then been overcome.

In the next section the possible approaches to large-scale integration are briefly discussed. The subsequent sections deal with design, fabrication and testing of the circuits and the details of interconnection strategy and mask manufacture. The article is concluded with a discussion of the results so far achieved as well as some considerations concerning the expected cost of discretionary wiring compared to other large-scale integration techniques as applied to storage systems.

Approaches to large-scale integration

In the field of semiconductor storage, present-day technology enables one to make at reasonable yield a static store of about 256 bits. As techniques improve this figure will increase but the rate of increase is unlikely to be high. On the other hand, the main use for semiconductor stores will be in computers where even a small machine might require as much as 250 000 bits of storage (i.e. two thousand 128-bit units). As mentioned in the introduction the packaging of each small unit and the assembly of many packages into a large system is expensive. Therefore some new techniques are required for integration.

Two main approaches have been put forward for large-scale integration. In one, the multi-chip system, small storage units are fabricated by conventional technology, but instead of mounting each one in a package a number of units are mounted on one appropriate substrate, using some special technique such as "beam leads" or "flip chips". An interconnection pattern is provided on this substrate so that the small units are correctly joined together to form a larger store. These substrates can then be built up to form large stores using techniques similar to those involved in the assembly of printed-circuit boards. The number of bits in each unit and the number of units on a substrate are determined by the yields and costs of the processes involved. At present, substrates containing one or two thousand bits are feasible.

The second approach, namely discretionary wiring, is designed to increase the size of the store that can be

A. F. Beer, B.A.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

made on one silicon chip. This is done by providing, on the silicon chip, more circuits than are actually used, probe testing each one and then devising and applying an interconnection pattern to join the good circuits while by-passing faulty ones. In this way a large store (equivalent to the completed substrate in the multi-chip system) can be made in one package, and the amount of assembly necessary for the main computer store is minimized.

The size of each circuit (i.e. the number of bits in each circuit) and the size of the final store are determined as in the multi-chip system, by the yields and costs of the processes. For our MOST store it was decided that the final size should be 1024 bits, and that each circuit should consist of one bit only. The number of excess bits required is a function of bit yield and the interconnection scheme available. In our case, an initial array of 40×40 bit circuits was found to be satisfactory.

A great deal of discussion has been recorded in the literature on the relative merits of the multi-chip, discretionary-wired and fixed-wired systems [1]. This paper will not add to that material beyond pointing out that because of the simple repetitive nature of the store, the computation and generation of an interconnection pattern for discretionary wiring is here quite straightforward. As a result the costs of the processes are expected to be fairly low and a fully competitive product appears to be possible.

The circuits

The bit circuits

Although the use of discretionary wiring leads to less stringent requirements on yield, a relatively high bit-circuit yield is still needed as will be seen in the section on the interconnection algorithm. Furthermore, because of the excess circuits provided on the chip, discretionary wiring requires more silicon area than other systems. Therefore, it is desirable to employ processes leading to both high yield and high packing density. This leads directly to the choice of a single type of MOS transistor as the active device since it has much simpler processing and higher packing density than either bipolar transistors or complementary MOS transistors (i.e. combinations of *N*- and *P*-channel MOSTs in one circuit). Because the *P*-channel MOS transistor is normally of the required enhancement type it was chosen in spite of its lower gain compared to *N*-channel MOSTs.

The circuit chosen for the storage cells is shown in *fig. 1*. It is made up of two cross-coupled inverters consisting of two active transistors (T_1 and T_4) and two load transistors (T_2 and T_3). An extra transistor

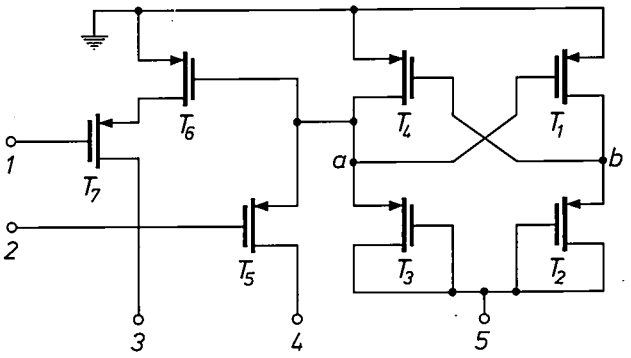


Fig. 1. The storage-cell circuit ("bit circuit"). It consists of two cross-coupled inverters (MOS transistors T_1-T_4), a MOS transistor for writing into the cell (T_5) and two MOS transistors for reading (T_6 and T_7). *1* word-read line, *2* word-write line. *3* sense line. *4* digit line. *5* supply-voltage line.

T_5 is provided for writing into the cell, and two further transistors, T_6 and T_7, are used for reading-out.

The gains of the active and load transistors have to be proportioned to give the appropriate logic levels, dissipation and speed. If node *a* is set to the "one" state, its voltage is $-V - V_{th}$, where $-V$ is the supply voltage and V_{th} is the transistor threshold voltage with substrate bias. This must be sufficient to make transistor T_1 conduct hard enough for the voltage at node *b* to be more positive than $-V_{th}$, i.e. T_4 is off. A ratio of about 10 : 1 in the gains of the active transistors (T_1, T_4) to the gains of the load transistors (T_2, T_3) is adequate for this. Having fixed the gain ratios of the transistors the absolute values of the gains are then adjusted to give acceptable speed and dissipation.

Setting of the bit is achieved by applying a negative pulse to the gate of T_5 (word-write line) and thereby forcing node *a* up to zero ("0" state) or down to $-V - V_{th}$ ("1" state) by the appropriate voltage on the drain of T_5 (digit line). The time required to set the cell is determined by the rate at which the capacitance attached to node *b* can be charged to $-V-V_{th}$ through the load MOS transistor (T_2) or discharged to zero through the active MOS transistor (T_1). Since T_2 has substantially lower gain than T_1, the negative-going transition of node *b* will be the slower — i.e. setting the cell in the "0" state. The capacitance at node *b* is made up of the gate capacitance of T_4, the drain capacitance of T_1 (or source of T_2), some portion of the channel capacitance of T_2 and some stray capacitance due to aluminium connections.

The dissipation of the circuit is determined by the "on" resistance of the load transistors. When this resistance is increased, the node *b* capacitance (exclud-

[*] This article is concerned only with static storage. Using dynamic-circuit techniques, 1024-bit units are available, but these techniques are not considered here. (Cf. the previous article in this issue.)

[1] See for example: R. L. Petritz, Current status of large scale integration technology, IEEE J. Solid-State Circuits SC-2, 130-147, 1967, and D. A. Hodges, Large-capacity semiconductor memory, Proc. IEEE 56, 1148-1162, 1968.

ing strays) is decreased if the gain ratio is maintained by reducing the length of the devices rather than increasing source-drain separation. Hence the speed of the circuit remains constant as dissipation is varied until the transistors become so small that the stray capacitances start to dominate. The time constant of node *b* starts to increase at about 3 mW dissipation and by 1 mW is increasing extremely rapidly. The bit has been designed to operate at 2 mW where the rise time has increased only a small amount. Node *b* then requires about 150 ns for a negative-going transition, while the opposite transition takes place in about 50 ns.

The slow transition could be improved by incorporating a further transistor in the circuit similar to T_5 but connected to node *b* and with its gate coupled to that of T_5. Then both nodes could be forced up or down through these low-impedance transistors and only 50 ns would be required for either transition. However, such a circuit would require an additional digit line leading to increased complexity in interconnections.

Reading of the cell is accomplished by current-sensing off-chip. The current derives from the buffer amplifier T_6 and the gating transistor T_7. This minimizes direct loading of the bistable and offers a reasonably fast non-destructive read-out. Since it is desirable to use bi-polar transistors for sensing this current, they have to be supplied off-chip and may serve for other store planes as well in a main store of many planes.

The circuit has been simulated on an Elliot 503 computer [2]. This has shown that with the circuit chosen, wide tolerances of transistor parameters are permissible. The circuit will operate reliably if the factor β [3] is in the range 0.34 ± 0.07 A/V²m and if the threshold voltage V_{th} is in the range -3.5 ± 1.5 V, provided the supply voltage is at least a factor of three greater than V_{th}.

The selection matrices

If the chip containing the 1024 bits is mounted in a package as a simple 32-word 32-bit store, a total of 130 leads must be provided between the chip and the package: 32 each of read, write, sense and digit plus earth and supply. This is clearly unsatisfactory but can be improved while still retaining the word organization of the store if selection matrices are provided on the chip to reduce the number of word-read and word-write connections. However, it is then necessary for the processing of the selection matrices to be compatible with that of the MOS transistors, and the simplest means to achieve this is to use MOS transistors as the active devices in the matrices as well, even though this is likely to lead to slower speeds than bipolar devices.

A number of matrices have been studied. The complete binary selection system shown in *fig. 2* reduces 32 word-write or word-read lines to one input and ten gate leads, but because there are five MOS transistors in series it is very slow. This matrix can be improved by increasing the branching at two of the levels to four each, so that there are only three branching levels ($2\times4\times4$) and hence only three transistors in series. This gives a delay time of about 450 ns which is still rather long.

The matrix finally chosen is shown in *fig. 3*. It requires one more lead (twelve in total) than the complete binary system and is an incomplete matrix in that some additional selection off-chip is needed to separate the eight input leads. However, in a large system requiring several store planes this can be coupled with the selection required between store planes.

The complete store has independent read and write matrices. A complete read-write cycle occupies about 450 ns. Reading alone, including selection, requires about 150 ns.

Silicon processing

The basic processing is largely conventional. The silicon is 3 Ωcm N on N^+ epitaxial material oriented in the ⟨100⟩ direction. The N layer is 10 μm thick and the N^+ substrate is 150 μm thick. The epitaxy is required to ensure a low resistance through the material as the earth line is the common substrate.

A thick (800 nm) steam-grown oxide is provided to mask off the diffusions. The source and drain regions are boron-diffused from oxidized slices of boron nitride. After reoxidizing the boron-diffused windows and opening new ones, a phosphorus diffusion is provided to ensure good ohmic contact to substrate of those regions which are earthed. At the same time the phosphorus diffusion provides channel stoppers in those regions where parasitic MOS transistors might otherwise occur.

After the boron and phosphorus diffusions, the gate regions and source and drain contact areas are stripped of the thick oxide, and a new oxide layer is grown at 1200 °C in a double-walled tube in very dry oxygen. Immediately following the oxidation (1 hour giving 200 nm oxide), the gas is changed to dry nitrogen for a five-minute anneal at the same temperature. After cooling a phosphorus deposition drive-in is carried out for stabilization. The resulting phosphorus glass is about 15 nm thick.

Contact windows are opened in the thin oxide using double photoresist methods, and aluminium is evaporated using an electron-bombarded source. Following definition of the aluminium, the slice is annealed at 450 °C in wet nitrogen in order to remove excess surface states.

A second dielectric layer is then deposited on top of

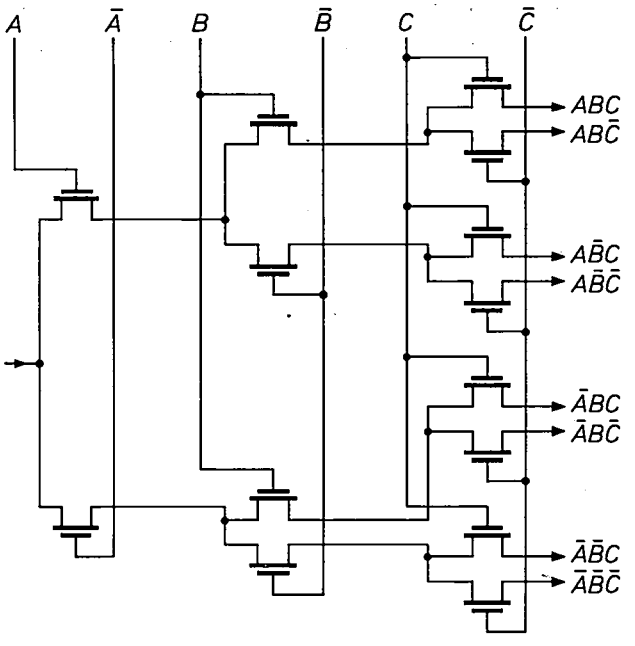


Fig. 2. Three of the five stages of a complete binary selection circuit for the reduction of 32 word-write or (word-read) lines to one input. There are ten gate leads (A, $\bar{A}$ etc.) in the complete matrix. A system like this is very slow.

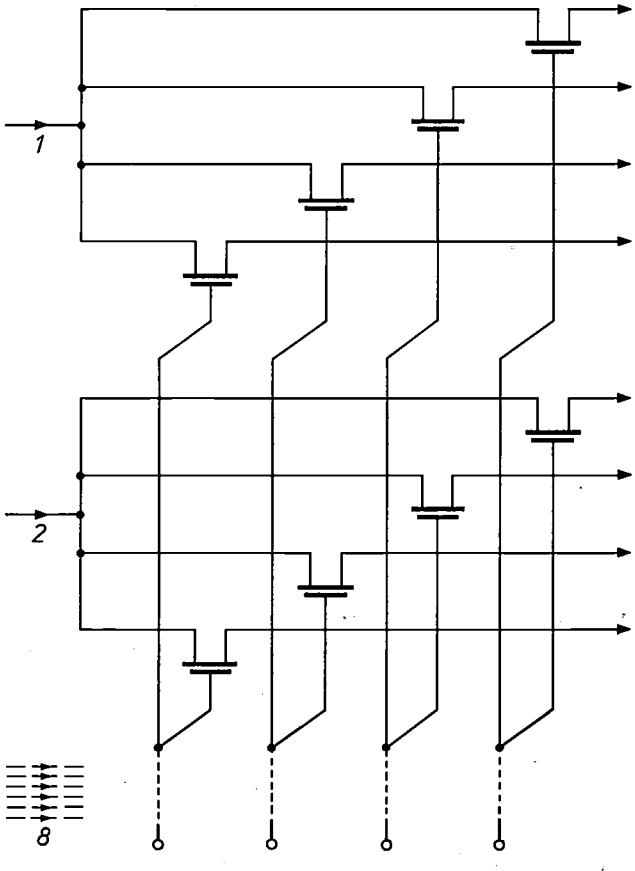


Fig. 3. The selection circuit applied in the MOST store consists of eight equal groups of four MOS transistors each. Two of these groups are shown. The total number of input leads is 12, i.e. one more than in the complete binary circuit shown in fig. 2.

the slice. At present, the preferred dielectric is SiO_2, deposited from the reaction of SiH_4 and O_2, but r.f. sputtered SiO_2 has also been used.

Holes are etched in the top oxide to the aluminium underneath, after which a second layer of aluminium is evaporated and probe pads defined. *Fig. 4* shows a part of a slice processed to this stage and a cross-section through one bit circuit. Each bit circuit occupies about 270×220 μm and is repeated at 300 μm steps in both directions. It can be seen that probe pads are applied to the word-write, word-read, digit, sense and supply-voltage lines, the earth line being completed through the substrate. The digit, sense and supply-voltage lines are each provided with two contacts to the top region. Only one is used for probing but both are needed for interconnections subsequently.

When the good circuits have been determined and the interconnection mask made, aluminium is re-evaporated and the interconnection pattern defined. This is an extremely critical operation since any fault occurring after testing usually means the rejection of the entire chip.

A number of techniques are being examined in order to improve the interconnection yield — that is, the yield of chips free from interconnection defects. Present results suggest that it will not be impossible to use aluminium alone but a higher success rate may be achieved by the use of two metals, such as molybdenum and gold, evaporated one on the top of the other, and defined in different etches. This will tend to reduce the incidence of short circuits between adjacent lines without increasing the incidence of open circuits in any one line.

Testing

In a production system it would be sufficient to carry out a series of go/no go tests on each element of the store and to use the results as the basis for the interconnection pattern. Such a procedure could be very fast. For example, by having two probe machines alternately driven by one tester, and having the probes connected simultaneously to four bit circuits at each measuring position, two complete 40×40 arrays could be measured in about two minutes. At the present stage of development however it is necessary to note the results of each measurement and analyse each fault in order to adjust the processing of the slices to give the best result.

The system used at present consists of a high-speed probe-test machine coupled to a data logger. Each

[2] L. F. Gee, D. W. Parker and P. Swift, An M.O.S.T. store, Information Processing 68, Proc. IFIP Congress, Edinburgh 1968, Vol. 2, pp. 778-782.

[3] The significance of the factor β and the voltage V_{th} is discussed in the article by J. A. van Nielen in this issue, page 209.

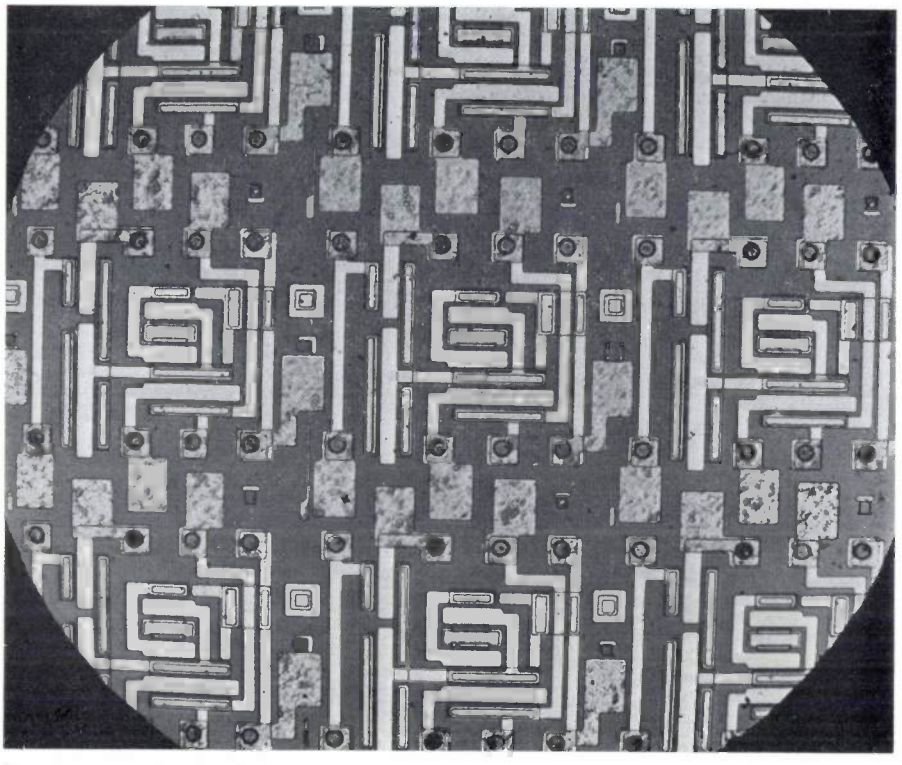


Fig. 4*a*

Fig. 4*b*

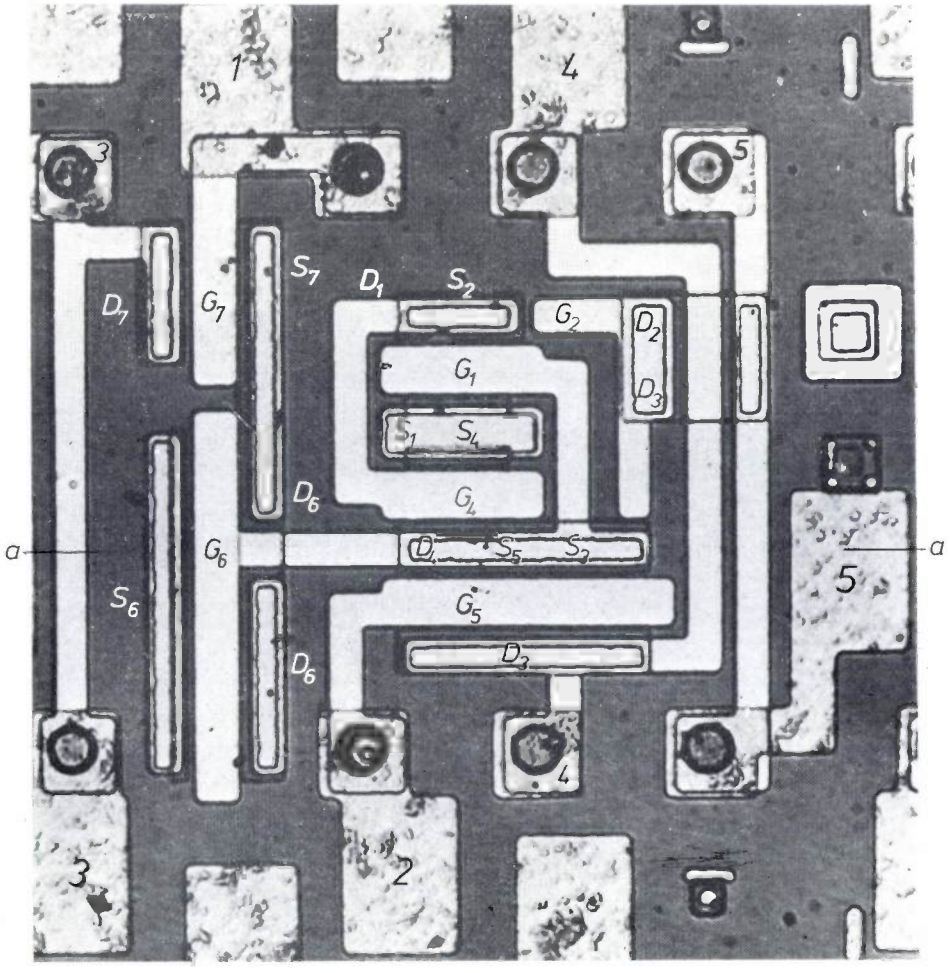


Fig. 4. *a*) Photomicrograph of a part of a silicon slice ready for probing the bit circuits. Nine of these circuits are entirely or partly visible. Each of the five connections of every circuit is provided with a probe pad. Magnification 130×.

b) A part of (*a*) extending over one bit circuit. *S*, *G* and *D* are the source, gate and drain of the different transistors; the subscripts and the numbers on the probe pads correspond with those in fig. 1.

c) Cross-section through one bit circuit (not to scale) along the line *a-a* drawn in (*b*).

O_1, O_2 and O_3 are the different oxide layers, *Ph* is a protecting phosphate-glass layer and *St* a channel stopper.

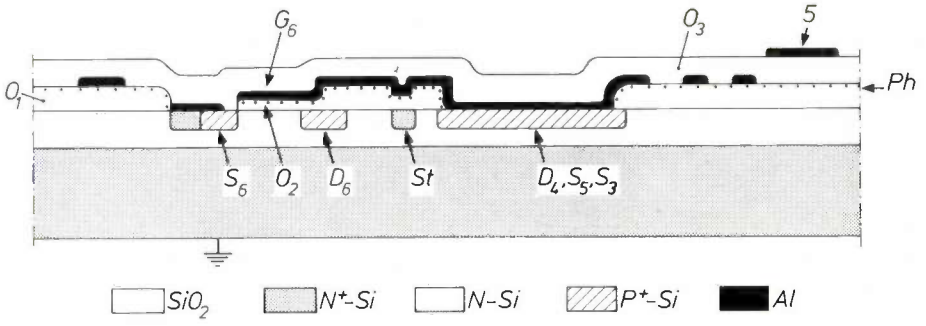


Fig. 4c

device is tested and the results recorded on punched tape by the logger. The completed data tape is then analysed by computer. Mean values and deviations are calculated for each measurement. These are then recorded on the print-out along with a slice map. On the slice map the result of the measurements on each device is recorded in a spatial position corresponding to the position of the devices on the silicon slice.

The measurements made on the bit circuits are basically functional. The digit voltages required to set the bit in the "1" and "0" states are measured along with the bit current and sense output current in each state. The leakage currents to the substrate of the sense, digit, word-write, word-read and supply lines are also measured. The measurements are checked against a specification, and the slice map and the interconnection pattern produced from the result.

Although these measurements are satisfactory for testing bit circuits, it is not always possible to determine the cause of failure from them. For this reason, it is convenient to fabricate discrete transistors using the same processes as for the bits. From the measurement on these transistors one can deduce the values of the transistor parameters in the circuits. In the early stages of development large numbers of arrays of such discrete transistors were processed on separate slices, and then measured and analysed in order to find the optimum processing conditions. Latterly the transistors provided for the selection matrices have been used satisfactorily for this purpose.

The interconnection algorithm

The results of the tests are used to derive an interconnection pattern which will join good bits but bypass bad ones. The basis of the algorithm [4] which determines this pattern is described briefly in this section.

As mentioned already the storage module is arranged as a word-organized unit of 32 words each of 32 bits. Hence the word-write and word-read lines have to be connected to all 32 bits in one word, while the sense and digit lines are connected to the corresponding bits in each word. The supply-voltage line is common to all bits. For convenience the word-write and word-read lines are considered to run horizontally and the sense and digit lines vertically (i.e. from the top to the bottom of the page).

The first step in the algorithm is to reject any column of bits in which there are more than eight faults and then to put in the vertical lines in the remaining columns. These lines by-pass faulty bits but do not deflect from one column to the other. Then, starting at the top left-hand corner, the horizontal lines are put in one word at a time. As the lines move across from column to column, they are allowed to deflect up or down so that each non-rejected column contributes one bit to each word. The deflections are limited in that for each move from one column to the next, the lines may move up or down by one step only. The program is arranged so that a deflection upwards is tried first and downwards tried last.

As the lines move across they may reach a point where a single deflection is insufficient to get around a faulty bit. In that case one or more previous moves are erased, and deflection put in where allowed in order that the faulty area can be by-passed. This approach means that it will often be necessary to by-pass good bits in order to avoid some faulty ones.

One result of this is that the last good bit in a column may be used for a word other than the last (i.e. 32nd) word. There is not then a usable bit available in that column for the next word. The computer program is arranged to wipe out the entire pattern and then reject that column. Vertical connections are again put in and the horizontal ones started. This procedure continues until 32 horizontal lines can be made or more than eight columns are rejected in which case that array is rejected. The entire procedure requires between 30 and 45 seconds of computing time on an Elliot 503 computer.

The usefulness of the algorithm has been checked by simulating random faults in arrays of various sizes

[4] This algorithm was designed by D. W. Parker of Mullard Research Laboratories. For a more extensive description see: D. W. Parker, Discretionary wiring of a MOST storage system, IEE Conf. Publ. No. 30, Integrated Circuits, pp. 192-208, 1967.

by computer, and then calculating the interconnection pattern for each array. *Fig. 5* shows the yield of completed 32×32 arrays as a function of initial array size with element fault rate as a parameter. It can be seen that over 99% of initial 40×40 arrays can be interconnected to form 32×32 stores when the bit-fault rate is 10%. Thus, although 540 extra bits are provided, on average only 160 can be faulty if a high yield of completed stores is to be obtained.

A variant of this algorithm has been tried in which the horizontal lines may deflect up or down by two or one steps. As expected, this gives a less stringent requirement on fault rate but in general requires slightly more silicon area between bits. At present this is the preferred version. The results of this algorithm are shown by the dashed lines on fig. 5.

The program is arranged in such a way that it will try to find an interconnection pattern for any size of array up to 40×40. This enables one to get a smaller array pattern successfully when the fault distribution is such that a full 32×32 is unsuccessful.

In practice, when a complete 32×32 store with selection is required, the discrete transistors which will form the selection matrices are considered in the same way as the bits. One column of 40 transistors is provided for the word-write matrix on one side of the bit array and a second, identical, column for the word-read matrix on the other side. The program then produces a 34×32 interconnection pattern with the constraint that the two outermost columns must be used.

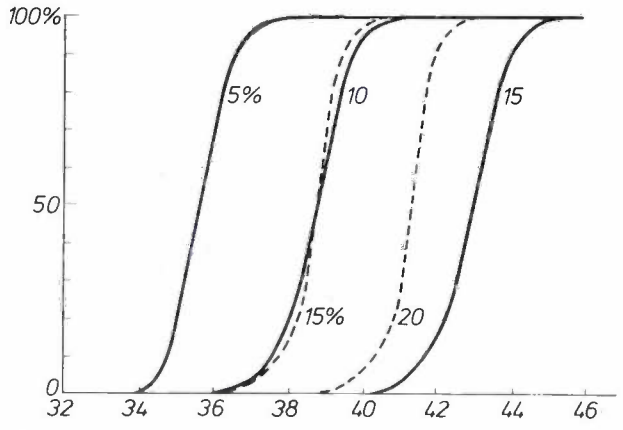


Fig. 5. The yield of completed 32×32 arrays as a function of the initial array size with the element fault-rate as a parameter. The solid lines apply for an interconnection algorithm in which the word-write (word-read) lines may deflect up or down by one step only. The dashed lines refer to an algorithm in which these lines may deflect by one or two steps.

Interconnection-mask generation

The interconnection pattern, designed by the method described in the preceding section, has now to be made into a mask suitable for defining aluminium on the silicon slice. Since each slice will in general require a different mask, its generation must be fast and inexpensive. An opto-mechanical line-writing machine [5] has been developed and appears to meet these requirements.

The basic principle of this machine is illustrated in *fig. 6a*. A slit in an aperture plate is imaged by a series of light pulses on to a photographic plate mounted on a moving table. The light pulses are synchronized to the moving table in such a way that an overlap of the image occurs on the photographic plate. As a result a continuous line can be drawn in the direction of table movement.

Horizontal lines (that is, normal to the direction of table movement) can be built up by a series of slits, each with a shutter as shown in fig. 6*b*. As the table moves

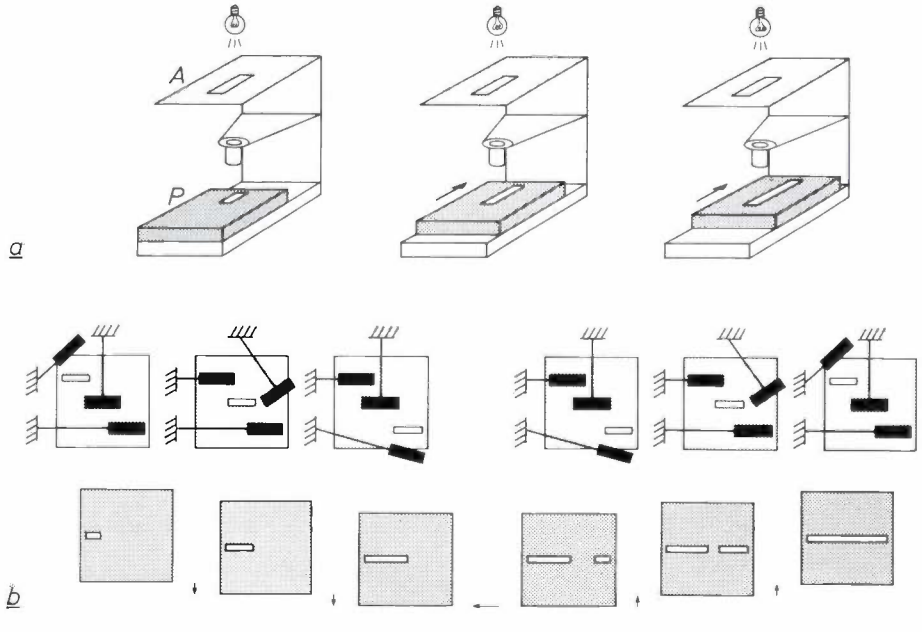


Fig. 6. Schematic representation of the generation of the "vertical" (*a*) and "horizontal" (*b*) lines of a mask suitable for defining of the interconnection pattern. A reduced image of a slit in an aperture plate *A* illuminated by a pulsed light source (*a*) or by opening a shutter (*b*) is formed on a photographic plate *P* mounted on a moving table.

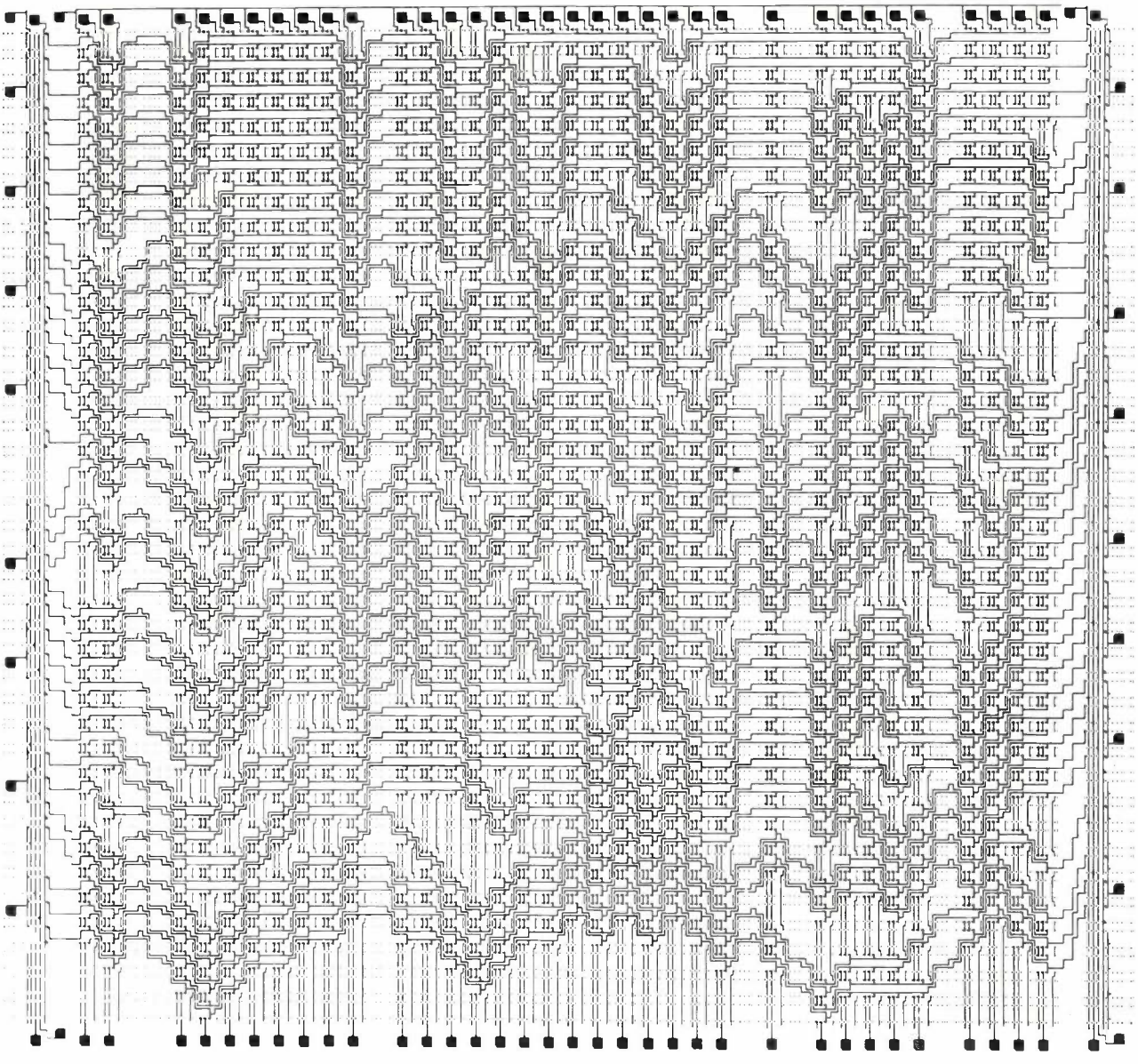


Fig. 7. Mask based on a single-deflection algorithm for defining the interconnections of a 32×32 store, including selection matrices and bonding pads.

down an image is formed by one slit at a time. At the end of travel of the table, it indexes along and moves back in the opposite direction (up in the diagram). The slits are again imaged one at a time so that horizontal lines are built up.

The shutters are operated by means of piezo-electric bilaminar flexure elements ("Bimorphs") which deflect on application of voltage. The bimorphs and the table movement are driven by a punched tape produced by the interconnection program.

By suitable positioning of horizontal and vertical slits, combinations of lines are built up as the table moves up and down and indexes along. For the store, two horizontal lines (word-write and word-read) are required. These must be capable of deflecting vertically in order to avoid faulty areas. The store also requires three vertical lines — sense, digit, and supply voltage — which either make contact to good bits or go over bad ones by-passing the contact areas. The supply lines, which are organized vertically, are finally joined together to one bonding pad. The sense and digit lines all go to separate pads while the word-write and word-read lines are taken through the selection matrices to bonding pads. All the pads are made by the machine as well.

A complete 32×32 interconnection mask based on the algorithm described in the previous section is shown in *fig. 7*. This mask includes selection matrices

[5] This machine was designed by I. H. Lewin of Mullard Research Laboratories.

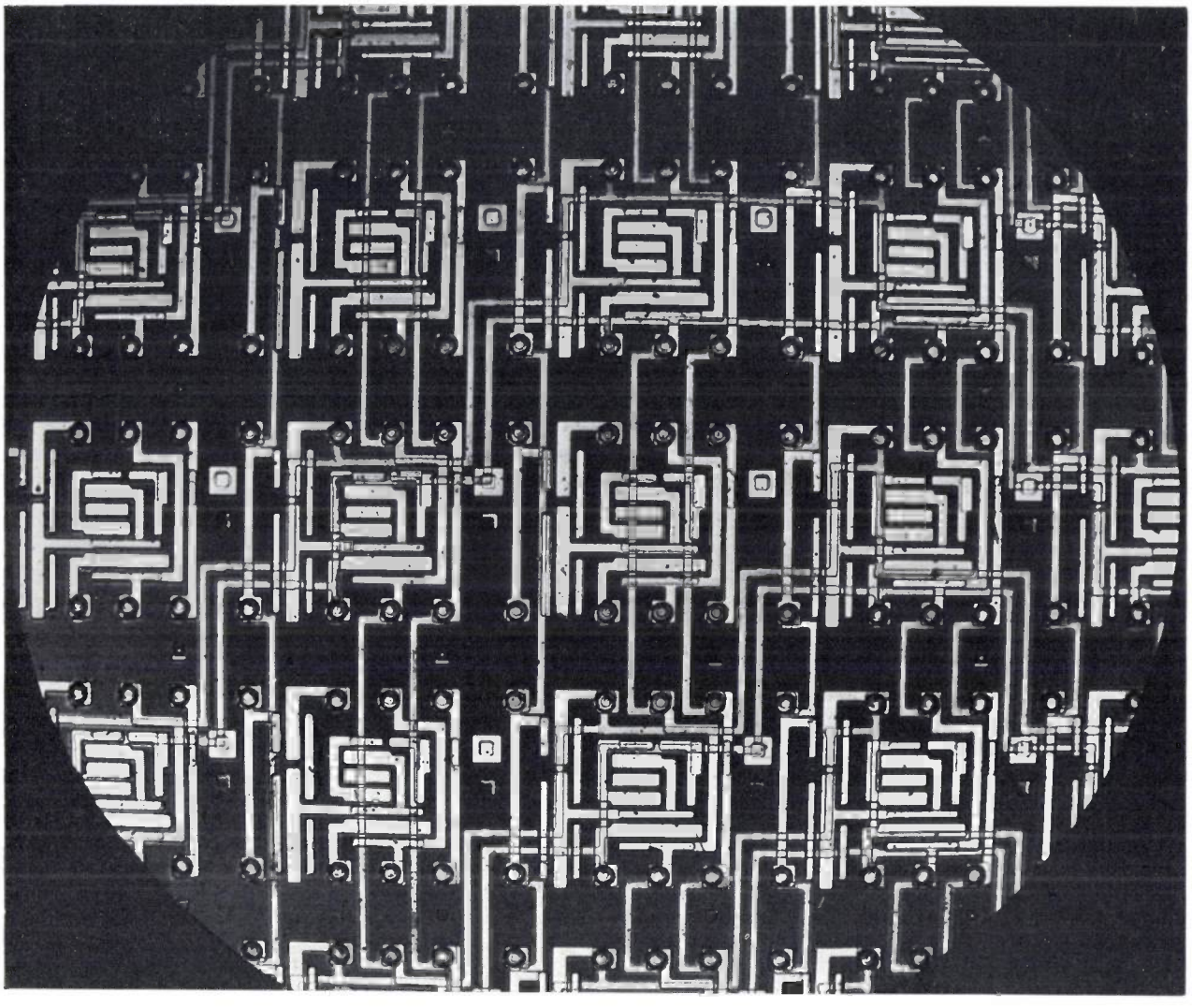


Fig. 8. Photomicrograph of a part of a slice on which the interconnections have been defined. Magnification 130×.

and all necessary bonding pads. The regular array of eight dots to be seen on the figure are the points where contact is made on the second oxide to the metallization underneath.

Once the mask has been generated, it is used in the normal way to define the interconnections on the top oxide. It then only remains to mount and bond the slice to an appropriate header before final testing. *Fig. 8* shows detail of the interconnections defined on the slice.

Results

All the arrays of discrete MOS transistors, made for control purposes, have been tested and the results analysed. The MOST arrays are normally 32×32 with two transistors in each position so that 2048 devices are measured on each slice. A good array typically has a yield of over 90% of working devices, and about 85-90% yield of devices within specification. By examination of the slice map for each array, it has been found that in general catastrophic faults are randomly distributed even amongst closely grouped devices. It should, however, be noted that the extreme edges of the slices are not used. Working devices out of specification are somewhat more likely to occur in groups. Some of these can be accounted for by variations in oxide thickness and oxide charge, but the majority and virtually all the catastrophic faults arise from photolithographic errors.

Substantial numbers of arrays of bit circuits have also been made. Neglecting those slices where gross defects are apparent, the remainder have typical bit yields of between 60 and 85%, with about 20% of the slices in the range 70 to 85%. Extensive analysis of these arrays has shown that the majority of faults are due to incorrect photolithographic processing and faulty areas on the masks themselves.

In the introduction it was mentioned that in the early 32×32 stores some of the bits were inoperative because they did not make contact with the intercon-

nection lines. However, the interconnection lines were found to be free of faults, and they can be aligned to the slice with an error of less than 3 μm at any point. Detailed examination has shown that the interconnection lines on the top oxide can become open-circuited due to the sharp steps in the oxide. The difficulty has been overcome by careful attention to the evaporation of the aluminium and to the routing of the lines so that the worst areas are avoided.

The electrical behaviour of the working portions of these stores have been investigated and the measurements largely confirm the results predicted by the computer simulation. The main difference is that the slow transition (from "1" to "0") is some 50 ns faster than predicted.

As the bit yield is now sufficiently high to give some full 1024 stores, the main technological problem that remains is concerned with the deposition and definition of the interconnection pattern, since any fault introduced at this stage will result in the rejection of the entire chip.

In order to obtain the highest possible yield of interconnections, the processing is arranged so that the minimum possible disturbance is caused to the bits after probe testing. Nevertheless, some faults will continue to arise.

It is as yet too early to predict the interconnection yield with any certainty, but the experiments so far carried out suggest that the definition of interconnections is comparable in difficulty to any one of ten discrete steps in producing a circuit. This leads directly to the result that the interconnection yield is given by $Y^{0.1n}$, where Y is the yield of individual storage-cell circuits and n is the number of cells to be connected.

On the basis of the results obtained so far on bit yields, the known costs of silicon processing and the interconnection-mask generation and with the help of the expression for the interconnection yield just mentioned, extensive cost analyses have been carried out to find an answer to the crucial question of whether discretionary wiring does offer a cost advantage over other approaches.

The cost study predicts that for any given yield of individual bits there is a specific size of encapsulated final array that gives a minimum cost per bit. As one might expect, both the minimum cost and the array size corresponding to that cost are a function of the wiring strategy used. The discretionary-wiring approach leads to a lower cost per bit and a larger array at that cost than a fixed-wiring approach as long as the bit yield is less than about 99.5%.

These cost indications are based on the system described in this article. However, once the bit yield is reasonably high a modified discretionary-wiring strategy gives a further reduction in cost. At high bit yields the testing and subsequent interconnection of bits one at a time is wasteful. It then becomes practicable to test a unit which is larger than one bit and accept or reject this whole unit. The testing, pattern generation and pattern definition all become simpler and some silicon area is saved as well since the bits can be packed more tightly when they do not need to be separately tested. On the other hand, some loss of yield will result since one faulty bit in a unit causes the rejection of that entire unit. Nevertheless, our cost studies suggest that a 16-bit unit discretionary system leads to a more economical product than the single-bit system once the bit yield is about 95% and it is cheaper than the fixed-wired system until this yield exceeds about 99.9%. This latter figure is likely to be beyond our capabilities for some years.

As mentioned in the beginning of this article, there exists an alternative method of making large arrays by using a multi-chip approach in which smaller fixed-wired arrays are mounted on to an insulating substrate on which a fixed interconnection pattern has been provided. The costs in this system depend on the testing, mounting and bonding of the chips onto the substrate. There is not yet sufficient information available on the techniques involved to be able to predict the costs of the multi-chip system relative to discretionary wiring.

There seems, however, little doubt that the discretionary-wiring approach based on techniques described in this paper offer a simple and economical means of achieving large arrays of interconnected elements such as bits in a computer memory.

The author wishes to thank the United Kingdom Ministry of Technology for permission to publish this paper, which results from work supported by a normal cost-sharing contract under the Ministry's Advanced Computer Technique Project.

Summary. The ever increasing scale on which integrated circuits are being used for the logic functions in computers makes it desirable that integrated circuits suitable for stores should also be available. These circuits are inexpensive but combining them to form larger units is not. It is important therefore that the circuits should be made as large as permissible. An MOS store now in the process of development has a capacity of 32 words, each of 32 bits, with independent selection circuits for reading and writing and with a maximum read-write cycle of 450 ns. It is made by forming a matrix 40 × 40 storage elements each consisting of seven MOS transistors, on a silicon chip, testing these elements one at a time and then connecting them in a pattern that leaves out the faulty elements (discretionary wiring). The yield of operating elements on any one chip is now between 60 and 85% (between 70 and 85% for 20% of the chips), which is sufficiently high to enable a 32 × 32-bit store to be made up from a number of them. Making the connections is no more difficult than any of the ten steps needed to make the circuits. Once the yield of good elements per chip exceeds 95%, it will be unnecessary to check the elements individually. As long as this proportion is under 99.9%, stores with discretionary wiring will probably be cheaper to produce than those with fixed wiring.

Volume 31, 1970, No. 7/8/9 pages 205-296 Published 1st July 1971

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No. 10

Characteristic luminescence

During the past five years many new phosphors have been discovered that have found immediate application in the fields of lighting and television. A common property of these phosphors is that the luminescent centre is an ion of one of the rare-earth metals. A considerable number of the new phosphors were discovered at Philips Research Laboratories.

The discovery of new phosphors has greatly stimulated fundamental research into the numerous puzzling aspects of phosphors, concerning for example the colour and efficiency of the emission. Up to a few years ago, for instance, there was no explanation for the fact that the colour of the emission from Eu^{3+} phosphors is orange or red, depending on the host lattice. At first sight this is baffling, because the emission is caused by electron transitions between deep-lying atomic orbits upon which the host lattice has little or no influence. Since compounds containing Eu^{3+} ions are used as red phosphors in colour television screens, it is of great practical importance to understand this phenomenon.

As regards the efficiency of phosphors, a distinction must be made between the luminescence where the absorption and emission both take place in the same ion or group of ions, and the luminescence where the emitting centre is other than that at which the energy is absorbed. In the first case it is possible to identify the ions or groups of ions that will in principle give emission and the host lattices in which they will give the highest efficiency. This identification is still based on empirical rules, but they can be shown to be theoretically reasonable with a surprisingly simple model. The theory underlying the transfer of energy in the second case is well established, and this understanding can be successfully used to predict the efficiency of phosphors.

The present article, consisting of three parts, presents much new insight and data gathered mainly in the last five years. Contrary to our practice in the past, these parts are not published in separate numbers of our Review but form the complete contents of the number before you. The editors wish to emphasize in this way the educational aspect of articles summarizing recent advances in science and technology.

Characteristic luminescence

G. Blasse and A. Bril

I. The absorption and emission spectra of some important activators

Introduction

Phosphors are materials capable of emitting radiation when subjected to ultraviolet radiation, X-rays, electron bombardment, friction or some other form of excitation. This emission is known as luminescence [1]. In a tubular fluorescent lamp, for example, the energy of the mercury line at 253.7 nm is converted into radiation covering the whole visible region. In a television tube the energy of fast electrons is converted into visible radiation.

In the course of the years a great deal of research has been devoted to the colour and the quantum efficiency of the luminescence of phosphors, two factors which obviously dictate to an important extent the usefulness of a given phosphor. In recent years great strides have been made in this field. The present article gives a broad review of the fundamental research done on phosphors that show characteristic emission. In such phosphors the emission comes from luminescent centres as a result of an electron transition that, in principle, would also be possible if the centre were situated in free space instead of in a crystal lattice. Nevertheless, as will be seen, the crystal lattice does play an important part.

The physical processes involved in the phenomenon of characteristic luminescence are presented schematically in *fig. 1*. The figure shows part of a crystal M in which two kinds of foreign ions or ionic groups (centres) are incorporated. One centre of each type is shown, marked A and S. We assume that the host lattice, that is a crystal without these centres but otherwise of the same composition, absorbs no radiation. The centre in the right half is raised to an excited state as a result of radiation absorbed by that centre. The centre returns to the ground state by giving up the excitation energy as radiation or as heat. The former case is referred to as luminescence, and the centre involved is called an activator.

It is also possible to excite the activator A by indirect means. If, for example, we want to excite a phosphor with the 253.7 nm radiation from a mercury lamp, and if A does not absorb this radiation, the excitation can nevertheless take place via the centre S which does absorb this radiation. In some cases the host lattice itself plays the part of S. The excited centre S can return to the ground state in three ways: by radiation, by the dissipation of the excitation energy in the form of heat, and by transfer of the excitation energy to A. In the latter case the excitation energy absorbed by S, or part of it, is emitted by A. S is then often referred to as a sensitizer of the luminescence from A, although it may itself also act as activator.

In the first part of this article we shall examine the possible excitation and emission transitions and the influence which the crystal lattice can exert on these transitions. In the second part we shall deal with the quantum efficiency of the activator for direct excitation

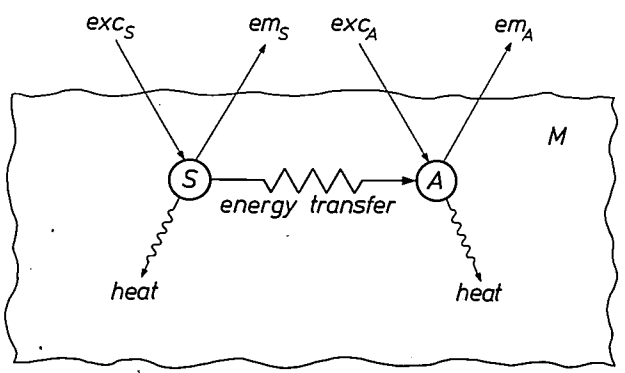


Fig. 1. Diagrammatic representation of luminescence. Incorporated in a host lattice M are an activator A and a sensitizer S. The host lattice does not absorb incident radiation. The activator A can absorb radiation (exc_A). This excitation is followed by emission (em_A) and/or by the dissipation of heat. The activator can also be excited via the sensitizer S. In that case S absorbs the radiation (exc_S) and then transfers excitation energy to A. Emission and/or heat dissipation from S are also possible.

Dr. G. Blasse, formerly with Philips Research Laboratories, Eindhoven, is now Professor of Solid State Chemistry at the University of Utrecht. Dr. A. Bril is with Philips Research Laboratories, Eindhoven. On 26th August 1970 Professor Blasse was awarded the Gold Medal for Research of the Royal Netherlands Chemical Society for his work on ferrites and on luminescence.

[1] The terms fluorescence and phosphorescence have been much longer in use than luminescence, although with different meanings in different branches of science. We have therefore avoided their use here. Particulars of the terminology used in this field will be found in G. F. J. Garlick, Handbuch der Physik, ed. S. Flügge, Vol. XXVI, Springer, Berlin 1958, p. 1.

of the activator. In the third part we shall take a look at cases of luminescence where the excitation energy is not absorbed by the activator itself.

In what follows, the role of activator is often played by one of the ions of the rare-earth metals (R.E. ions) or of the closely related elements yttrium (Y) and scandium (Sc).

Research on these phosphors in particular has considerably advanced the understanding of characteristic luminescence, since the properties of these phosphors can be studied on simple model compounds. This is possible because of the close similarity between the chemical properties of these ions. The host lattice may be, for instance, a compound of the ions La^{3+}, Y^{3+} or Lu^{3+}. The latter ions do not absorb ultraviolet radiation. Rare-earth ions, for example Eu^{3+} or Tb^{3+}, are now substituted for a small proportion of the host-lattice ions. These R.E. ions occupy in the host lattice the crystallographic sites of La^{3+}, Y^{3+} or Lu^{3+} in a virtually random distribution. This too is attributable to the similarity between their chemical properties. It is possible in this way to make phosphors whose chemical constitution is well defined. As will be seen in the following sections, this has important advantages.

The energy-level diagram of the ions of the rare-earth metals

The characteristic properties of the R.E. ions are attributable to the presence in the ion of a deep-lying 4f shell (see Appendix I, page 313) which is not entirely filled. The electrons of this shell are screened by the outer electron shells, and as a result they give rise to a number of discrete energy levels. Since the presence of the crystal lattice scarcely affects the positions of these levels, there is a close resemblance between the energy-level diagram of the free ion and that of the incorporated ion.

The 4f shell may contain 14 electrons. *Table I* shows the most common valencies of the R.E. ions and the number of 4f electrons in the ground state of the relevant ions. The energy-level diagrams for Ce^{3+}, Eu^{2+}, Eu^{3+}, Gd^{3+} and Tb^{3+} are given in *fig. 2*. These energy-level diagrams have been chosen here as examples because they are the simplest ones and are at the same time representative of all the types encountered. In most R.E. ions the number of levels is fairly large, except in Ce^{3+} and Eu^{2+} (and Yb^{3+}). The Ce^{3+} ion has only one 4f electron, and this gives rise to two energy levels: in the one state the orbital and spin moments of the electron are parallel ($^2F_{7/2}$), and in the other state antiparallel ($^2F_{5/2}$). As the number of electrons increases, there is in general a rapid increase in the number of possible states.

Fig. 2 shows that in addition to the discrete 4f levels there are other levels present. These are represented schematically as broad, hatched bands. The energy levels of these bands depend to a great extent on the lattice.

The bands referred to fall into two groups. In the first group one of the 4f electrons is raised to the higher 5d level: $4f^n \rightarrow 4f^{n-1}5d$. This 5d orbit lies at the surface of the ion, and can therefore be strongly influenced by the lattice. In the Eu^{2+} ion, the $4f^65d$ level lies so low that the $4f^7$ levels present (except for the ground level) are completely overlapped (fig. 2). In the second group one of the electrons of the surrounding anions of the lattice has jumped across to the 4f orbit of the central R.E. ion. Since this is a transfer of charge, the state is referred to as a charge-transfer state. Obviously the position of this energy band depends on the nature of the surrounding ions.

What is it that determines whether the energetically lowest band corresponds to a $4f^{n-1}5d$ state or to a charge transfer state? The answer to this question is bound up with the fact that a state with a completely or half-filled electron shell is very stable. If we compare, for example, the trivalent ions with one another, we get the following picture. The excited states of Gd^{3+} ($4f^7$, hence half-filled) lie at a high energy level (fig. 2). In the case of Tb^{3+} ($4f^8$, half-filled plus one) the 4f shell readily releases an electron, and the transition $4f^8 \rightarrow 4f^75d$ takes place at relatively low energy, while in the case of Eu^{3+} ($4f^6$, half-filled less one) the 4f shell readily accepts an electron and thus the charge-transfer state has a low energy.

Having seen which states play a part in the R.E. ions

Table I. The ions of the rare-earth metals and the number of 4f electrons in their respective ground states.

ion	number of 4f electrons
La^{3+}	0
Ce^{3+}	1
Ce^{4+}	0
Pr^{3+}	2
Nd^{3+}	3
Pm^{3+}	4
Sm^{2+}	6
Sm^{3+}	5
Eu^{2+}	7
Eu^{3+}	6
Gd^{3+}	7
Tb^{3+}	8
Tb^{4+}	7
Dy^{3+}	9
Ho^{3+}	10
Er^{3+}	11
Tm^{3+}	12
Yb^{2+}	14
Yb^{3+}	13
Lu^{3+}	14

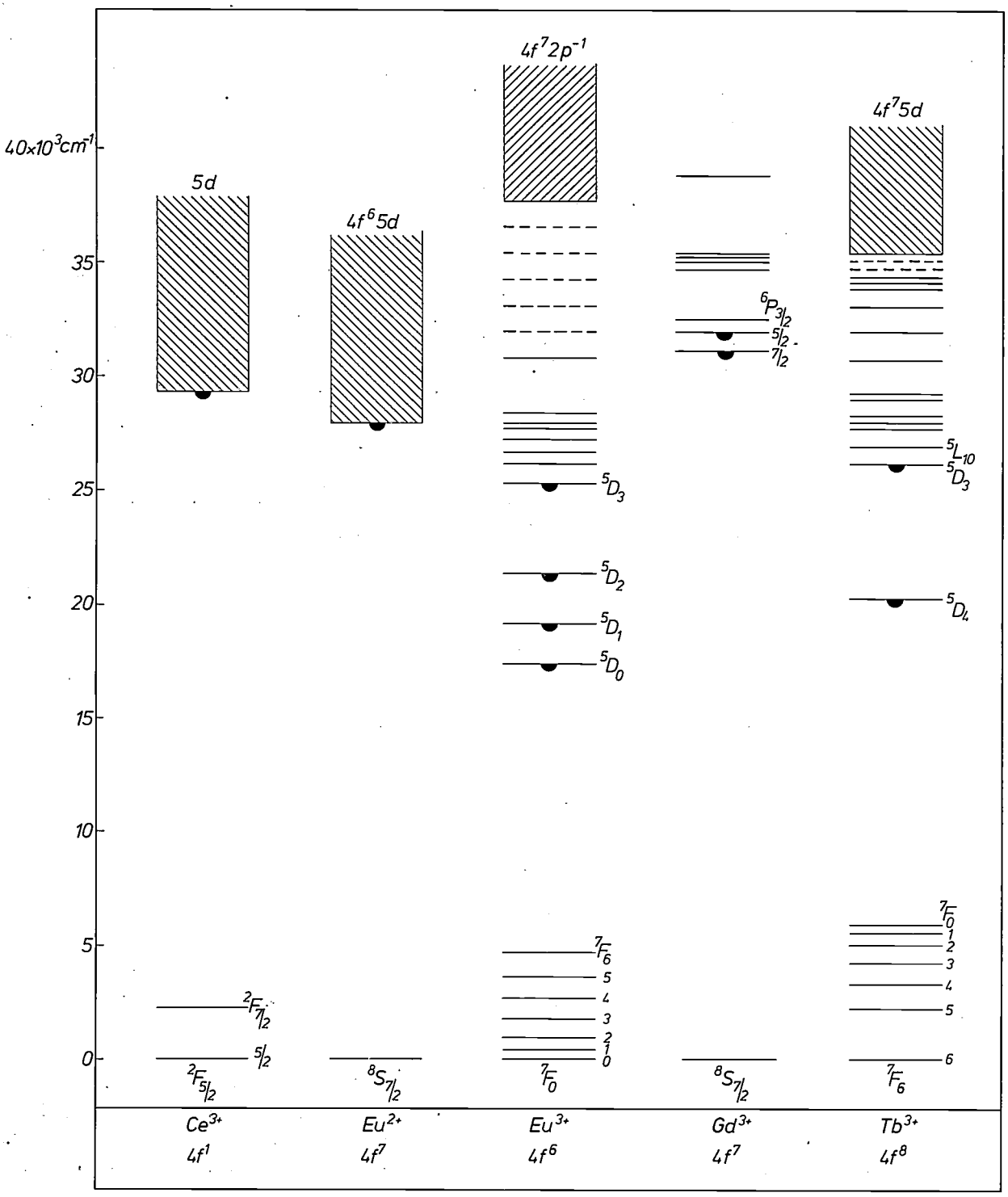


Fig. 2. Energy-level diagram of some ions of rare-earth metals in oxide host lattices (the energy is plotted vertically in cm^{-1}). Horizontal lines indicate the narrow 4f levels. Where the levels are not well known they are shown as dashed lines. The hatched broad bands correspond to charge-transfer states (Eu^{3+}) or $4f^{n-1}5d$ states (Ce^{3+}, Eu^{2+}, Tb^{3+}). For Gd^{3+} these states have such a high energy that they cannot be shown in the figure. Levels labelled with black half-circles are levels from which luminescence is observed.

and considered the basic structure of the energy-level diagrams of these ions, we shall deal in the following sections with the optical transitions between the levels present in the energy-level diagram of R.E. ions.

Situations will be encountered where the electric-dipole transition between two levels is allowed, and others where such a transition is forbidden. It will be seen that in the latter case, apart from magnetic-dipole

radiation, electric-dipole radiation is nevertheless frequently observed, albeit very much weaker. We shall look at the conditions in which a forbidden transition partly ceases to be forbidden. We shall devote considerable attention to the red emission of the Eu^{3+} ion, not only because it concerns a very simple and instructive case, but also because of the great practical importance of Eu^{3+} phosphors in their application as the red phosphor on colour television screens [2].

Optical transitions involving a 5d level or a charge-transfer state

Let us look first at 4f-5d transitions. These transitions are allowed for the emission and absorption of electric-dipole radiation (see Appendix II, page 313) and therefore correspond to an intense optical absorption. It may be derived from fig. 2 that this absorption lies in the ultraviolet part of the spectrum for the ions mentioned in the figure (Ce^{3+}, Eu^{2+}, Tb^{3+}). *Figs. 3* and *4* give reflection and excitation spectra for the garnets $Y_3Al_5O_{12}$, $(Y,Ce)_3Al_5O_{12}$, $(Y,Tb)_3Al_5O_{12}$ and $(Y,Tb)_3Ga_5O_{12}$ being respectively the host lattice without activator and the host lattice with Ce^{3+} and with Tb^{3+} as activator). Both the reflection spectra and the excitation spectra give a picture of the absorption, and we see that in the activated crystals there is indeed strong absorption in the UV. It is noticeable here, particularly in the excitation spectra, that this absorption takes place in a number of discrete bands. This may be explained as follows. The excited 5d state is strongly influenced by the electric field of the surrounding ions, that is to say, the crystal field [3]. This crystal field has the effect of splitting the 5d level into a number of levels which are roughly 15 000 to 20 000 cm^{-1} apart. The number of these levels is determined by the crystallographic symmetry at the position of the rare-earth ion.

Since the crystal-field splitting varies considerably from one lattice to another, so too does the spectral position of the absorption bands appertaining to a particular 4f-5d transition (fig. 4). In *Table II* this is illustrated for the Ce^{3+} ion.

Now let us see what happens when an activator in which 4f-5d transitions take place is excited in the corresponding absorption bands in the UV.

In the case of Tb^{3+}, excitation in the 4f-5d absorption bands is followed by green emission. As a result of absorbing UV radiation, the ion is raised to a $4f^75d$ state; it then decays stepwise from this state to the $^5D_3$ or the $^5D_4$ state, or both (see fig. 2), thereby giving up to the lattice small quanta of energy, or phonons. Because of the large distance between these states and the ^7F levels, the process stops here and the ion then returns to the ground state by emitting radiation (luminescence). Although the position of the 4f-5d absorption and excitation bands depends to a very great extent on the nature of the lattice, the (green) emission does not. This, of course, is because the emission is the consequence of a transition between 4f levels (in prin-

Table II. Position of some absorption and excitation bands of Ce^{3+} in various host lattices (in 10^3 cm^{-1}).

$(Y,Ce)BO_3$	27.4	29	40.8	43.5
$(La,Ce)BO_3$	30.8	36.9	41.5	
$(Y,Ce)PO_4$	32.8	34.2	39.6	
$(Y,Ce)_3Al_5O_{12}$	22.0	29.4	37	44
(Y,Ce)OCl	31.6	35.8		

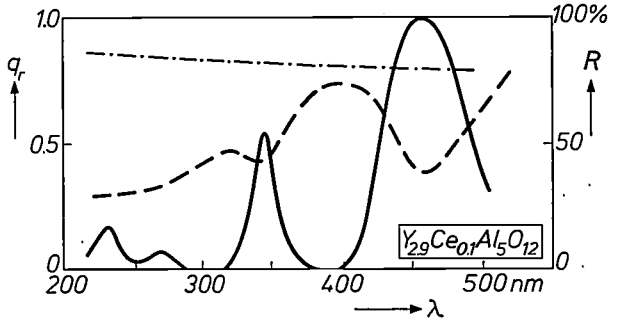


Fig. 3. The chain-dotted line gives the reflection spectrum (the reflectance R in % as a function of wavelength λ) of the host lattice $Y_3Al_5O_{12}$. Spectral regions of low reflection correspond to regions of high absorption. As can be seen, the host lattice shows practiclaly no absorption in the spectral region considered. The dashed line indicates the reflection spectrum of the Ce^{3+}-activated phosphor $Y_{2.9}Ce_{0.1}Al_5O_{12}$, which has two absorption bands. The solid line gives the excitation spectrum of the Ce^{5+} luminescence of $Y_{2.9}Ce_{0.1}Al_5O_{12}$: the relative quantum yield q_r of the luminescence is plotted as a function of the wavelength of the exciting radiation. (q_r is the ratio of the number of emitted quanta to the number of incident excitation quanta, multiplied by a factor such that the maximum becomes unity.) Each excitation band corresponds to an absorption band. The excitation spectrum gives a clearer picture of the absorption bands than the reflection spectrum. The Ce^{3+} absorption bands correspond to the various 4f-5d transitions. The distance between them in the spectrum is equal to the crystal-field splitting of the 5d level [3].

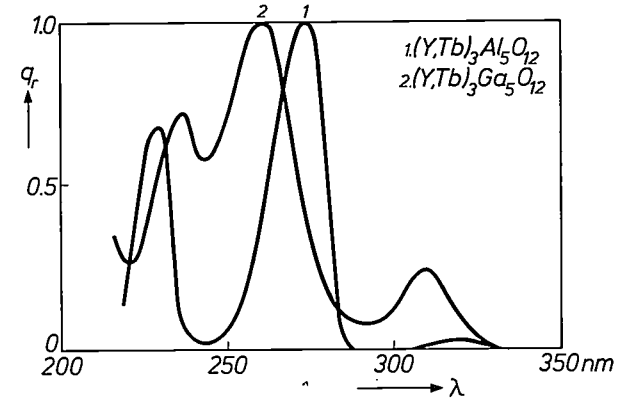


Fig. 4. Excitation spectrum of the Tb^{3+} emission of the compounds $Y_{2.9}Tb_{0.1}Al_5O_{12}$ and $Y_{2.9}Tb_{0.1}Ga_5O_{12}$. The excitation bands correspond to absorption resulting from a 4f-5d transition. It can be seen that the substitution of Ga for Al causes a marked difference in crystal-field splitting.

[2] See for example A. Bril and W. L. Wanmaker, Philips tech. Rev. 27, 22, 1966.

[3] For a treatment of the crystal field and its influence on the energy levels, see P. F. Bongers, Philips tech. Rev. 28, 13, 1967.

ciple a strictly forbidden transition for electric-dipole radiation, but this will be discussed presently).

The situation as far as the Ce^{3+} ion is concerned is entirely different. Excitation in the 4f-5d absorption bands is followed by emission from the 5d states themselves. Contrary to the case of Tb^{3+}, the emission here depends strongly on the lattice. *Fig. 5* shows various emission spectra of Ce^{3+} phosphors. Emission of a very short wavelength is found with $(Y,Ce)PO_4$, whereas the emission of $(Y,Ce)_3Al_5O_{12}$ is of very long wavelength. The latter phosphor exhibits emission not only from the lowest 5d state but also from the state immediately above it. It is also noticeable that the emission bands are split to some extent. This splitting is a consequence of the fact that the emissive transition takes place both to the $^2F_{7/2}$ level and to the $^2F_{5/2}$ level.

unusual for characteristic luminescence, in which phosphors can be made with the same anions and the same activator in different host lattices so as to produce different emissions covering almost the entire visible spectrum. This is due in the first place to the difference in crystal-field splitting of the 5d level.

We shall now consider the optical absorption caused by a transition to a charge-transfer state. The Eu^{3+} ion shows absorption of this type. Some examples of reflection spectra are presented in *fig. 7*. These transitions, too, correspond to allowed optical transitions. Unlike the 4f-5d transitions, however, there is no distinct splitting in the absorption spectra (cf. figs. 3 and 4). The relation between the energy level of the charge-transfer absorption band and the nature of the host lattice can be explained here in the following way. In

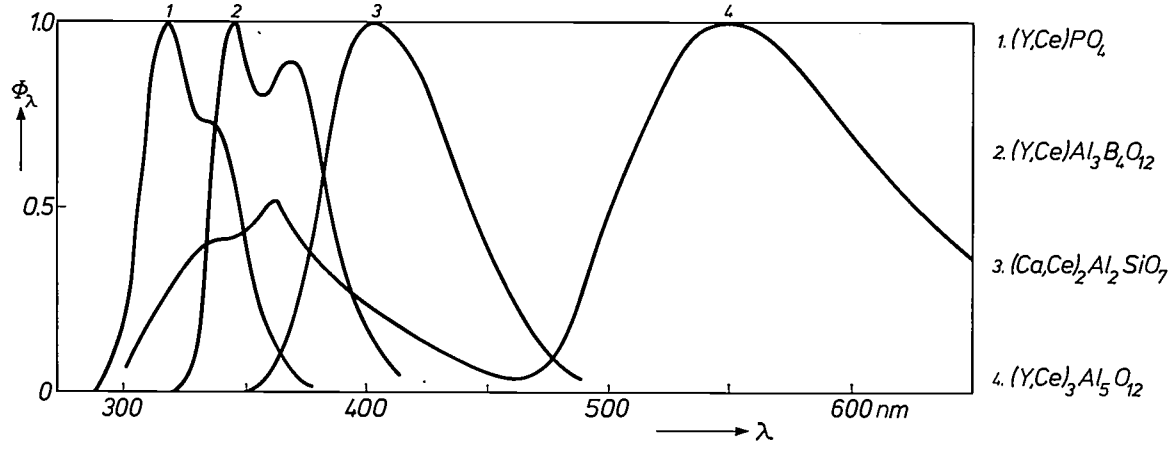


Fig. 5. Emission spectrum of some Ce^{3+} phosphors for excitation with 254 nm radiation. The quantity Φ_λ is the relative spectral radiance. The maxima are put equal to 1. Note that in $(Ca,Ce)_2Al_2SiO_7$, trivalent Ce^{3+} ions are situated at the sites of divalent Ca^{2+} ions.

The average lifetime of the excited 5d state of the Ce^{3+} ion is very short, being between 10^{-7} and 10^{-8}s. After the excitation (UV radiation or fast electrons), the Ce^{3+} ion returns very rapidly to the ground state; the intensity of the emission therefore decreases rapidly when the excitation stops. This property makes the Ce^{3+} phosphors very suitable for application in tubes for flying-spot scanners and in index picture tubes for colour television [4].

In the case of the Eu^{2+} ion, the excited 4f-5d band overlaps nearly all $4f^7$ levels, and in this respect it strongly resembles the Ce^{3+} ion. There are some conspicuous differences, however. Firstly, the emission band is not split (see *fig. 6*). This is due to the fact that there is only one 4f level below the $4f^65d$ band ($^8S_{7/2}$). Secondly the $4f^65d$ state has a lower energy than the 5d state of the Ce^{3+} ion. Consequently the emission bands of the Eu^{2+} ion in various host lattices lie closer to the visible part of the spectrum. We then have the situation, the crystal lattice the Eu^{3+} ion will be surrounded by negative anions, for example O^{2-} ions. A free O^{2-} ion is unstable and breaks up into an O^- ion and an electron, giving up energy in the process; in a crystal lattice, however, an O^{2-} ion is stabilized by the surrounding positive ions. The smaller the radius and the higher the charge of these ions, the larger this stabilization will be and the more energy it will cost to remove an electron from the O^{2-} ion. We may therefore expect the charge-transfer absorption band of an Eu^{3+} ion surrounded by anions to lie at shorter wavelength (i.e. higher energy) if the cations of the lattice are smaller and carry a higher charge. The first case is illustrated in *Table III*. In the emission process of the Eu^{3+} ion the charge-transfer level plays no part, since the ion decays from the charge-transfer level via a number of 4f levels to the ^5D levels, from which the ground state is reached by the emission of radiation (fig. 2). More will be said about this under the next heading.

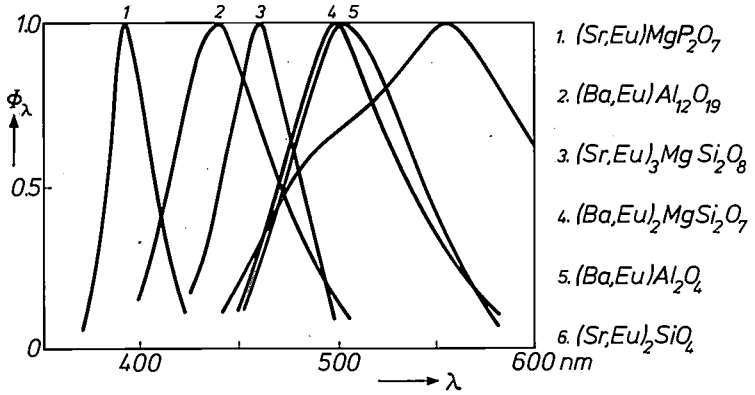


Fig. 6. Emission spectrum of some Eu^{2+} phosphors for excitation with 254 nm radiation.

Optical transitions between 4f levels

Electric-dipole transitions between 4f levels of rare-earth ions are in principle strictly forbidden. This is because the *parity* of the wave function of the electrons does not change (Laporte's selection rule, see Appendix II).

We shall now consider in particular the transitions between the ^5D and ^7F levels of the Eu^{3+} ion. The electric-dipole transition between these levels is forbidden not only because of the above-mentioned Laporte's selection rule, but also because the spin quantum number S of the total angular momentum changes (from 2 to 3). As a result of this the transition is forbidden both for electric- and for magnetic-dipole radiation.

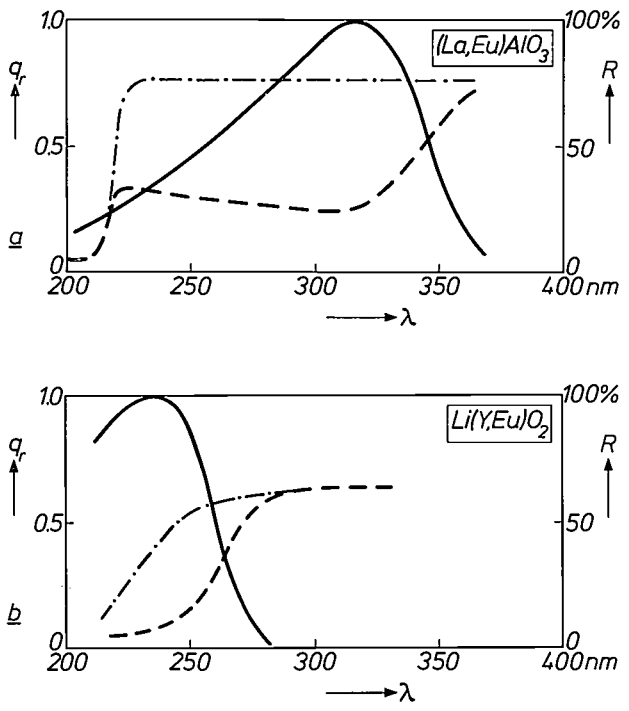


Fig. 7. *a*) Reflection spectrum of the host lattice $LaAlO_3$ (chain-dotted line) and of (La,Eu)AlO_3 (dashed). The solid line is the excitation spectrum of the Eu^{3+} luminescence of (La,Eu)AlO_3. The absorption and excitation bands at about 310 nm correspond to the charge-transfer absorption. R is the reflectance, q_r the relative quantum yield. *b*) As *a*, but now for Li(Y,Eu)O_2. The charge-transfer absorption now lies at much higher energy (shorter wavelength).

Table III. Position of the charge-transfer level of Eu^{3+} and its dependence on the nature of the host lattice, particularly on the size of the cations. The phosphors are divided into two groups, each with their own crystal structure.

phosphor	charge transfer level (10^3 cm^{-1})	ionic radii (Å)
(La,Eu)OBr	30.7	Br^- 1.95; La^{3+} 1.14
(La,Eu)OCl	33.3	Cl^- 1.81; La^{3+} 1.14
(Gd,Eu)OCl	35.0	Gd^{3+} 0.97
(Y,Eu)OCl	35.4	Y^{3+} 0.92
$Na(La,Eu)O_2$	36.0	Na^+ 0.94; La^{3+} 1.14
$Na(Gd,Eu)O_2$	41.1	Na^+ 0.94; Gd^{3+} 0.97
$Li(Y,Eu)O_2$	42.0	Li^+ 0.68; Y^{3+} 0.92
$Li(Lu,Eu)O_2$	43.0	Li^+ 0.68; Lu^{3+} 0.85

How, then, can the relevant transitions nevertheless be observed? No more than a very brief summary of the underlying theory [5] can be given here.

The *spin prohibition* is not strict because the description of the ^7F levels as states with six parallel spins is not entirely correct. Because of spin-orbit coupling it is necessary to consider what we call ^7F states as being composed of a pure ^7F state with a slight "admixture" of the pure ^5D state. Consequently this spin prohibition no longer applies so strictly. The same applies to the free ion.

The *parity prohibition* can be lifted only by the influence of the crystal lattice. Just as the spin prohibition was cancelled by the slight mixing of the ^7F state with the ^5D state as a result of spin-orbit coupling, so too can the parity prohibition be cancelled by mixing the $4f^6$ configuration with a state possessing a different parity. The interaction responsible for this is formed by the odd crystal-field terms [5], that is to say those terms that change sign on inversion with respect to the R.E. ion. If the R.E. ion is located at a site that is a centre of symmetry in the relevant crystal lattice, then the odd crystal field terms are absent and the parity prohibition cannot be lifted.

In that case only magnetic-dipole transitions are possible. The selection rule here is: $\Delta J = 0, \pm 1$

[4] Shortly to appear in this journal.

[5] See for example B. G. Wybourne, Spectroscopic properties of rare earths, Interscience, New York 1965.

(except that $J = 0 \rightarrow J = 0$ is forbidden). J is the total orbital angular momentum and appears in the notation as a subscript, e.g. ^7F_J (here J can have values ranging from 0 to 6). If, then, the Eu^{3+} ion is situated at a centre of symmetry and is brought into the $^5D_0$ state (fig. 2), the only possible transition accompanied by the emission of radiation is $^5D_0 \rightarrow {}^7F_1$ (magnetic-dipole emission; initial level $J = 0$, final level $J = 1$, $\Delta J = +1$). *Fig. 8* shows the emission spectrum of an Eu^{3+} ion situated at a centre of symmetry. As expected, this spectrum consists of emission lines that correspond to the $^5D_0$-$^7F_1$ transition. The colour of this emission is orange. Since the 4f levels are discrete, we must expect discrete emission lines, and these are in fact observed. The figure also shows that in the case of Eu^{3+} in Ba_2GdNbO_6 only one $^5D_0$-$^7F_1$ line is found. For Eu^{3+} in $NaLuO_2$ and $Gd_2Ti_2O_7$ several $^5D_0$-$^7F_1$ emission lines are found. What is the reason for this disparity?

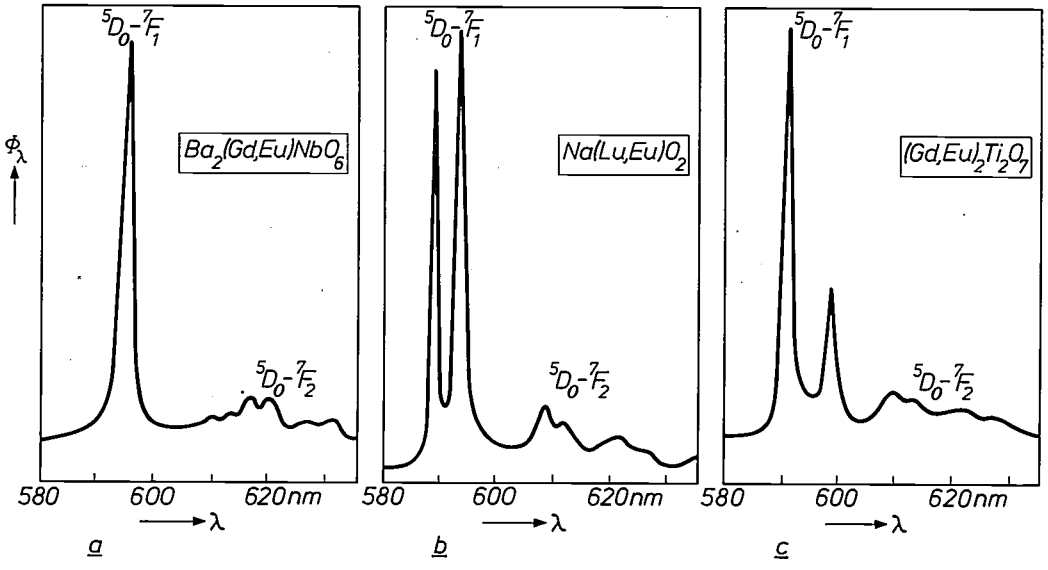


Fig. 8. Emission spectrum (linear scale) of the Eu^{3+} ion when it occupies a site with a centre of symmetry. Excitation with 254 nm radiation, *a*) in Ba_2GdNbO_6, *b*) in $NaLuO_2$, *c*) in $Gd_2Ti_2O_7$. The colour of the emission is orange.

We have already mentioned above that ionic energy levels can be split by the field of the surrounding ions (crystal-field splitting). For the 5d level the splitting is considerable (this orbit is at the surface of the ion). Crystal-field splitting is also found for 4f levels but, since the 4f electrons are well screened from the environment, the splitting is much smaller. For d electrons the splitting may amount to a few 10 000 cm^{-1}, but for the 4f electrons it may be no more than a few 100 cm^{-1}. Now a level with $J = 0$ is a single, non-degenerate state and cannot therefore be split. A level with $J = 1$ is triply degenerate and can be split. The manner of splitting depends on the symmetry of the crystal field. A field possessing cubic symmetry permits triple degeneration (equivalent x-, y- and z-axes) and does not cause splitting. Tetragonal and trigonal fields cause splitting into two levels; fields with lower symmetry cause splitting into three levels.

In Ba_2GdNbO_6 the Eu^{3+} ion occupies the position of Gd^{3+}. This is a crystallographic site with cubic sym-

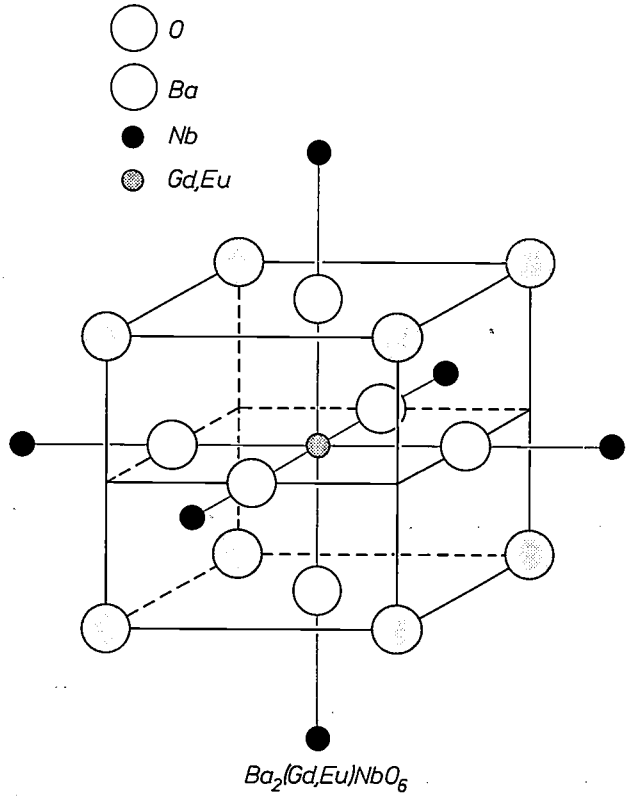


Fig. 9. Crystal structure of Ba_2GdNbO_6.

metry (see *fig. 9*). The $^7F_1$ level is therefore not split. The $^5D_0$ level ($J = 0$) can never be split. The emission transition $^5D_0$-$^7F_1$ therefore consists of one line. In $NaLuO_2$ (*fig. 10*) and $Gd_2Ti_2O_7$ the symmetry at the location of the Eu^{3+} ion is trigonal. The $^7F_1$ level is split into two sublevels, and the emission transition $^5D_0$-$^7F_1$ therefore consists of two lines.

(fig. 8*b*) and in $NaGdO_2$ (fig. 11*a*). Both host lattices crystallize in the rock-salt structure (fig. 10; standard examples NaCl, MgO). All cations here are located at the centre of an octahedron with six O^{2-} ions at the corners (i.e. at a centre of symmetry). In $NaLuO_2$ ($NaGdO_2$) the Mg^{2+} ions in MgO are replaced by Na^+ and Lu^{3+} (Gd^{3+}) ions. The monovalent and trivalent

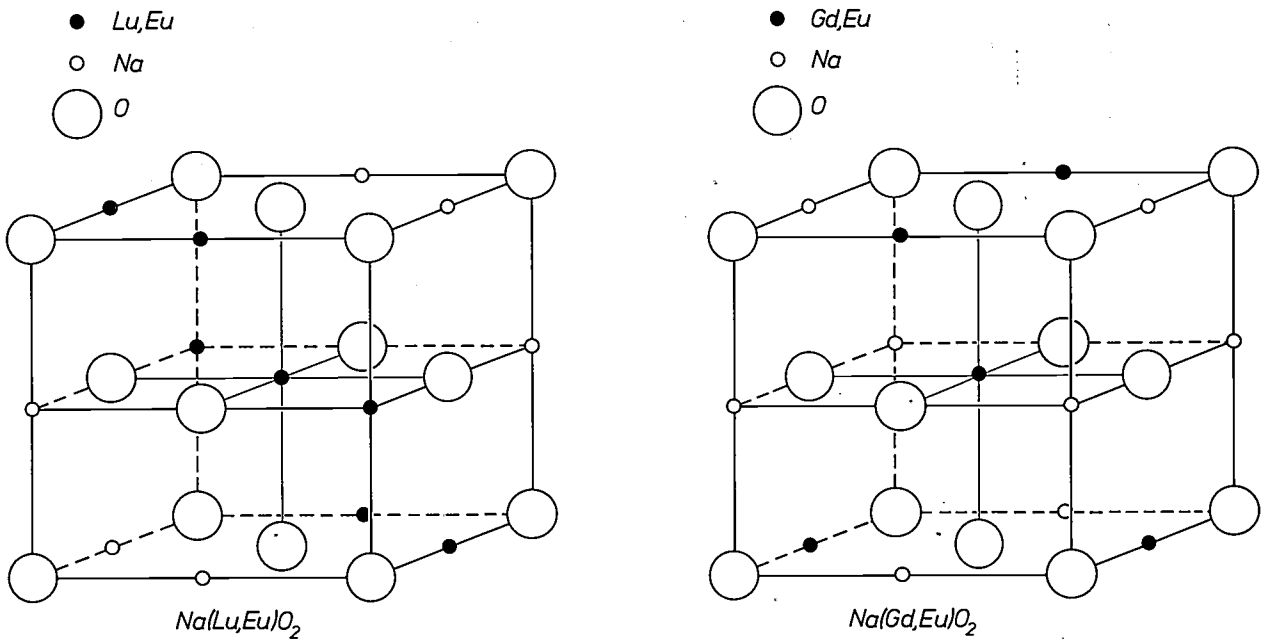


Fig. 10. Crystal structure of $NaLuO_2$ and $NaGdO_2$ (schematic). To make the relation between the two structures clear, the unit cube of the rock-salt structure is drawn rather than the unit cells. Deformations of the ideal structure are not represented.

Let us now consider the situation where the Eu^{3+} ion occupies a crystallographic site that does not coincide with a centre of symmetry. In this case not only magnetic-dipole transitions are possible but also electric-dipole transitions. The latter are known as forced electric-dipole transitions (the parity prohibition having had to be forced). The forced electric-dipole transitions are similarly subject to selection rules, viz. $\Delta J \leqq 6$. If, however, $J = 0$ for the initial or final level, then $\Delta J = 2$, 4 or 6.

In the example we have chosen (emission starting from the $^5D_0$ level of the Eu^{3+} ion) we have $J = 0$ for the initial level. We may therefore expect the following electric-dipole transitions: $^5D_0$-$^7F_2$, $^7F_4$, $^7F_6$ with, in addition, $^5D_0$-$^7F_1$ (magnetic-dipole transition). The transitions $^5D_0$-$^7F_0$, $^7F_3$, $^7F_5$ will necessarily be of low intensity, and this is in fact observed [6].

Fig. 11 gives some examples of emission spectra of the Eu^{3+} ion in host lattices where it occupies a site which is not a centre of symmetry. The colour of the emission from these phosphors is red. It is interesting to compare the emission spectra of Eu^{3+} in $NaLuO_2$ ions, however, are ordered over the cation sites (superstructure). This differs for the combinations $Na^+ + Lu^{3+}$ and $Na^+ + Gd^{3+}$ (see fig. 9). Owing to the difference in superstructure, Eu^{3+} occupies a centre of symmetry in $NaLuO_2$ but not in $NaGdO_2$. This seemingly minor difference in structure has a considerable influence on the relative intensities of the Eu^{3+} emission lines. In $NaGdO_2$ the electric-dipole lines predominate and the colour of emission is red; in $NaLuO_2$ they are absent and the colour of emission is orange.

A comparison of the emission spectra also shows that the Eu^{3+} ion in $NaLuO_2$ does show some emission at the position of the $^5D_0$-$^7F_2$ lines. This emission consists of weak, fairly broad lines. The relevant transitions occur because the ions of the host lattice are not stationary, as hitherto assumed, but in fact vibrate. These vibrations can cause a deviation from pure inversion symmetry, which means that the electric-dipole transitions are no longer forbidden.

In $YAl_3B_4O_{12}$ the Eu^{3+} ion occupies the centre of a

[6] G. Blasse, A. Bril and W. C. Nieuwpoort, J. Phys. Chem. Solids 27, 1587, 1966.

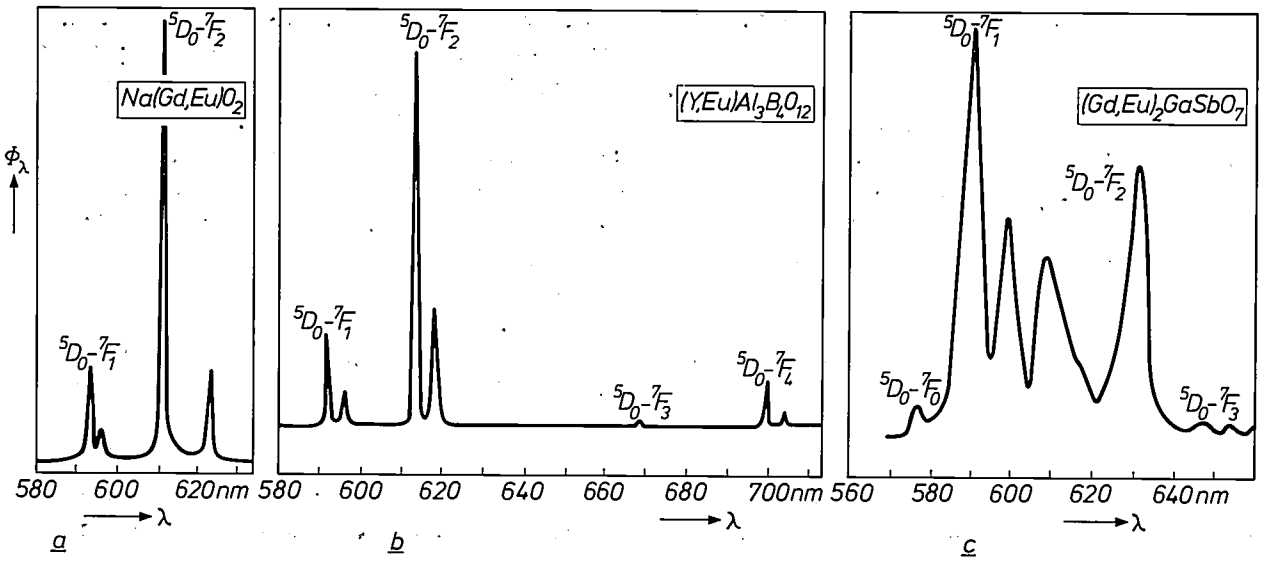


Fig. 11. Emission spectrum (linear scale) of the Eu^{3+} ion when not located at a centre of symmetry. Excitation with 254 nm radiation, *a*) in $NaGdO_2$ (cf. fig. 8*b*), *b*) in $YAl_3B_4O_{12}$ (a somewhat larger spectral region is drawn here) and *c*) in Gd_2GaSbO_7 (cf. fig. 8*c*). The colour of the emission is red.

trigonal prism. With such an environment there cannot of course be any centre of symmetry. Nevertheless the emission spectrum of Eu^{3+} in $YAl_3B_4O_{12}$ (fig. 10*b*) differs only slightly from that of Eu^{3+} in $NaGdO_2$ (fig. 10*c*), which still shows, to a rough approximation, a centre of symmetry. This indicates that even slight deviations from inversion symmetry have a marked influence on the intensity of the electric-dipole transitions of R.E. ions.

The same appears from the emission spectrum of Eu^{3+} in Gd_2GaSbO_7 (fig. 10*c*). This host lattice can be thought of as being derived from $Gd_2Ti_2O_7$ by substituting for two Ti^{4+} ions a Ga^{3+} and an Sb^{5+} ion. The Eu^{3+} ion in $Gd_2Ti_2O_7$ shows orange emission, but in Gd_2GaSbO_7, where the next-nearest neighbours of the Eu^{3+} ion are different, the emission is red.

It is worth noting that the lifetime of the luminescent $^5D_0$ level is about 10^{-3} s. This is approximately 10^5 times longer than the lifetime of a level that luminesces via an allowed electric-dipole transition, and roughly equal to the value expected for a magnetic-dipole transition. This illustrates just how strictly forbidden the 4f-4f transitions are.

Up to now we have not considered which state has to be mixed in the $4f^6$ configuration in order to break through the parity prohibition. The most likely assumption is that this is the 5d state. As far as the Eu^{3+} ion is concerned we have a different opinion [7]. In the case of Eu^{3+} the 5d state lies at high energy. The mixing of states is the more important the smaller the energy difference between the mixed states. It therefore seems obvious in the case of the Eu^{3+} ion to assume that the charge transfer state is the one that is mixed in the 4f states. The effect will be greater (more electric-dipole radiation) the lower the energy of the charge-transfer state. This is in fact the case (see *Table IV*).

Considerable support for our view comes from the fact that the emission of the Eu^{3+} ion in fluoride host lattices always consists to a large extent of magnetic-dipole radiation, even if the Eu^{3+} ion is located at a position without inversion symmetry (see Table IV). Below energies of 50 000 cm^{-1}, the Eu^{3+}-$(F^-)_n$ complex does not show broad absorption bands. The charge-transfer band is shifted with respect to that in oxides towards a much higher energy. This is attributable to the fact that the F^- ion is very much more electronegative than the O^{2-} ion. It will thus cost more energy to remove an electron from the F^- ion than from an O^{2-} ion. Admixture with a state possessing the opposite parity will not therefore occur to any great extent, and the emission will contain a great deal of magnetic-dipole radiation.

Let us now return to the oxides. It also appears to be possible to introduce into the lattice the charge-transfer

Table IV. Ratio of the intensity of the magnetic-dipole emission ($^5D_0$-$^7F_1$) to that of the total electric-dipole emission for Eu^{3+} in various host lattices, and its dependence on the position of the lowest-lying, strong absorption bands. In the last column the nature of the group showing this absorption is indicated between brackets.

host lattice	ratio	position of lowest-lying strong absorption band (10^3 cm^{-1})
YF_3	0.85	ca. 50 (Eu^{3+}-F^-_n)
$ScPO_4$	0.62	ca. 48 (Eu^{3+}-O^{2-}_n)
YPO_4	0.43	ca. 45 (Eu^{3+}-O^{2-}_n)
YVO_4	0.15	31.2 (VO_4)
$ScVO_4$	0.12	29.8 (VO_4)

state of other groups that are neighbours of the Eu^{3+} ion. An example is the charge-transfer state of the vanadate group in the familiar phosphor $(Y,Eu)VO_4$ (see Table IV). This phosphor gives red emission. On the other hand the phosphor $(Y,Eu)PO_4$, which has the same crystal structure, gives orange emission. In the vanadate phosphor the low-lying (33 000 cm^{-1}) charge-transfer state of the vanadate group is introduced. The phosphate phosphor contains no such low-lying charge-transfer state. The lowest-lying state that can then be considered is the charge-transfer state of the Eu^{3+} ion itself (45 000 cm^{-1}). Owing to its much higher energy, however, this state is much less effective in breaking down the parity prohibition.

We see then that, if we choose the host lattice appropriately, we can control the colour of emission from Eu^{3+} phosphors. Orange emission ($^5D_0$-$^7F_1$) is only possible if the Eu^{3+} ion in the lattice occupies a site with a centre of symmetry. Red emission ($^5D_0$-$^7F_2$) is only possible provided the Eu^{3+} ion does not occupy a site that is a centre of symmetry in the crystal structure and also provided that a state exists with allowed optical transitions that is not too high above the ground state.

Some surprises are found, however, one example being the phosphor with the composition $Sr_{2-2x}Na_xEu_xTiO_4$. In Sr_2TiO_4 part of the strontium is replaced by Na^+ and Eu^{3+}. This phosphor has an orange emission [8]. The emission spectrum (*fig. 12*) shows that this is not the result of $^5D_0$-$^7F_1$ emission but of strong $^5D_0$-$^7F_0$ emission, which, together with the $^5D_0$-$^7F_2$ emission, causes the orange colour. This transition ($J = 0 \rightarrow J = 0$) is a most strictly forbidden one. To explain the observed intensity one must assume a linear crystal-field term. In this article we shall not go any further into this phenomenon.

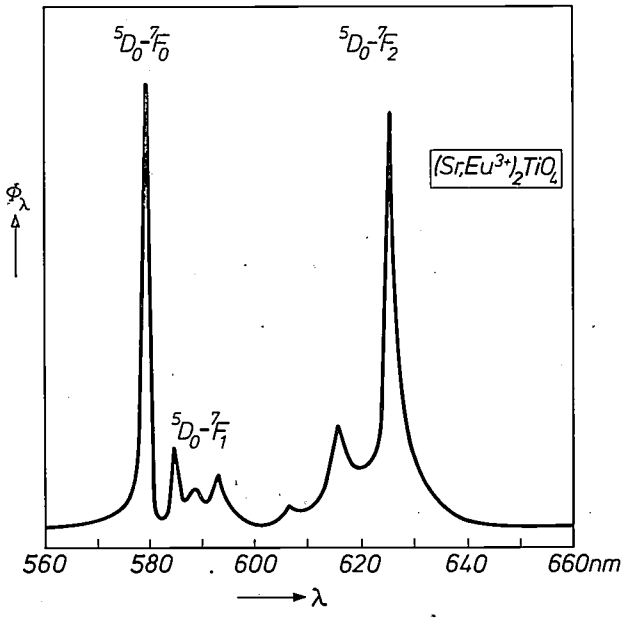


Fig. 12. Emission spectrum (linear scale) of the Eu^{3+} ion in Sr_2TiO_4. Excitation with 365 nm radiation. The colour of emission is orange. Note the very intense $^5D_0$-$^7F_0$ transition.

The case we have discussed in the foregoing ($^5D_0$-^7F_J emission of Eu^{3+}) is a fairly simple one, because the $^5D_0$ state is not split by the crystal field and because a simple selection rule applies to the electric dipole transitions. More complex cases, such as for example the emission of Tb^{3+}, will not be considered in this article. In any case, the structure-dependence in such cases is by no means so striking.

In this first part of our article we have looked at the energy diagram and associated optical transitions of a number of rare-earth ions. These diagrams and transitions are nowadays reasonably well known. The influence of the crystal lattice on the situation and intensity of absorption and emission bands or lines can also be well understood in qualitative terms. In the next part we consider the efficiency of the luminescence.

Appendix I. The notation of the quantum states

We give here a short description of the notation of the quantum states, both for the individual electrons in their different orbits and for the electron cloud as a whole. In describing the quantum states of electrons we use the principal quantum number n and the azimuthal quantum number l, which characterizes the orbital angular momentum. The principal quantum number is expressed in figures, the azimuthal quantum number in lower-case letters (s, p, d, f, etc. for $l = 0, 1, 2, 3$, etc.). We speak, for example, of 4f electrons or 4f orbitals. Exponents are used to indicate the number of electrons of each type of which the electron cloud is composed, for example $(1s)^2\ (2s)^2\ (2p)^6 3s$ or simply $1s^2 2s^2 2p^6 3s$.

In describing quantum states of the whole cloud we use in addition to the azimuthal quantum number L of the *total* orbital angular momentum, the quantum number S of the *total* spin angular momentum and the quantum number J of the *total* angular momentum. These quantum states are described with the symbol $^{2S+1}L_J$. The quantity $2S + 1$ indicates the number of values which J can assume ($|L - S|, |L - S + 1|, \ldots L + S$), referred to as its multiplet character. The value of L is again expressed in letters, but now in the upper case (S, P, D, F etc. for $L = 0, 1, 2, 3$, etc.). Thus, when $S = \frac{1}{2}$, $L = 3$ and $J = 7/2$, we speak of the $^2F_{7/2}$ level.

Appendix II. Transition probability, selection rules

The radiation emitted during the transition of an electron from one quantum state to another is an oscillation of the electromagnetic field in the space around the atom. The frequency ν of this oscillation is given by Bohr's equation: $h\nu = E_1 - E_2$; here h is Planck's constant, and E_1 and E_2 are the energies of the initial and final states, respectively. The nature of this radiation can further be characterized by treating the oscillating field as composed of contributions from an oscillating electric or magnetic dipole, quadrupole, octupole, etc. If the probability of the transition between two quantum states is large, we speak of an "allowed"

[7] G. Blasse and A. Bril, J. chem. Phys. 50, 2974, 1969.
[8] W. C. Nieuwpoort and G. Blasse, Solid State Comm. 4, 227, 1966.

transition, which may be accompanied by intense radiation. If the probability of the transition is very small or zero, the transition is described as "forbidden". If the transition takes place through the emission of radiation, its probability per unit time is equal to the reciprocal of the lifetime of the initial quantum state.

Whether or not a transition is allowed depends on whether the changes of the various quantum numbers corresponding to the transition obey certain selection rules. It makes a difference here what kind of radiation is absorbed or emitted, whether for example it is electric-dipole radiation, magnetic-dipole radiation, electric-quadrupole radiation, etc. Thus it is possible for a particular transition to be forbidden for electric-dipole radiation but allowed for magnetic- or for electric-quadrupole radiation. This is also bound up with the symmetry properties of the wave functions that describe the initial and final states (see below). The maximum value of the transition probability for electric- and for magnetic-dipole radiation is approximately 10^8 and 10^3 per second, respectively. The magnetic-dipole radiation is much less intense than the electric, and is observable only when the electric-dipole radiation is strictly forbidden. The intensity of quadrupole and higher multipole radiation is so much less that it can in general be disregarded.

Parity

The symmetry properties referred to above are known as the parity of the functions. A wave function $f(x, y, z)$ is said to have *even* parity (or parity + 1) when

$$f(-x, -y, -z) = f(x, y, z)$$

and *odd* parity (or parity −1) when

$$f(-x, -y, -z) = -f(x, y, z).$$

In the latter case the integral of the function over the whole coordinate space is always 0, since the contributions of points such as (x, y, z) are exactly compensated by those from $(-x, -y, -z)$.

In wave mechanics the probability of a transition between an initial state with wave function ψ_1 and a final state with wave function ψ_2 can be described by equations of the form

$$\int \psi_1(r)\, g(r)\, \psi_2^*(r) dr.$$

Here $g(r)$ is a function characterizing the type of radiation whose probability of emission is to be calculated. Thus $g(r)$ for electric-dipole radiation is a first-degree function of r, for electric-quadrupole radiation a second-degree function of r and for magnetic-dipole radiation a function of $r dr/dt$. Since the integral can differ from 0 only if the function $\psi_1 g(r) \psi_2^*$ is even, electric-dipole radiation — $g(r)$ odd — can only be emitted during transitions between states for which $\psi_1 \psi_2^*$ is also odd, that is to say when ψ_1 and ψ_2 differ in parity (Laporte's selection rule). It is also immediately clear from this why electric-dipole transitions between two terms with the same electron configuration are forbidden, and why for the same transition electric-multipole radiation can have a markedly different transition probability from magnetic-multipole radiation of the same order.

Summary I. The article begins with a survey of the physical processes that can take place in a phosphor showing characteristic emission. A discussion then follows of the absorption and emission spectra of a number of ions of the rare-earth metals. The energy-level diagrams of these ions consist partly of discrete levels ($4f^n$ states) and partly of broad bands ($4f^{n-1}5d$ and/or charge-transfer states). Transitions between the ground state and the latter states in the broad bands are permitted. For Ce^{3+} and Eu^{2+} these transitions are also observed in emission. Electric-dipole transitions between the 4f levels are strictly forbidden. Nevertheless their occurrence in both absorption and emission is observed. The selection rules involved and the breaking of these rules are illustrated in the context of the Eu^{3+} ion. This ion shows a red or an orange emission, depending on the nature of the host lattice. This phenomenon is explained in terms of well-established theories on transitions between 4f levels.

II. The efficiency of phosphors excited in the activator

As explained in part I of this article, phosphors can be excited in two fundamentally different ways. In the one case the excitation energy is absorbed by the activator itself; in the other case the excitation energy is absorbed elsewhere, after which it must be transferred to the activator. In this second article we shall be concerned exclusively with the first case.

The conversion efficiency of a phosphor can be numerically expressed in various ways. We shall refer only to the quantum efficiency, that is to say the ratio of the number of quanta emitted by the phosphor to the number of quanta it absorbs. Phosphors of technical interest have quantum efficiencies of 70 to 90%.

The problems we shall touch on in this part of the article are among the most difficult and least well understood problems of luminescence. We shall deal first with the question of why certain ions (or groups of ions) luminesce and others do not. We shall then go on to show that the chemical composition of the environment of a luminescent centre can strongly influence the efficiency. Finally, relations will be sought between the efficiency and the other physical properties of the centre.

The configurational-coordinate model for characteristic luminescence

Since the end of the thirties, various models have been proposed to explain the presence or absence of characteristic luminescence. These models are based on what is termed the configurational-coordination diagram of the luminescent centre. We shall start by considering this type of diagram.

transition, which may be accompanied by intense radiation. If the probability of the transition is very small or zero, the transition is described as "forbidden". If the transition takes place through the emission of radiation, its probability per unit time is equal to the reciprocal of the lifetime of the initial quantum state.

Whether or not a transition is allowed depends on whether the changes of the various quantum numbers corresponding to the transition obey certain selection rules. It makes a difference here what kind of radiation is absorbed or emitted, whether for example it is electric-dipole radiation, magnetic-dipole radiation, electric-quadrupole radiation, etc. Thus it is possible for a particular transition to be forbidden for electric-dipole radiation but allowed for magnetic- or for electric-quadrupole radiation. This is also bound up with the symmetry properties of the wave functions that describe the initial and final states (see below). The maximum value of the transition probability for electric- and for magnetic-dipole radiation is approximately 10^8 and 10^3 per second, respectively. The magnetic-dipole radiation is much less intense than the electric, and is observable only when the electric-dipole radiation is strictly forbidden. The intensity of quadrupole and higher multipole radiation is so much less that it can in general be disregarded.

Parity

The symmetry properties referred to above are known as the parity of the functions. A wave function $f(x, y, z)$ is said to have *even* parity (or parity + 1) when

$$f(-x, -y, -z) = f(x, y, z)$$

and *odd* parity (or parity −1) when

$$f(-x, -y, -z) = -f(x, y, z).$$

In the latter case the integral of the function over the whole coordinate space is always 0, since the contributions of points such as (x, y, z) are exactly compensated by those from $(-x, -y, -z)$.

In wave mechanics the probability of a transition between an initial state with wave function ψ_1 and a final state with wave function ψ_2 can be described by equations of the form

$$\int \psi_1(r)\, g(r)\, \psi_2^*(r) dr.$$

Here $g(r)$ is a function characterizing the type of radiation whose probability of emission is to be calculated. Thus $g(r)$ for electric-dipole radiation is a first-degree function of r, for electric-quadrupole radiation a second-degree function of r and for magnetic-dipole radiation a function of $r dr/dt$. Since the integral can differ from 0 only if the function $\psi_1 g(r) \psi_2^*$ is even, electric-dipole radiation — $g(r)$ odd — can only be emitted during transitions between states for which $\psi_1 \psi_2^*$ is also odd, that is to say when ψ_1 and ψ_2 differ in parity (Laporte's selection rule). It is also immediately clear from this why electric-dipole transitions between two terms with the same electron configuration are forbidden, and why for the same transition electric-multipole radiation can have a markedly different transition probability from magnetic-multipole radiation of the same order.

Summary I. The article begins with a survey of the physical processes that can take place in a phosphor showing characteristic emission. A discussion then follows of the absorption and emission spectra of a number of ions of the rare-earth metals. The energy-level diagrams of these ions consist partly of discrete levels ($4f^n$ states) and partly of broad bands ($4f^{n-1}5d$ and/or charge-transfer states). Transitions between the ground state and the latter states in the broad bands are permitted. For Ce^{3+} and Eu^{2+} these transitions are also observed in emission. Electric-dipole transitions between the 4f levels are strictly forbidden. Nevertheless their occurrence in both absorption and emission is observed. The selection rules involved and the breaking of these rules are illustrated in the context of the Eu^{3+} ion. This ion shows a red or an orange emission, depending on the nature of the host lattice. This phenomenon is explained in terms of well-established theories on transitions between 4f levels.

II. The efficiency of phosphors excited in the activator

As explained in part I of this article, phosphors can be excited in two fundamentally different ways. In the one case the excitation energy is absorbed by the activator itself; in the other case the excitation energy is absorbed elsewhere, after which it must be transferred to the activator. In this second article we shall be concerned exclusively with the first case.

The conversion efficiency of a phosphor can be numerically expressed in various ways. We shall refer only to the quantum efficiency, that is to say the ratio of the number of quanta emitted by the phosphor to the number of quanta it absorbs. Phosphors of technical interest have quantum efficiencies of 70 to 90%.

The problems we shall touch on in this part of the article are among the most difficult and least well understood problems of luminescence. We shall deal first with the question of why certain ions (or groups of ions) luminesce and others do not. We shall then go on to show that the chemical composition of the environment of a luminescent centre can strongly influence the efficiency. Finally, relations will be sought between the efficiency and the other physical properties of the centre.

The configurational-coordinate model for characteristic luminescence

Since the end of the thirties, various models have been proposed to explain the presence or absence of characteristic luminescence. These models are based on what is termed the configurational-coordination diagram of the luminescent centre. We shall start by considering this type of diagram.

Various configurational-coordination diagrams are shown in *figs. 13, 14* and *15*. The potential energy of the luminescent centre in the crystal lattice is plotted as a function of the configurational coordinate r. To see what r represents, we take a metal ion M^{n+} surrounded by four O^{2-} ions at the corners of a tetrahedron. These ions will vibrate, that is to say oscillate in relation to one another while the centre of mass of the system remains in its place. An example of such a vibration is what is termed the symmetric valence vibration. Here the M^{n+} ion remains stationary while the O^{2-} ions vibrate in phase along the M-O bonding axis. When drawing the configurational-coordinate diagram it is assumed (on not unreasonable grounds) that we need only take this symmetric valence vibration into account. The quantity r then represents the distance M-O.

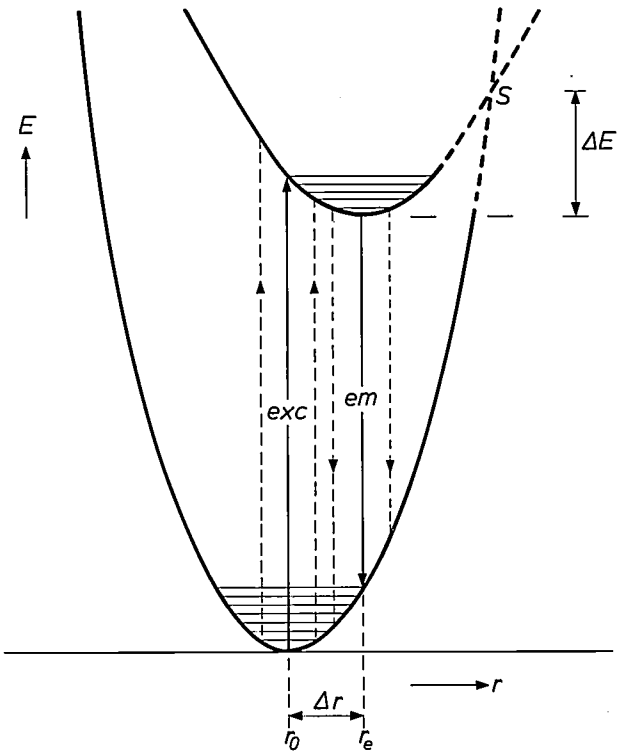


Fig. 13. Configurational-coordinate diagram of a luminescent centre. The potential energy E of the centre in the lattice is plotted as a function of the configurational coordinate r for the ground state and the first excited state. In practice r is identified with the distance between the central cation and the surrounding anions. Vibrational states are represented schematically by horizontal lines in the parabolae. The excitation and emission transitions correspond to *vertical* transitions between the two curves. Since $\Delta r \neq 0$, the emission shows a Stokes shift (wavelength of the emission is longer than that of the excitation). Since the centre can be in various vibrational states both at the ground level and at the excited level, and since $\Delta r \neq 0$, the transitions occur in a broad band of energies (schematically indicated by vertical dashed lines). Non-radiative return from the excited state to the ground state is possible via the point of intersection S of the two curves. This requires an activation energy ΔE, which can be supplied at higher temperatures (thermal quenching of the emission). In the region where the two parabolae intersect, the curve is marked by dashes since the situation is actually more complicated than is indicated here. This is due to interaction between the ground and the excited state at the situation of the intersection point. Our treatment is not invalidated by this.

At a temperature of absolute zero the luminescent centre will occupy the lowest vibrational level of the ground state. The ions surrounding the central ion vibrate about their equilibrium positions situated at a

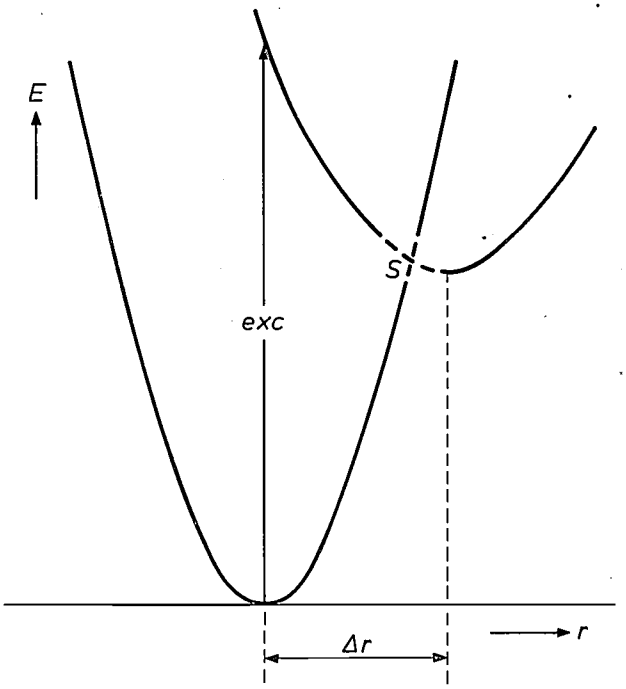


Fig. 14. The Seitz model for explaining the absence of luminescence; see fig. 13. The minimum of the curve for the excited state lies outside the curve for the ground state; luminescence is then not possible.

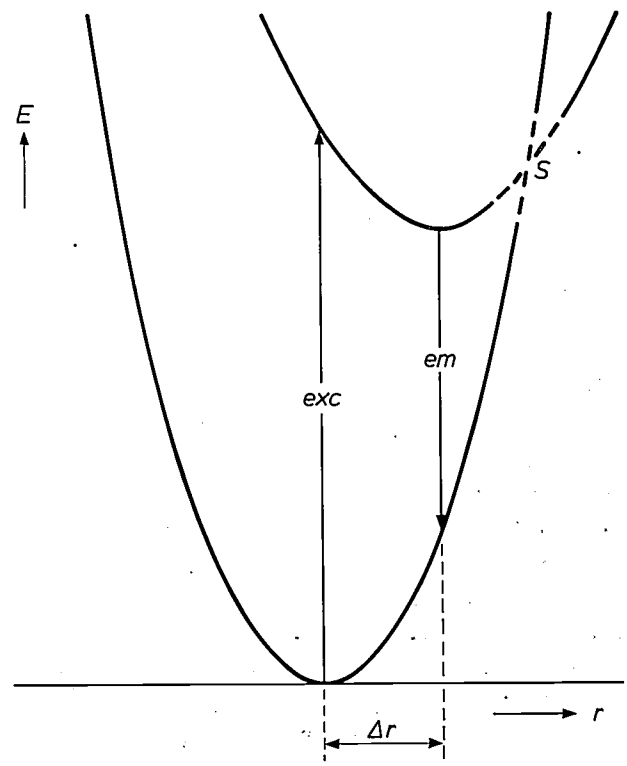


Fig. 15. The Dexter-Klick-Russell model for explaining a low luminescence efficiency or the absence of luminescence. The intersection point S of the two curves lies below the vibrational level reached after excitation. The non-radiative return to the ground state now requires no activation energy (as it did in fig. 13).

distance r_0 from the central ion. At higher temperature, higher vibrational levels may be occupied. In fig. 13 the horizontal lines represent vibrational states. Due to the absorption of radiation of the appropriate wavelength (in our case very often UV radiation) the centre is raised to an excited state. Since the equilibrium distance r_e of the excited state will not in general be equal to that of the ground state, and since the centre may be at different vibrational levels, this transition will correspond to a fairly broad absorption band. The fact that the optical absorption corresponds to a vertical transition in fig. 13 is attributable to the rapid nature of electronic transitions as compared to vibrational movements, which involve the (heavier) nuclei.

nescence of some phosphors depends on temperature. It can be seen that the luminescence decreases with rising temperature.

With the aid of the simple model in fig. 13 (the Mott-Seitz model) we can therefore explain

a) the broad-band character of the emission and absorption of many centres;
b) the Stokes shift of the emission;
c) the temperature dependence of the emission.

If now the equilibrium configuration of the excited state lies outside the curve of the ground state (fig. 14), then after excitation the intersection point of both curves is reached before the above-mentioned equilibrium configuration, and the system relaxes non-

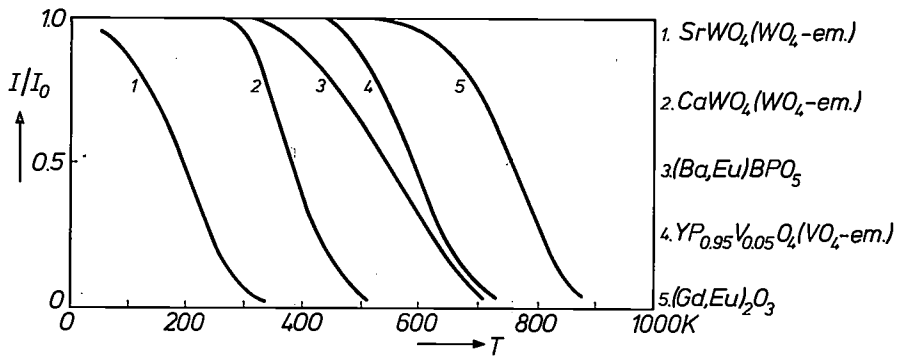


Fig. 16. Thermal quenching of the luminescence. The relative intensity of the luminescence from a number of phosphors, obtained by excitation with 254 nm radiation, plotted as a function of absolute temperature. The relative intensity is the ratio of the intensity I to the value I_0 approached asymptotically as T approaches the absolute zero point.

Once in an excited state, the system will relax towards the equilibrium state (of the excited level) by dissipating heat. From this state or nearby levels the system returns to the ground state, thereby emitting radiation. The emission too, therefore, consists of a broad band. Line emission is found only in the exceptional case where the configurational-coordinate curves are identical in shape and have the same equilibrium distance, as for example in the case of the rare-earth ions. Because of the above-mentioned heat dissipation, the emission always lies at a lower energy than the absorption. This displacement of emission with respect to absorption is known as the Stokes shift.

From the configurational-coordinate diagram in fig. 13 we can now understand also why the emission will be quenched at higher temperature. If the luminescent centre is in the equilibrium configuration of the excited state, it may also, as a result of thermal activation, occupy a vibrational level situated at the point of intersection S of the curves representing the excited and ground states (activation energy ΔE). Having arrived here, the centre will return non-radiatively to the equilibrium configuration of the ground state, dissipating heat in the process. *Fig. 16* shows the way in which the luminescence of some phosphors depends on temperature. It can be seen that the luminescence decreases with rising temperature.

radiatively to the ground state. No emission is then possible. The radiationless return to the ground state is temperature-independent. This is the model which Seitz proposed to explain the absence of luminescence in certain cases [9]. In other words, the condition for the absence of luminescence is a large difference between the equilibrium distance of the excited state and that of the ground state.

D. L. Dexter, C. C. Klick and G. A. Russell later proposed a different model [10]. This shows that even under less rigorous circumstances than in fig. 14 non-radiative transitions to the ground state may occur (fig. 15). The characteristic feature of the situation in fig. 15 is that the intersection point S of the two curves is lower than the level reached after excitation. When, after excitation, the system now relaxes, while vibrating, to the equilibrium position of the excited state, the intersection point of the two curves is passed. Here too, a temperature-independent, radiationless return to the ground state can take place.

We learn from these models that the difference Δr between the equilibrium configuration of the excited state and that of the ground state must be small if luminescence is to occur.

Which ions exhibit characteristic luminescence?

Taking as our starting assumption that the value of Δr is the principal criterion for the presence or absence of luminescence, we looked for a method of learning something about Δr. The equilibrium distance of the excited state is known in only a few cases. It proved possible to make an estimate of Δr with the aid of the radii of electron orbits as calculated by J. T. Waber and D. T. Cromer [11], and in this way we were able to propose a criterion for the occurrence or non-occurrence of luminescence [12]. Only the absolute value of Δr is important in this respect: in other words, we are only interested in the shift of the curve of the ground state relative to that of the excited state. A few cases will serve to illustrate our method.

We consider first of all the Sb^{3+} ion, a well-known luminescent centre. The ground state of the electrons in the outer shell is 5s. Upon excitation one of the 5s electrons is raised to 5p; the emission corresponds to the reverse transition. For all elements Waber and Cromer have calculated radii that correspond to the maximum in the charge density of an electron orbit. For the 5s orbit of Sb the radius is 0.97 Å, for the 5p orbit 1.16 Å. We made the assumption that the difference between the electron orbit radii, 0.19 Å, is equal to the difference Δr between the equilibrium distances of the luminescent centre.

Let us now look at the Sb^{5+} ion. Characteristic luminescence has never been observed from this ion. The ground state is $4d^{10}$. Excitation consists in raising one of the d electrons to the 5s or 5p orbit. For the 4d orbit Waber and Cromer gave 0.44 Å, for 5s 0.97 Å and for 5p 1.16 Å. The minimum value of Δr that follows from this is 0.53 Å. This is a much greater value than we found for Sb^{3+} (0.19 Å). According to the models in figs. 13, 14 and 15, this difference in Δr might explain why Sb^{3+} luminesces and Sb^{5+} does not.

We applied calculations of this type to all ions or groups of ions where the electron transitions involved in the absorption and emission are known. Some examples are presented in *Table V*. It was found that a necessary but not sufficient condition for the occurrence of luminescence is that the calculated absolute value of Δr must be smaller than 0.3 Å. This condition is not sufficient because the calculation of Δr applies to the free activator ion or ionic group, neglecting the influence of the surrounding lattice, which can be very great (see below). We shall now comment on the other examples in Table V.

Table V. Excitation transitions of some luminescent centres and the electron-orbit radii used for calculating Δr.

centre	excitation transition	electron-orbit radius (Å) ground state	electron-orbit radius (Å) excited state	Δr
$Nb^{5+}(O^{2-})_n$	2p-4d	4p(Nb) 0.59	4d(Nb) 0.75	0.16
$W^{6+}(O^{2-})_n$	2p-5d	5p(W) 0.58	5d(W) 0.75	0.17
Sb^{3+}	5s-5p	5s(Sb) 0.97	5p(Sb) 1.16	0.19
Sb^{5+}	4d-5s	4d(Sb) 0.44	5s(Sb) 0.97	0.53
Ce^{3+}	4f-5d	5p(Ce) 0.82	5d(Ce) 1.07	0.25
Eu^{2+}	4f-5d	5p(Eu) 0.74	5d(Eu) 0.98	0.24
$Eu^{3+}(O^{2-})_n$	2p-4f	5p(Eu) 0.74	5p(Eu) 0.74	0
Mn^{2+}, $Mn^{4+}(O^{2-})_n$	2p-3d	3d(Mn)	3d(Mn)	0

The Ce^{3+} ion. As explained in part I, the excitation is a 4f-5d transition and the emission a 5d-4f transition. Since the 4f shell lies inside the ion, we take the radius for the ground state to be that of the outer orbit of the ion in the ground state. This is the 5p orbit with $r = 0.82$ Å. For the 5d orbit $r = 1.07$ Å, so that $\Delta r = 0.25$ Å. And, indeed, the Ce^{3+} ion shows luminescence in many lattices. Similar reasoning applies to the Eu^{2+} ion.

The WO_4^{2-} group. In the ground state the outer filled orbit of the tungsten ion is 5p ($r = 0.58$ Å). The excitation (absorption) band is the consequence of a charge-transfer process: one of the 2p electrons of the oxygen ions goes to the empty 5d orbit of the W^{6+} ion. This orbit has a radius of 0.75 Å, so that $\Delta r = 0.17$ Å. The tungstate group indeed shows luminescence. The niobate group behaves analogously (charge transfer from the 2p orbits of the oxygen ions to the empty outer 4d orbit of the Nb^{5+} ion).

The Eu^{3+} ion. This ion too can be excited in a charge-transfer state. Since the charge transfer now takes place from the 2p orbits of the oxygen ions to the 4f shell situated inside the Eu^{3+} ion, Δr according to our method of calculation is zero. The result $\Delta r = 0$ must obviously be interpreted to mean that Δr is very small. Very high quenching temperatures are in fact found for the luminescence of Eu^{3+}. The same applies to the well-known activators Mn^{2+} and Mn^{4+}, where the charge is transferred from the anions to the already partly filled 3d shell of the manganese ion.

It is really rather surprising that such a rough method works so well. For we are now in a position, with the criterion that Δr must be smaller than 0.3 Å, to select those groups of ions that are capable in principle of exhibiting luminescence. The absence of characteristic luminescence in other cases can thus be attributed to a too high value of Δr.

As stated, the occurrence of luminescence depends to a very great extent on the nature of the host lattice. Thus, the WO_4 group in $CaWO_4$ luminesces very efficiently, both at low temperature (77 K) and at rela-

[9] F. Seitz, Trans. Faraday Soc. **35**, 74, 1939.
[10] D. L. Dexter, C. C. Klick and G. A. Russell, Phys. Rev. **100**, 603, 1955.
[11] J. T. Waber and D. T. Cromer, J. chem. Phys. **42**, 4116, 1965.
[12] G. Blasse, J. chem. Phys. **48**, 3108, 1968.

tively high temperature (300 K). In the isomorphous $BaWO_4$, however, this group shows a weak luminescence only at temperatures of 77 K and lower. The Eu^{3+} ion too is an efficient luminescent centre in $Y_3Al_5O_{12}$, whereas in $LaAlO_3$ the efficiency is very much lower. We shall now take a closer look at this effect of the host lattice.

Dependence of the efficiency of luminescence on the host lattice

In the present state of knowledge it is not possible to state in *quantitative* terms how the efficiency of the luminescence depends on the host lattice. Proceeding from the idea developed above that it is the magnitude of Δr that determines the *quenching temperature* of the luminescence, and hence also the efficiency at room temperature, we were able to indicate a rough relationship between the quenching temperature of the luminescence and the radius and charge of the cations of the host lattice [13]. Unlike the previous theoretical treatment, where only the absolute value of Δr was important, in this treatment the sign of Δr plays a significant part.

In fig. 13 it is assumed that Δr is positive, in other words that the luminescent centre expands after excitation. However, Δr may be negative as well as positive. This was shown long ago by F. E. Williams [14] in his pioneering work on (K,Tl)Cl. The Tl^+ ion situated at a potassium site in the host lattice KCl is the luminescent centre. In the ground state the Tl^+ ion has two 6s electrons in its outer shell; in the state from which emission occurs the ion has one electron in the 6s orbit and one electron in the 6p orbit. The emission and excitation thus correspond here to a transition in which there is no transfer of electrons to another ion (unlike a charge-transfer transition, where an electron transfers from a neighbouring anion to the cation).

Upon excitation of the Tl^+ ion the electron-charge distribution of the ion moves somewhat farther away from the nucleus (due to the transition of one of the electrons from $6s \rightarrow 6p$). The negative charge cloud becomes more diffuse and as a result the cation effectively assumes a greater positive charge. It therefore attracts the negative ions more strongly, so that the equilibrium distance of the luminescent centre in the excited state is smaller than that of the ground state. Williams's calculations showed that in the case of Tl^+ in KCl, Δr has a negative value and is 0.2 Å. The reasoning adopted applies to all cases where the luminescent cation itself is excited. These include, for example, the 4f-5d transitions of the ions of the rare-earth metals.

A positive value of Δr is to be expected when the anion is excited. The electron cloud becomes more diffuse and the anion thus effectively assumes a greater positive charge (that is to say becomes less negative) and therefore attracts the cations less strongly, so that the equilibrium distance becomes greater. The only case of excitation of anion electrons of interest to us is that of the charge-transfer transitions already dealt with.

In the considerations that follow, we shall divide the luminescent centres into two groups, those with $\Delta r > 0$ (excitation of anion electrons) and those with $\Delta r < 0$ (excitation of cation electrons). We shall now see how Δr depends on the size of the host lattice ion for which the activator ion has been substituted and on the magnitude of the charge and size of the host-lattice ions surrounding the activator.

If the activator ion is larger than the host-lattice ion which it replaces, e.g. Eu^{3+} (ionic radius 0.98 Å) or Ce^{3+} (1.07 Å) in an Lu^{3+} host lattice (0.85 Å), the environment of the activator will be compelled to expand in order to make room for the activator. If the activator is raised to the excited state, and if this is accompanied by an increase of the equilibrium distance (anion excitation, $\Delta r > 0$), then the environment of the activator will have to expand yet further. Since this expansion costs energy, the lattice will tend to oppose the expansion of the luminescent centre, in other words Δr will be relatively small.

The opposite is the case if the activator is located at a site which is occupied in the host lattice by a larger ion, for example Eu^{3+} (0.98 Å) in an La^{3+} compound (1.14 Å). Upon excitation in the charge-transfer absorption band of the Eu^{3+} ion ($\Delta r > 0$) we shall then find Δr to be relatively large.

If the excitation of the activator ion occurs by an electronic transition at the ion itself ($\Delta r < 0$), the situation is reversed as compared with that involving an activator with $\Delta r > 0$ (charge transfer). If, for example, the site in the lattice occupied by an activator with $\Delta r < 0$ is too small, then the environmental expansion that occurs for the activator in the ground state is partly reversed by excitation. In that case Δr is not constrained to remain small.

The second of the factors just mentioned that determine Δr, the influence of the cations surrounding the luminescent centre, may be sketched as follows. Small, highly charged cations will give the host lattice great bonding strength. In such a rigid lattice it is evident that Δr will be relatively small (irrespective of whether Δr is positive or negative). If the lattice contains large cations of low charge, the bonding in the lattice will be weak. Such a lattice can thus comply with the tendency of the activator to expand or shrink upon excitation. The absolute value of Δr in this case will therefore be relatively large.

Table VI summarizes these conclusions. We shall now take a number of examples to show that the

Table VI. The relation between the quenching temperature T_q of the emission and the radius and charge of the host lattice cations, in accordance with the thermal-quenching model proposed in this article.

radius and charge of cations	$\Delta r < 0$ (e.g. Tl^+, Eu^{2+}, Ce^{3+})	$\Delta r > 0$ (e.g. Eu^{3+}, WO_4)
activator ion larger than host-lattice ion	T_q low	T_q high
activator ion smaller than host-lattice ion	T_q high	T_q low
host lattice with small, highly charged cations	T_q high	T_q high
host lattice with large cations of low charge	T_q low	T_q low

Table VII. Relation between the quenching temperature of the emission of the WO_4 group and the radius of the other ions of the lattice.

compound	quenching temperature (K)	ionic radius (Å)	
$CaWO_4$	410	Ca^{2+}	0.99
$SrWO_4$	280	Sr^{2+}	1.12
$BaWO_4$	100	Ba^{2+}	1.34
Li_2WO_4	300	Li^+	0.68
Na_2WO_4	100	Na^+	0.94

Table VIII. Some properties of the isomorphous phosphors $R.E._{1.9}Eu_{0.1}SO_6$. The radius of the Eu^{3+} ion is 0.98 Å.

rare earth	Lu	Y	Gd	La
radius $R.E.^{3+}$ ion (Å)	0.85	0.92	0.97	1.14
position of charge-transfer absorption band (nm)	270	270	275	290
quantum efficiency q (%) for excitation with this wavelength	60	55	50	35
quenching temperature (K)	750	750	700	600

experimental results are in good agreement with the predictions in Table VI. The experimental quantity is in all cases the quenching temperature of the luminescence; the prediction relates to Δr. We have already shown that a small value of Δr corresponds to a high quenching temperature and a large value of Δr to a low quenching temperature. To limit the number of variables we shall as far as possible compare host lattices possessing the same crystal structure.

Highly charged ions with inert-gas structure

In oxides the groups consisting of highly charged cations with an inert-gas structure and surrounding oxygen ions, such as the vanadate group $V^{5+}O_4$ and the tungstate groups $W^{6+}O_4$ and $W^{6+}O_6$, are known luminescent centres. The absorption band in which these centres can be excited corresponds to the transition of an electron from the surrounding O^{2-} anions to the central ion (charge transfer by means of anion excitation, hence $\Delta r > 0$). We have found [15] that the quenching temperature of the emission from centres of this type depends to a great extent on the other cations in the lattice, in the manner predicted in Table VI. Some examples are listed in *Table VII*. These illustrate in particular the influence of the surroundings of the activator.

For Ti^{4+} especially, a great deal of information is available concerning the influence which the size of the space occupied by the activator ion Ti^{4+} in the host lattice exerts on the quenching temperature. These data are due to F.A.Kröger [16]. He found the highest quenching temperatures for Ti^{4+} ($r = 0.68$ Å) at Si^{4+} sites ($r = 0.42$ Å) in silicates. In germanates at Ge^{4+} sites ($r = 0.53$ Å) and in stannates at Sn^{4+} sites ($r = 0.71$ Å) the quenching temperature of the emission from the Ti^{4+} ions is appreciably lower.

We see then that for this group of activators our model provides a good explanation for the relation between the experimentally determined quenching temperature and the ionic radius of the cations in the host lattice. We shall now turn to some rather more complicated cases.

Trivalent europium (Eu^{3+})

Like the above-mentioned group of activators, the Eu^{3+} ion can be excited in a charge-transfer absorption band (but also in discrete 4f levels). The difference compared with the foregoing group is that the emissive transition is not the reversed-excitation transition [17] (see I, fig. 2, p. 306). If the excitation takes place in the sharp 4f levels, then Δr is zero, because in this case the configurational-coordinate curves have the same shape and the same equilibrium distance. Thermal de-excitation is thus practically impossible. If, on the other hand, the excitation takes place in the charge-transfer state (see fig. 19), Δr has a positive value. The activation energy for non-radiative de-excitation is thus lower than in the case of excitation in the 4f levels, and therefore the quenching temperature is lower.

It is a known fact that the quenching temperature of the Eu^{3+} emission for excitation in the charge-transfer absorption band decreases upon an increase in the radius of the cation for which Eu^{3+} has been substituted in the host lattice (see *Table VIII*). One would also expect this from Table VI, since Δr is positive. Non-

[13] G. Blasse, J. chem. Phys. **51**, 3529, 1969.
[14] F. E. Williams, J. chem. Phys. **19**, 457, 1951.
[15] G. Blasse and A. Bril, Z. phys. Chemie Neue Folge **57**, 187, 1968.
[16] F. A. Kröger, Some aspects of the luminescence of solids, Elsevier, Amsterdam 1948.
[17] G. Blasse and A. Bril, Philips Res. Repts. **23**, 461, 1968.

radiative losses in the Eu^{3+} centre are attributed in this model to a direct, non-radiative transition from the charge-transfer state to the ground state.

This explains why the quenching temperature of the emission of the well-known phosphor $(Y,Eu)_2O_3$ (about 800 K) is much lower upon excitation in the charge-transfer absorption band than upon excitation in the discrete 4f levels. In the latter case non-radiative de-excitation is virtually impossible [18].

In accordance with Table VI, the Eu^{3+} emission is found to have a high quenching temperature in the case of small highly-charged cations in the host lattice. This is found, for example, in the Eu^{3+} phosphors where the host lattice is a borate or a silicate.

We see, then, that the prediction given in Table VI is also borne out for the Eu^{3+} ion. We shall now consider the case where Δr is negative after excitation, taking the Eu^{2+} ion as an example.

Divalent europium (Eu^{2+})

The excitation band of the Eu^{2+} ion corresponds to a 4f-5d transition. This is thus a case of cation excitation, so that we have $\Delta r < 0$. All efficient Eu^{2+} phosphors contain small highly-charged cations. Examples are $(Sr,Eu)_2P_2O_7$, $(Sr,Eu)Al_2O_4$, $(Ba,Eu)Al_{12}O_{19}$ and $(Ba,Eu)BPO_5$ [19]. The quenching temperature of these phosphors is relatively high, as one would expect.

What now will be the effect of the size of the ion which has been replaced by Eu^{2+} ($r = 1.13$ Å)? In practice these will be the ions Ca^{2+} (0.99 Å), Sr^{2+} (1.12 Å) and Ba^{2+} (1.34 Å). From Table VI we would expect Δr to decrease in the case of the Eu^{2+} ion, going from Ca^{2+} to Ba^{2+}. However, a complication arises. If we put an Eu^{2+} ion at one of the Ba^{2+} sites of a Ba^{2+} host lattice, the site will be too large for the Eu^{2+} ion, and therefore Δr will be relatively small (since $\Delta r < 0$). But in the neighbourhood of the Eu^{2+} ion there are other Ba^{2+} ions. These are very large, so that in the environment of an activator with $\Delta r < 0$ the influence of the size of the surrounding ions and the influence of the size of the substituted ion will oppose one another.

This is reflected in the experimental results (*Table IX*). In host lattices where the concentration of the alkaline-earth metal ions is relatively low (e.g. $BaAl_{12}O_{19}$, $BaPBO_5$, $(Sr,Ba)_2MgSi_2O_7$), the quenching temperature of the Eu^{2+} emission increases, going from Ca to Ba. The influence of the alkaline-earth metal ions as neighbours of the Eu^{2+} centre is evidently of little significance here. If the concentration of alkaline-earth ions increases, the quenching temperature going from Ca to Ba ($Ba_3MgSi_2O_8$) shows a smaller increase. Scarcely any difference is found in the quenching temperature of the Eu^{2+} emission in Ba_2SiO_4, Sr_2SiO_4 and Ca_2SiO_4.

Table IX. Quenching temperature of the Eu^{2+} emission in various groups of isomorphous host lattices [19].

host lattice	quenching temperature (K)
$CaAl_{12}O_{19}$	420
$SrAl_{12}O_{19}$	460
$BaAl_{12}O_{19}$	670
$CaBPO_5$	325
$SrBPO_5$	550
$BaBPO_5$	670
$Ca_2MgSi_2O_7$	275
$Sr_2MgSi_2O_7$	305
$Sr_{0.5}Ba_{1.5}MgSi_2O_7$	350
$Ca_3MgSi_2O_8$	505
$Sr_3MgSi_2O_8$	520
$Ba_3MgSi_2O_8$	545
Ca_2SiO_4	390
Sr_2SiO_4	410
Ba_2SiO_4	420

Here too, it can be said that the experimental results are in good agreement with Table VI. It is difficult, however, to make any prediction concerning the quenching temperature, since two mutually opposing effects are involved. This is particularly the case with the Ce^{3+} and the Tb^{3+} ions [13].

Cerium (Ce^{3+}) and terbium (Tb^{3+})

Like the Eu^{2+} ion, the Ce^{3+} ion and the Tb^{3+} ion can be excited in a 4f-5d transition, so that $\Delta r < 0$. Whereas in the case of Eu^{3+}, where $\Delta r > 0$, there is a clear relationship between the quenching temperature and the size of the ion for which Eu^{3+} has been substituted (La^{3+}, Gd^{3+}, Y^{3+}, Lu^{3+}, see Table VIII), we have never been able to find any such relation for Ce^{3+} and Tb^{3+} substituted for the same ions. This result is, however, in agreement with our model. It should also be mentioned that Ce^{3+} and Tb^{3+} are examples of activators whose emission shows a high quenching temperature in silicates, borates, phosphates etc. Here too, the small highly-charged ions clearly exert a marked effect.

Summarizing it can be said that we have a relation that gives us some idea of the connection between the quenching temperature of the emission of an activator on the one hand and of the size and charge of the host-lattice cations on the other.

For activators that can be excited by charge transfer from anion to cation, we have found another relation which also provides information on the quenching temperature of the emission. This will now be discussed.

[18] See e.g. M. J. Weber, Phys. Rev. **171**, 283, 1968.

[19] G. Blasse, W. L. Wanmaker and J. W. ter Vrugt, J. Electrochem. Soc. **115**, 673, 1968.
G. Blasse and A. Bril, Philips Res. Repts. **23**, 201, 1968.
G. Blasse, A. Bril and J. de Vries, J. inorg. nucl. Chem. **31**, 568, 1969.

Influence of the energy level of the charge-transfer state on the quenching temperature

Luminescent centres that can be excited by charge transfer from the anions to the central cation show as a rule a higher emission quenching temperature the shorter the wavelength of the charge transfer absorption band. We are concerned here with chemically diverse activators such as Eu^{3+}, NbO_6, WO_6 and uranyl (UO_2^{2+}). Some examples are given in Table VIII and *fig. 17*.

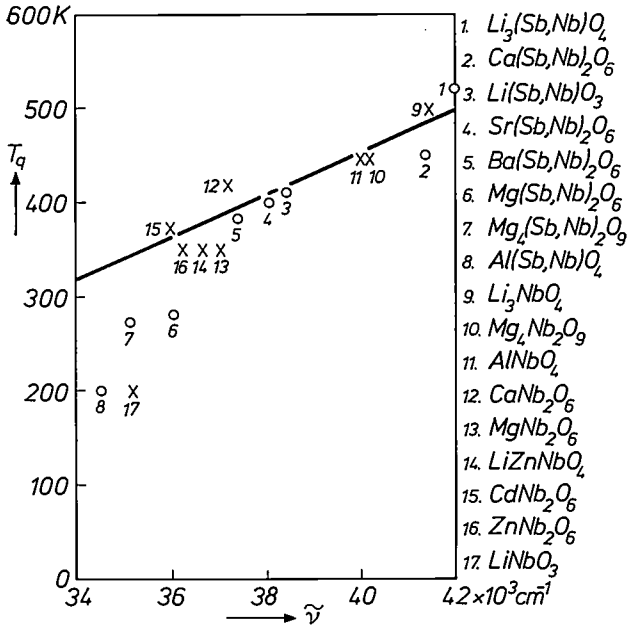


Fig. 17. Quenching temperature T_q of the niobate emission in various host lattices as a function of the wave number ν of the maximum of the charge-transfer absorption band. In the region 42 000-37 000 cm^{-1} a linear relation exists. The crosses are measured points for niobates, the circles for antimonates activated with Nb.

In fig. 17 the quenching temperature of the emission of the niobate group is plotted as a function of the position of the charge-transfer absorption band. Over a wide region the relation found is in fact a linear one. This relates to host lattices with completely different crystal structures. We observed further that the quantum efficiency of the phosphors that lie on the straight line is in any case high at 100 K, whereas that of the phosphors well below the straight line is low even at 100 K. If the absorption band of the niobate group lies below 34 000 cm^{-1}, no luminescence is observed at all [12]. Similar effects were observed with the tungstate group and the Eu^{3+} ion.

Using the simple configurational-coordinate diagram in figs. 13 and 15 it is possible to get some idea of what is happening. *Fig. 18* shows another configurational-coordinate diagram. Various possibilities are shown for the excited state. The shape of the curves is the same and the value of Δr is constant. The curves *MS* correspond to the Mott-Seitz model, which predicts a high luminescence efficiency for 0 K. The curve *D* corresponds to the Dexter-Klick-Russell model, which allows low efficiencies for 0 K.

We assume that the Mott-Seitz model is valid for the phosphors that lie on the straight line in fig. 17. Their absorption band lies at 37 000 cm^{-1} or higher. Their efficiency at low temperature is high. The curve D_{crit} in

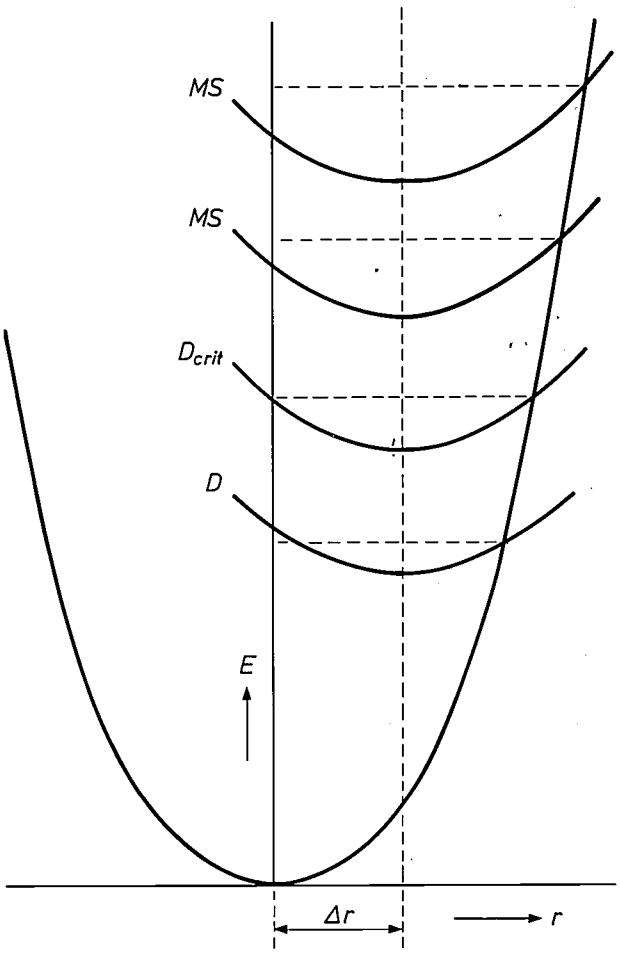


Fig. 18. If the minimum of the curve for the excited state lies within the curve of the ground state, there are two situations possible for the relative positions of the intersections of the first curve with the ordinate axis and with the other curve. In the case of curves *MS* the point of intersection with the ordinate axis is lower than the other intersection point (Mott-Seitz model), in the case of curve *D* it is higher (Dexter-Klick-Russell model). The curve D_{crit} represents the borderline case. The relation between figures 17 and 18 is discussed in the text.

fig. 18 corresponds roughly to the niobate group, which has the absorption band at 37 000 cm^{-1}. If this band lies at lower energy, then the Dexter-Klick-Russell model is apparently applicable. It is even possible to work out this picture quantitatively [12].

It cannot be said, however, that the relation between

quenching temperature and position of the absorption band is explained. We have rather an illustration here of a transition from the Mott-Seitz model to the Dexter-Klick-Russell model under the influence of the host lattice. It should also be noted that the relation is valid only for octrahedrally surrounded, highly-charged ions with an inert-gas structure (NbO_6 and WO_6 groups) and not for the tetrahedrally surrounded ions (VO_4 and WO_4 groups), which are frequently encountered. The variation of Δr for the smaller tetrahedral site probably plays a much more important part than for the octahedral site, which is in fact somewhat too large. What is essential in the model presented in fig. 18, is that Δr remains constant.

The work described in the foregoing led us to predict a number of possible ways in which non-radiative losses might take place in the Eu^{3+} centre of Eu^{3+} phosphors. These will now be discussed.

Non-radiative losses in the Eu^{3+} centre

Fig. 19 gives a schematic (and incomplete) configurational-coordinate diagram for the Eu^{3+} ion in oxides. In drawing the diagram we started from the assumption

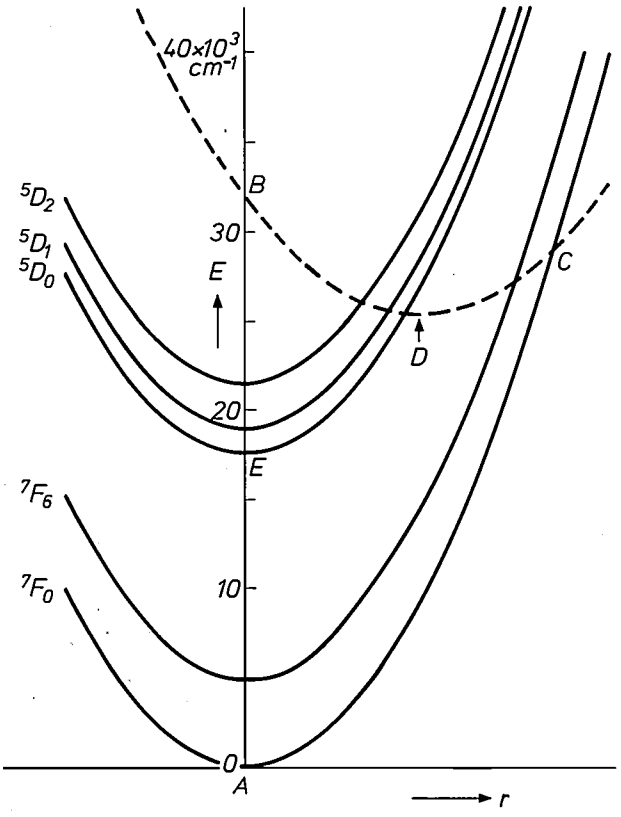


Fig. 19. Schematic configurational-coordinate diagram for the Eu^{3+} ion (in $LaAlO_3$, for example). The curves for the 4f levels (solid line) have the same shape. For clarity only a few 4f levels are shown. The dashed line refers to the charge-transfer state. As explained in the text, non-radiative de-excitation is possible by way of $D \rightarrow E$ (followed by a luminescent transition $E \rightarrow A$), $D \rightarrow C \rightarrow A$, and at higher temperature also $E \rightarrow D \rightarrow C \rightarrow A$.

that the curves relating to the 4f levels all have the same equilibrium distance. This is a very reasonable assumption, because differences only occur deep in the rare-earth ion, and these do not influence the chemical bond. The curve relating to the charge-transfer state, however, is considerably shifted with respect to that for the 4f states ($\Delta r > 0$). We have shown that in this case there are three possible ways in which non-radiative losses may occur [20]. These may be understood as follows.

Upon excitation the system is raised to the charge-transfer state (AB in fig. 19). Giving up vibrational energy to the lattice the system relaxes to point D, the equilibrium state of the charge-transfer state. In this process point C is passed, the point where the curve of the ground state intersects the curve of the excited state. This is the first possible way in which non-radiative losses may take place. Such losses are temperature-independent. This process is identical with what takes place in the Dexter-Klick-Russell model (fig. 15). If the process does not take place and the system arrives at D, then the following may happen.

At low temperature the only possible means of further relaxation is a transition from the charge-transfer state to the 4f levels $^5D_0$, $^5D_1$, $^5D_2$ etc., so that the system arrives at E (luminescence from point D has never yet been observed). Luminescence then takes place from the $^5D_0$ level (E).

At higher temperature the system may choose another route. It may reach C thermally via D and then return to A non-radiatively (Mott-Seitz model). This is the second possible form for a radiationless process. At a sufficiently high temperature this process ($D \rightarrow C \rightarrow A$) will become more probable than the process that gives rise to luminescence ($D \rightarrow E \rightarrow A$). This is in fact observed (I/I_0 in *fig. 20*). Upon excitation in the charge-transfer absorption band, the Eu^{3+} emission is quenched at a particular temperature.

Again at higher temperature (though lower than the quenching temperature) there will always be a small fraction of the excited centres that choose the route that leads to E. The decay time of the luminescence from this level has been measured as a function of temperature (τ/τ_0 in fig. 20). It is equal to the average time which the ion remains at the luminescent level. As soon as non-radiative processes become operative, the lifetime of the level and hence the decay time of the luminescence will become shorter. This only occurs at much higher temperatures than that of the quenching of the luminescence (fig. 20), indicating that relaxation can also take place along the path $E \rightarrow D \rightarrow C \rightarrow A$ (fig. 19). This process is the third possible mode for non-

[20] G. Blasse, A. Bril and J. A. de Poorter, J. chem. Phys. **53**, 4450, 1970 (No. 12).

radiative losses. It calls for an activation energy equal to the difference between the levels E and D, i.e. the energy needed for a return to the charge-transfer state.

Which of the three types of non-radiative losses occur depends both on the temperature and on the energy level of the charge-transfer state. The occurrence of the two temperature-dependent losses can be prevented by a sufficiently low temperature (fig. 20). If the efficiency of the luminescence is still found to be lower

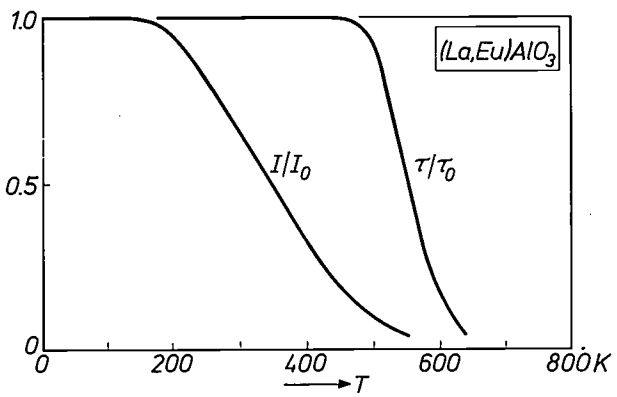


Fig. 20. Relative decay time τ/τ_0 and relative intensity I/I_0 of the emission of the Eu^{3+} ion in $LaAlO_3$ as a function of temperature. The excitation takes place in the charge-transfer level. The quantities τ_0 and I_0 represent the values of τ and I respectively, extrapolated to $T = 0$ K. The temperature quenching shown by the I/I_0 curve corresponds to the process $D \rightarrow C \rightarrow A$ in fig. 19; the drop shown by the τ/τ_0 curve is due to the process $E \rightarrow D \rightarrow C \rightarrow A$ in fig. 19. This drop therefore occurs at higher temperature than that of I/I_0.

than 100% in spite of a very low temperature, we must attribute this to the temperature-independent losses (Dexter-Klick-Russell model). The condition then is that the energy level of the charge-transfer state must be so low that C lies below B (as drawn in fig. 19). Such a low energy for the charge-transfer state is only rarely encountered. Examples are Eu^{3+} in $LaAlO_3$ and the niobate phosphors, where the charge-transfer absorption band is lower than 37 000 cm^{-1} (fig. 17).

In most cases, however, the charge-transfer state lies at a higher energy, for example at about 40 000 cm^{-1}, and then C in fig. 19 will not be below B but above it, as indicated schematically in fig. 18. The temperature-independent losses are then impossible.

If the charge-transfer state has a higher energy there can be a second consequence. It should be remembered that if the curve BDC (fig. 19) shifts upwards, the curve through E remains in its place. For by manipulating the host lattice we can influence the situation of the charge-transfer level but not that of the 4f levels. This implies that if the charge-transfer state is sufficiently high, non-radiative losses from E are impossible at the temperature commonly used for the measurements (T about 1000 K). This is because the energy difference E-D has become too great to be overcome thermally, so that the route $E \rightarrow D \rightarrow C \rightarrow A$ is blocked. We are thus left with only one non-radiative process, the one involving thermal excitation from $D \rightarrow C$ (Mott-Seitz model). This distance, too, will become greater as the energy level of the charge-transfer state is higher, so that the quenching temperature of the emission increases when the charge-transfer state shifts towards higher energies. This has been observed in the case of Eu^{3+} (see Table VIII) and also for the octahedral niobate group (fig. 17).

Summarizing it can therefore be said that there is no essential difference between the processes associated with the Eu^{3+} and the niobate centres. The situation with the Eu^{3+} ion is only somewhat more complicated owing to the occurrence of the process $D \rightarrow E$, whereas with the niobate group the luminescence occurs directly from D.

In this part of the article we have dealt with the least understood aspects of the luminescence process. It need hardly be said that our considerations do not have the character of a theory. They may rather be described as a set of empirical rules that throw light on the question of which ions or ionic groups will in principle give emission and which host lattices can be used with most chance of success.

It is in a sense surprising that such a simple qualitative model as described here (figs. 13, 14 and 15) is capable of showing what may probably be the physical background of our empirical relations.

Summary II. This article considers the quantum efficiency of phosphors showing characteristic luminescence, for the case where the excitation energy is absorbed directly by the luminescent centre. The starting point of these considerations is the configurational-coordinate diagram (Mott-Seitz model, extended by the Dexter-Klick-Russell model). An empirical method is described by means of which the change of distance Δr between the central cation and the neighbouring anions, brought about by excitation of the centre, can be estimated from calculated values of electron-orbit radii. It is shown that for the occurrence of luminescence the absolute value of Δr calculated in this way must be smaller than 0.3 Å. The effect of the host lattice is estimated from the radius and charge of the cations of the host lattice. The treatment makes it possible to obtain some idea of the efficiency of the luminescence, and how it is influenced by the chemical composition of the luminescent centre and by the host lattice. Finally a relation is dealt with between the quantum efficiency and the situation of the charge-transfer state of a number of centres (e.g. Eu^{3+}, NbO_6). This relation can also be explained with the aid of the configurational-coordinate model.

III. Energy transfer and efficiency

Introduction

The radiation emitted by a phosphor originates from a centre (activator) which is incorporated in some way in a crystal lattice (host lattice). In the preceding parts of this article we have looked at the electron transitions inside such a centre and at the quantum efficiency when the activator is excited directly, that is to say where the activator itself absorbs the excitation energy. In this part we shall discuss the case where the excitation energy is not absorbed in the activator itself but in another centre, which then transfers the energy to the activator.

An extensively studied example of such energy transfer is the phosphor $(Ca,Sb^{3+},Mn^{2+})_5(PO_4)_3(F,Cl)$, familiar for its use in tubular fluorescent lamps. When the phosphor is excited by radiation with a wavelength of 254 nm from the mercury-discharge spectrum, this radiation is absorbed by the Sb^{3+} ion, and not by the host lattice or by the Mn^{2+} ion. However, the luminescence consists of emission both from Sb^{3+} (blue) and from Mn^{2+} (yellow). This is due to the fact that the Sb^{3+} ions transfer part of the absorbed energy to the Mn^{2+} ions. If the relative concentrations of the two types of ions are suitably chosen, it is even possible to effect a complete energy transfer.

As already mentioned in part I, the centre (ion or group of ions) which absorbs the radiation is called the *sensitizer*, and the centre to which the energy is transferred is the *activator*. We also pointed out that there is in fact no fundamental difference between these two kinds of centre. This is evident for example, from the ability of the Sb^{3+} ion to emit radiation itself.

Fig. 21 gives a schematic picture of a crystal lattice containing sensitizer ions S and activator ions A. We shall use this figure to illustrate the processes that can take place in such a crystal.

If a centre S has absorbed a quantum of the exciting radiation, four things can happen:

1) S luminesces itself (thus acting as an activator). The probability of this process will be called P_S^r.

2) S returns non-radiatively to the ground state, while dissipating heat to the lattice. The probability of this process will be called P_S^{nr}. Unless otherwise stated, we shall disregard this process.

3) S transfers its excitation energy to A. The probability of this energy transfer will be called P_{SA}. This process can be followed by emission from A, but also possibly by the radiationless return to the ground state.

4) S transfers its energy to another centre S. The probability of this process will be called P_{SS}.

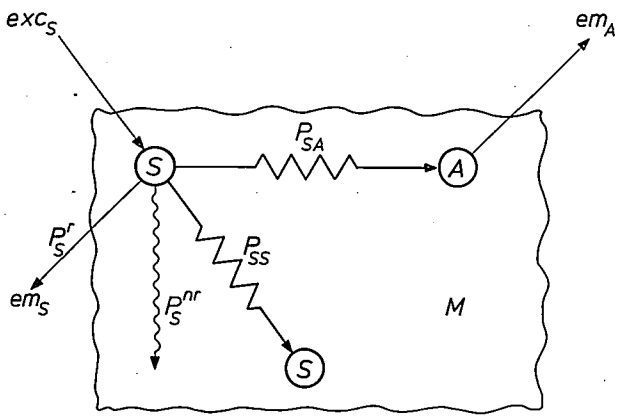


Fig. 21. Host lattice M with activator A and sensitizers S. The exciting radiation exc_S is absorbed by one of the centres S. This process is followed by one or more of the following processes: emission from S (em_S, probability P_S^r), non-radiative loss from S by heat dissipation (probability P_S^{nr}), transfer of energy to another centre of the type S (P_{SS}) and finally transfer of energy to a centre of the type A (P_{SA}). In the latter case A can emit radiation (em_A).

There are various methods that can be used to demonstrate the occurrence of energy transfer. One can, for example, measure the excitation spectrum of the emission from A. This is done by measuring the quantum yield of the emission from A (identified by its wavelength region) as a function of the wavelength of the incident radiation. A band in the excitation spectrum corresponds, of course, to an absorption band. If the excitation spectrum of the A emission shows the excitation bands of S in addition to those of A, this indicates energy transfer from S to A, since the excitation energy is absorbed by S and emitted by A. This is illustrated in *fig. 22* for energy transfer by the Ce^{3+} ion in $(Y,Ce,Tb)Al_3B_4O_{12}$. The excitation spectrum of the Tb^{3+} emission contains not only a band corresponding to excitation of the Tb^{3+} ion itself but also bands that correspond to excitation of the Ce^{3+} ion.

Another method of demonstrating energy transfer is to measure the decay time of the luminescence from S as a function of the concentration of A. If S is situated in the host lattice in an isolated position, the average lifetime τ_S of the excited state of S (i.e. the decay time of the luminescence) is equal to the reciprocal of P_S^r. If we now add A ions we make an extra process possible in which S can lose its excitation energy. As a result τ_S will become shorter and so too will the decay time of the luminescence from S. By measuring τ_S as a function of the concentration of A we can thus obtain information about P_{SA}.

The quantum efficiency q of the emission from A is defined in the case of excitation in S as the ratio of the number of quanta emitted by A to the number absorbed by S. If we want a high q we must ensure that $P_{SA} \gg P_S^r$. Now of course P_{SA} is a function of the distance r_{SA} between S and A. At low A concentrations, that is to say large r_{SA}, it is often difficult to make P_{SA} sufficiently large. As will later be shown, it is essential in many cases to keep the A concentration low. One can then still cause the energy of S ions to be transferred to A ions by increasing the S concentration. The energy then goes through the lattice from one S ion to the other (at least where $P_{SS} \gg P_S^r$) until an A ion is reached.

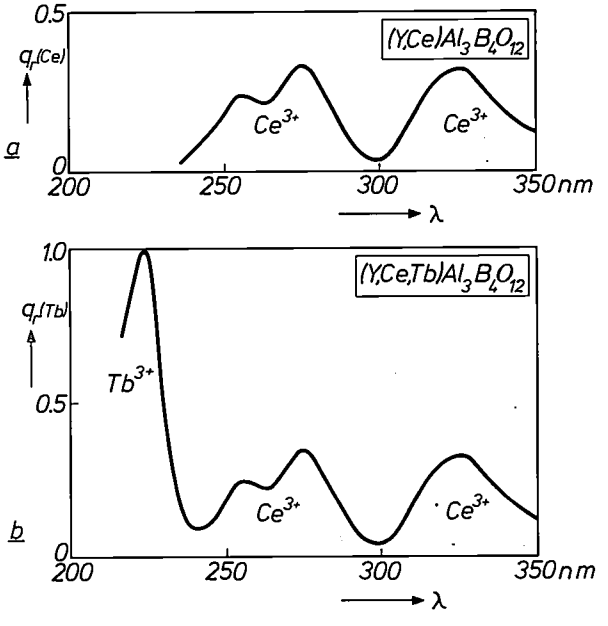


Fig. 22. *a*) Excitation spectrum of the Ce^{3+} emission of $(Y,Ce)Al_3B_4O_{12}$. The relative quantum yield q_r(Ce)of the Ce^{3+} emission is plotted as a function of the wavelength λ of the incident radiation. The excitation bands correspond to Ce^{3+} absorption bands. *b*) Excitation spectrum of the Tb^{3+} emission of $(Y,Ce,Tb)Al_3B_4O_{12}$. The relative quantum yield q_r(Tb) of the Tb^{3+} emission is plotted. This spectrum shows the same bands as the excitation spectrum of the Ce^{3+} emission, with in addition a band which is characteristic of Tb^{3+} itself (at about 225 nm). The latter corresponds to direct excitation of the Tb^{3+} centre; the bands first mentioned correspond to excitation of the Ce^{3+} centre followed by energy transfer from Ce^{3+} to Tb^{3+}.

Up to now it has been assumed that the symbols S and A represent ions or ionic groups incorporated in a non-absorbing host lattice. In many cases, however, S is an ion or ion group of the host lattice itself. In $(Y,Eu)VO_4$, for example, radiation with a wavelength of 254 nm is absorbed by the vanadate group. The emission, however, takes place in the Eu^{3+} ion, and a transfer of energy takes place from the vanadate group to the Eu^{3+} ion. In $Eu^{3+}(Al,Cr)_3B_4O_{12}$, the Eu^{3+} ion absorbs 254 nm radiation. Emission takes place in the Cr^{3+} ion. The latter two phosphors illustrate the fact that one and the same ion — here the Eu^{3+} ion — can play the role of either S or A, depending on the nature of the host lattice.

Concentration quenching

In order to obtain a high emission efficiency it would seem obvious to make the activator concentration as high as possible. In many cases, however, it is found that the emission efficiency decreases if the activator concentration exceeds a specific value known as the critical concentration. An example is to be seen in *fig.23*. This effect, called concentration quenching, may be explained in a number of cases as follows. If the concentration of the activator becomes so high that the probability of energy transfer exceeds that for emission, the excitation energy repeatedly goes from the one activator ion to the other. Now the host lattice is not perfect: it contains all kinds of sites where the excitation energy may, in some obscure way, be lost, such as at the surface, at dislocations, impurities, etc. In traversing the lattice the excitation energy will sooner or later encounter such a site where, dissipated as heat, it makes no contribution to the luminescence. The efficiency then decreases, in spite of the increase of the activator concentration.

In a similar way, concentration quenching for S centres can also take place. The value of the critical concentration of S centres provides information about P_{SS}: if the critical concentration is high, then P_{SS} is low and *vice versa*.

In the following section we shall take a closer look at the theory of the energy-transfer effect, and we shall conclude this article with the application of this theory to a variety of examples.

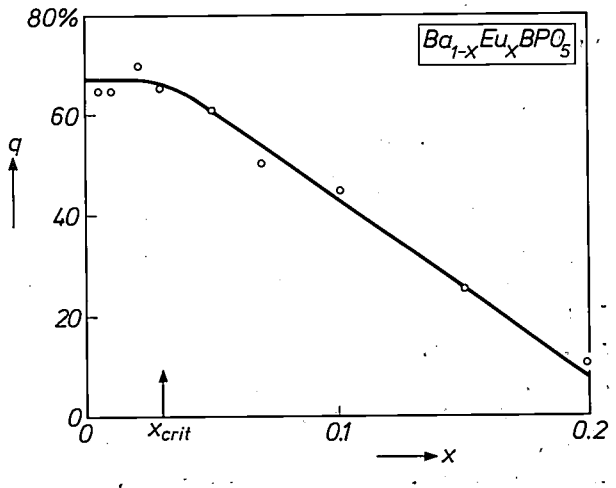


Fig. 23. Concentration quenching of the Eu^{2+} emission of $Ba_{1-x}Eu_xBPO_5$. For Eu^{2+} concentrations x which are greater than the critical concentration x_{crit} the absolute quantum efficiency q decreases with increasing x. (q is the ratio of the number of emitted quanta to the number of absorbed excitation quanta.)

Theory of the energy transfer

We shall consider here only those forms of energy transfer that involve no displacement of electric-charge carriers. We shall also disregard energy transfer by radiation (S radiates its energy and this is then absorbed by A). This case is seldom of importance in the phosphors of interest to us. The process most frequently observed is the non-radiative transfer of energy. The underlying theory was given by Th. Förster [21] and later worked out in more detail by D. L. Dexter [22].

In *fig. 24* we show schematically an energy-level diagram for an S and an A centre. We raise S from the ground state *1* to the excited state *2*, and we want the energy to be transferred from S (state *2*) to A. In other words, we want S to return from state *2* to state *1*, while A at the same time moves from state *1* to a higher energy level. The theory shows that this is only possible if one of the levels of A lies at the same height as level *2* of S (resonance). In the theory mentioned, such a transfer can take place in two essentially different ways.

In the first place the transfer can be brought about by the Coulomb interaction between all charged particles of S and A. If S and A are so far apart that their charge clouds do not overlap, this form of energy transfer is the only one possible.

If the charge clouds of S and A do overlap, however, another transfer process is possible by exchange interaction between the electrons of S and A. The essential difference between the previous process and this one is that here electrons are exchanged between S and A, whereas in the Coulomb interaction process the electrons remain with their respective ions or ionic groups.

A mathematical treatment of these mechanisms is outside the scope of this article [21] [22]. We will, however, discuss the result, because it gives some idea of what takes place in the process of energy transfer. We begin with energy transfer by Coulomb interaction, and consider the case where the dipole-dipole interaction is much stronger than that of multipoles of higher order, so that we can disregard the contributions of the latter. In that case the probability $P_{SA}(dd)$ of energy transfer from S to A is given by the expression:

$$P_{SA}(dd) = \frac{3\hbar^4 c^4}{4\pi K^2} \cdot \frac{Q_A}{\tau_S r_{SA}^6} \int f_S f_A \frac{dE}{E^4}. \quad (1)$$

In this expression:
$\hbar = h/2\pi$, where h is Planck's constant,
c is the velocity of light,
K the dielectric constant of the host lattice,
τ_S the decay time of the emission from S in the absence of A (this quantity is equal to the reciprocal of P_S^r).

The integral represents the overlapping of the normalized emission band $f_S(E)$ of S and the normalized absorption band $f_A(E)$ of A, both given as functions of photon energy E. Q_A is the integrated absorption of A.

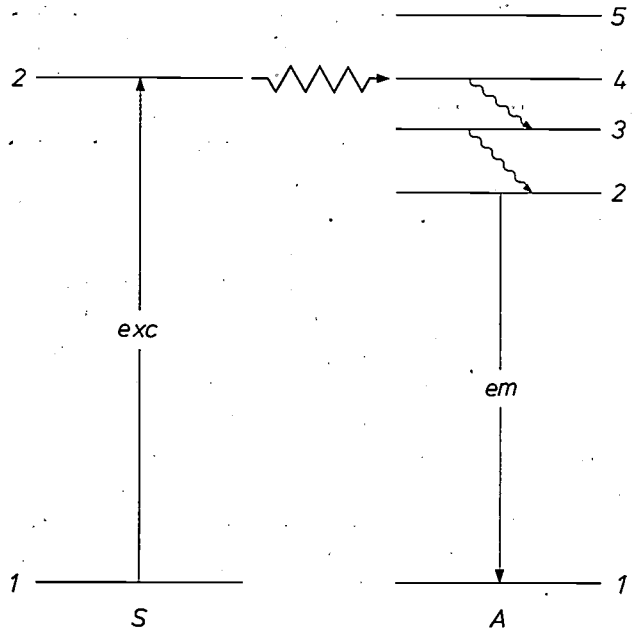


Fig. 24. Energy transfer from a sensitizer to an activator centre (S and A respectively). Due to the excitation ($\rightarrow exc$), S is raised from the ground state *1* to the excited state *2*. On transfer of the excitation energy to A, S returns to *1* and A goes to the energy level *4*. In the case illustrated, the return to the ground state takes place via two non-radiative transitions ($4 \rightarrow 3$ and $3 \rightarrow 2$) followed by a transition $2 \rightarrow 1$ where radiation is emitted. If A has no energy level very close to the level *2* of S, no energy transfer is possible.

It is possible to determine experimentally whether the levels involved are in resonance with one another by comparing the frequency of the emission band of S (transition $2 \rightarrow 1$ in S) with that of the relevant absorption band of A ($1 \rightarrow 4$ in A). The more these bands overlap, the better the resonance condition is fulfilled. Greater overlapping corresponds to a greater value of the integral in expression (1) and hence implies a higher energy-transfer probability. If the bands do not overlap, energy transfer is not possible.

Fig. 25 shows the overlapping of emission and absorption spectra. In fig. 24, the relevant levels are shown for simplicity as discrete lines; in practice, however, the spectra consist of bands possessing a certain width.

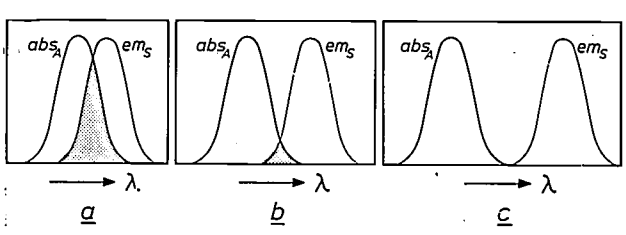


Fig. 25. The probability of energy transfer depends on the spectral overlapping of the emission band em_S of S and the absorption band abs_A of A. *a*) Considerable overlapping, *b*) moderate overlapping, *c*) no overlapping; in this case the transfer of energy from S to A is impossible.

Let us now return to equation (1) in order to examine more closely the part $Q_A/\tau_S r_{SA}^6$. We see that the transfer probability for electric dipole-dipole interaction depends on the absorption area of the relevant transition in A. The transfer probability is therefore greatest if the relevant transition is an allowed electric-dipole transition in A. The transfer probability also depends to a great extent on the distance between S and A.

If $Q_A = 0$ (forbidden electric-dipole transition in A) there can still be a certain transfer probability by interaction due to terms of higher order. The mathematical expressions for these are not fundamentally different from equation (1). For electric-dipole-quadrupole interaction the distance term now appears as the eighth power and Q_A represents the absorption area resulting from a quadrupole transition, etc. The resultant transfer probabilities are in some cases, surprisingly, scarcely less than those for electric dipole-dipole interaction [22].

We now want to know over what distances energy can be transferred by Coulomb interaction. For this purpose we fill in the constants in equation (1), taking for Q_A the value for an allowed electric-dipole transition and for the overlapping integral a value which corresponds to a fairly high overlap. We then find:

$$P_{SA}(dd) = (27/r_{SA})^6 \tau_S^{-1}. \quad . \quad . \quad . \quad . \quad (2)$$

In this equation the distance r_{SA} must be expressed in Ångström units. We must realize that τ_S^{-1} (where τ_S is the decay time of S in the absence of A) is equal to P_S^r, the probability of a radiative transition in S. When A centres also are present the probability of emission from S and the probability of transfer from S to A are therefore equal to one another if $r_{SA} = 27$ Å, an appreciable distance. This distance, called the critical distance for energy transfer, is denoted by the symbol r_{SA}^0. For $r_{SA} > r_{SA}^0$ the emission is almost exclusively in S. For $r_{SA} < r_{SA}^0$ energy transfer dominates, and is more important the smaller the value of r_{SA}.

We shall now discuss the equation for the probability of transfer by exchange interaction:

$$P_{SA}(ex) = \frac{2\pi}{\hbar} Z^2 \int f_S f_A dE. \quad . \quad . \quad . \quad (3)$$

This equation, of course, again contains the overlap integral. The quantity Z cannot be obtained directly from optical experiments; it is proportional to the exchange integral

$$\int \left\{\psi_A^e(r_1)\psi_S^0(r_2)\right\}^* \frac{e^2}{(r_1 - r_2)} \left\{\psi_A^0(r_2)\psi_S^e(r_1)\right\} dr_1 dr_2. \quad . \quad . \quad . \quad (4)$$

This expression contains the position coordinates r_1 and r_2 of the two electrons, and also the quantum-mechanical wave functions ψ of the two centres.

The product between the first set of curly brackets gives the final state: S is then in the ground state (ψ_S^0), A in the excited state (ψ_A^e). The product between the second pair of curly brackets gives the initial state: S is in the excited state (ψ_S^e), A in the ground state (ψ_A^0). The complex character of the exchange integral is a consequence of the fact that electron *1* is in the initial state at S but in the final state at A. The converse applies to electron *2* (exchange). If the charge clouds of S and A do not overlap, Z is zero and so too is $P_{SA}(ex)$. If there is some overlapping, however, then electrons *1* and *2* repel each other in the region of overlap between S and A. Because of this, exchange can take place. Since the density of charge clouds decreases exponentially with the distance of the electron to the nucleus, the dependence of Z upon distance will also be exponential and so too will that of $P_{SA}(ex)$. Significant overlapping of the charge clouds of two cations in a crystal lattice is found only between cations that are nearest neighbours (separation 3 to 4 Å). Exchange interaction is therefore limited to neighbouring cations in the lattice. The critical distance for this transfer will never be much greater than 4 Å.

We note that equation (3) does not comprise the optical properties of S and A (apart from the overlap integral). Exchange transfer, then, unlike transfer by Coulomb interaction, is not dependent on the oscillator strength or transition probability of the relevant transitions, and may even take place to a level from which a return to the ground state is strictly forbidden.

The equations for the probability of two modes of energy transfer may be summarized as follows. For transfer by Coulomb interaction we write:

$$P_{SA} = g_{SA} E_{SA}, \quad . \quad . \quad . \quad . \quad . \quad (5)$$

and for transfer by exchange interaction:

$$P_{SA} = f_{SA} E_{SA}. \quad . \quad . \quad . \quad . \quad . \quad (6)$$

In these two analogous equations, E_{SA} represents the overlap integral of the emission band of S and the absorption band of A. The quantity g_{SA} comprises the optical strengths of the relevant transitions and a distance-dependence of the type of r_{SA}^{-n}. The quantity f_{SA} is proportional to the overlapping of the charge clouds of S and A and therefore comprises an exponential distance-dependence.

We shall now apply this theory to a number of specific cases and show how strongly the luminescence properties may vary as a result of differences in the transfer probabilities.

[21] Th. Förster, Ann. Physik (6) **2**, 55, 1948.
[22] D. L. Dexter, J. chem. Phys. **21**, 836, 1953.

Specific examples

In *Table X* we have listed a number of phosphors and subdivided them as follows. The transfer from a centre S to a centre A takes place either over distances greater than the distance between the nearest cation neighbours ($SA+$), or over distances equal to or smaller than the distance between nearest neighbours ($SA-$). We make the same division for the transfer from one S centre to another. The phosphors then fall into four groups ($SS+$ and $SA+$, $SS+$ and $SA-$, $SS-$ and $SA+$, and $SS-$ and $SA-$). The probability of a high emission yield is of course greatest with a transfer of the type $SA+$ or of the type $SS+$, and certainly if both of them are possible at the same time.

Phosphors with SS+ and SA+

An example of a phosphor with $SS+$ and $SA+$ is (Y,Bi)VO_4. The vanadate group acts here as S, and the Bi^{3+} ion as A. Excitation of the VO_4 group by 254 nm radiation results in a yellow-green emission in the Bi^{3+} centre. We now look first at the luminescence properties of the VO_4 group. We find that YVO_4 itself shows no or scarcely any luminescence at room temperature. The explanation must be sought, as we have been able to demonstrate, in concentration quenching [23].

This may be deduced from the surprising fact that the VO_4 group at 20 °C — and even at much higher temperatures — does show emission if the concentration of the VO_4 groups is sufficiently reduced. This can be ascertained owing to the fact that YPO_4 is isomorphous with YVO_4 and readily forms mixed crystals with it. The phosphate group does not absorb the 254 nm radiation. It is found that in the mixed crystal series $YP_{1-x}V_xO_4$ for $0 < x \leqslant 0.2$ the efficiency of the blue vanadate luminescence under excitation with 254 nm radiation is high and constant. For $x > 0.2$, however, it decreases. The concentration of vanadate groups then increases such that the excitation energy travels through the lattice from one VO_4 group to another. In this way, with increasing vanadium content x an ever larger part of the excitation energy is trapped in lattice imperfections and so lost for the emission.

From our experiments we have derived a value for the critical distance r_{SS}^0 in the case of the transfer $VO_4 \rightarrow VO_4$; it amounts to about 8 Å [24]. *Fig. 26* shows the emission and excitation (absorption) spectra of the vanadate group. It can be seen that both bands have, in fact, some overlap. The absorption spectrum of the Bi^{3+} centre in YVO_4 is also shown; it overlaps even more markedly the VO_4 emission. We may therefore assume that P_{SA} will certainly be just as great and probably even greater than P_{SS}. This now explains the high efficiency of the Bi^{3+} emission from the phosphor (Y,Bi)VO_4 when excited in the VO_4 group, even when the Bi^{3+} concentration is low. As already remarked, in cases where $SS+$ and $SA+$ can occur together the efficiency of the A emission upon excitation in S will always be high, unless the S and A concentrations are both very low. For in that case not all excitation energy will be transferred from S to A, and in addition to emission in A we also observe emission in S. The composition $Y_{0.99}Bi_{0.01}P_{0.99}V_{0.01}O_4$ upon excitation in the vanadate group does in fact show both the blue vanadate emission and the yellow-green of the Bi^{3+} centre.

Table X, Some phosphors classified according to the nature of the atoms between which energy is transferred (S and A or S and S, denoted respectively by SA and SS) and also by the distance over which the energy transfer takes place. A plus sign following the symbols SS and SA means that the transfer takes place over distances larger than the distance between nearest-cation neighbours in the lattice; a minus sign means that the transfer takes place over distances equal to or less than the distance between nearest-cation neighbours.

	$SA+$	$SA-$
$SS+$	(Y,Bi,Eu)$Al_3B_4O_{12}$ $S = Bi^{3+}$; $A = Eu^{3+}$	(Y,Eu)VO_4 $S = VO_4$; $A = Eu^{3+}$
	(Y,Bi)VO_4 $S = VO_4$; $A = Bi^{3+}$	(Ce,Tb)BO_3 $S = Ce^{3+}$; $A = Tb^{3+}$
$SS-$	(Y,Tb)TaO_4 $S = TaO_4$; $A = Tb^{3+}$	(Y,Eu)NbO_4 $S = NbO_4$; $A = Eu^{3+}$
	Eu(Al,Cr)$_3B_4O_{12}$ $S = Eu^{3+}$; $A = Cr^{3+}$	(Ce,Tb)F_3 $S = Ce^{3+}$; $A = Tb^{3+}$

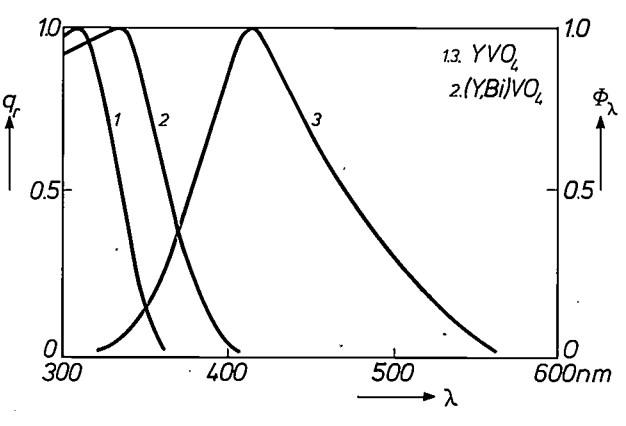


Fig. 26. Overlapping of emission spectra (spectral radiant power Φ_λ plotted as a function of the emitted wavelength λ and absorption spectra. Curve *3*: emission spectrum of the VO_4 emission from YVO_4. Curve *1*: excitation spectrum (q_r vs. λ; cf. fig. 22) of the same phosphor. We take this spectrum as a representation of the absorption spectrum; it may be concluded that there is some overlapping with curve *3*. Curve *2*: excitation spectrum of the Bi^{3+} ions of (Y,Bi)VO_4. This curve overlaps curve *3* even more.

An even more extreme example in this class of phosphors is (Y,Bi,Eu)$Al_3B_4O_{12}$. The Bi^{3+} ion here plays the role of *S* and the Eu^{3+} ion the role of *A*. Owing to the high value of the overlap integral (*fig. 27*) the $Bi^{3+} \rightarrow Bi^{3+}$ transfer can take place over a very considerable distance. In phosphors with the composition $Y_{1-x}Bi_xAl_3B_4O_{12}$ concentration quenching of the Bi^{3+} emission already takes place when *x* is about 0.005. The critical distance for the $Bi^{3+} \rightarrow Bi^{3+}$ transfer is no less than about 35 Å in this case [24]. The Bi^{3+} emission band not only substantially overlaps the Bi^{3+} absorption band but also the charge-transfer absorption band of the Eu^{3+} ion. The transition from the ground state to the charge-transfer state, an allowed transition, also has high absorptivity. For the $Bi^{3+} \rightarrow Eu^{3+}$ transition as well, the critical distance is therefore very considerable. In phosphors consisting of $YAl_3B_4O_{12}$ with low Bi^{3+} and Eu^{3+} concentration, the Eu^{3+} emission therefore shows a high efficiency on excitation in the Bi^{3+} ion.

To obtain phosphors with a high emission efficiency in *A* it is not necessary for both the *SS* and *SA* transfers to take place over a great distance. It is sufficient if one of them does this, as will appear from the following.

Phosphors with SS+ and SA−

A familiar phosphor from this class is (Y,Eu)VO_4. Here the VO_4 group is *S* and has already been discussed above. The role of *A* is played by the Eu^{3+} ion. As we have seen above, *SS+* applies in this case. We now look at the energy transfer $VO_4 \rightarrow Eu^{3+}$. The vanadate emission lies in the blue part of the spectrum. As argued in part I of this series, the Eu^{3+} ion in this spectral region has no allowed absorption transitions. In the absorption spectrum of Eu^{3+} in this region we do find, however, a number of very weak, sharp absorption peaks, which correspond to the strictly forbidden 4f-4f transitions within this ion. This implies that both Q_A and the overlap integral in equation (1) have low values, so that the probability of transfer from VO_4 to Eu^{3+} by Coulomb interaction will be small. This transfer, then, can only take place over a short distance, in other words we have here a case of *SA−*. An estimate of r_{SA}^0 gives roughly 4 Å [24]. For this distance P_{SA} is roughly 10^4 s^{-1}.

The probability of energy transfer from VO_4 to Eu^{3+} by exchange interaction is much greater. An estimate gives a value of about 10^7 s^{-1}. The fact that the transfer from groups such as vanadate, niobate and tungstate to rare-earth ions takes place by exchange interaction is evident from the angular dependence of the transfer. We have investigated this in a large number of cases, and we shall illustrate it here for the phosphors $(Y,Eu)_2WO_6$ and $(Gd,Eu)_2WO_6$ [25]. Upon excitation

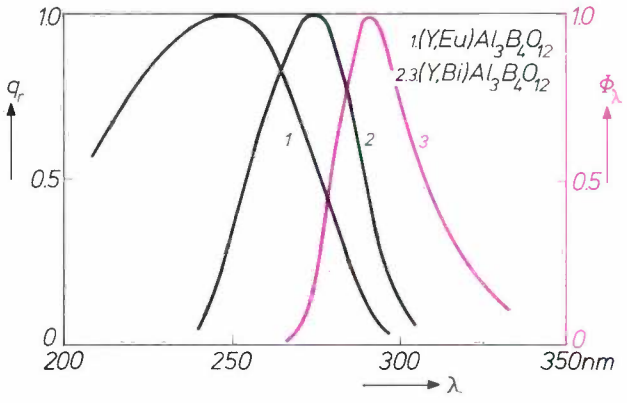


Fig. 27. The spectral overlap of the emission and excitation spectra of (Y,Bi)$Al_3B_4O_{12}$ (curves *3* and *2*) and of the emission spectrum of (Y,Bi)$Al_3B_4O_{12}$, and the excitation spectrum of (Y,Eu)$Al_3B_4O_{12}$ (curves *3* and *1*).

in the tungstate group the quantum efficiency of the Eu^{3+} emission shows a high value for $(Y,Eu)_2WO_6$ and a low value for $(Gd,Eu)_2WO_6$. This difference in efficiency proves to be attributable to the difference in the probability of energy transfer from the tungstate group to the Eu^{3+} ion in both lattices. The difference between the tungstate-Eu^{3+} configuration in Y_2WO_6 and Gd_2WO_6 is that the angle W-O-Eu^{3+} in Y_2WO_6 is about 180° and in Gd_2WO_6 about 90°. Angular dependence of the transfer probability does not follow from the theory for transfer by Coulomb interaction, but it does from the theory for transfer by exchange interaction. In the latter, as described, the overlapping of the charge clouds of the W^{6+} ion and of the Eu^{3+} ion plays an important part. It is known that this overlapping is much greater for the linear than for the rectangular configuration. This explains the higher probability of transfer from the tungstate group of the Eu^{3+} ion in $(Y,Eu)_2WO_6$. A transfer process of this type is of course limited to short distances.

Thus (Y,Eu)VO_4 is an example of a phosphor with *SS+* and *SA−*. The quantum efficiency of the Eu^{3+} emission for excitation in the vanadate group is high. As in the case of the phosphor (Y,Bi)VO_4, we can reduce P_{SS} by lowering the vanadate concentration. We find that in the system $(Y,Eu)V_{1-x}P_xO_4$ the efficiency of the Eu^{3+} emission upon excitation of the vanadate group decreases if *x* becomes greater than 0.8. In that case P_{SS} has become so low that it is no longer possible for all excitation quanta to reach the Eu^{3+} ion. We therefore find in fact that for $x > 0.8$ the blue vanadate emission occurs in addition to the red Eu^{3+} emission.

Another example of a phosphor with *SS+* and *SA−*

[23] G. Blasse, Philips Res. Repts. **23**, 344, 1968.
[24] G. Blasse, Philips Res. Repts. **24**, 131, 1969.
[25] G. Blasse and A. Bril, J. chem. Phys. **45**, 2350, 1966.

is $(Ce,Tb)BO_3$. The quantum efficiency of the Tb^{3+} emission for excitation in the Ce^{3+} ion is high. Physically this example is completely analogous to $(Y,Eu)VO_4$, even though the chemical differences are considerable.

The host lattice $CeBO_3$ shows rather a low efficiency. This too is caused by concentration quenching of the Ce^{3+} emission. If the Ce^{3+} concentration is lowered by replacing the Ce^{3+} ion in $CeBO_3$ by the La^{3+} ion, which does not absorb UV radiation, then efficient Ce^{3+} luminescence (maximum at ~390 nm) occurs when the Ce^{3+} concentration is sufficiently low. This situation is therefore entirely comparable with that in the case of $Y(P,V)O_4$. In $CeBO_3$ the excitation energy moves readily through the lattice. If no activator is present, the excitation energy is finally lost at an imperfect ion in the lattice. If Tb^{3+} is added, however, the energy is transferred to the Tb^{3+} ion [26]. This ion gives a green luminescence. For the same reasons as mentioned above in connection with the transfer $VO_4 \rightarrow Eu^{3+}$, the transfer $Ce^{3+} \rightarrow Tb^{3+}$ takes place by exchange interaction.

The third manner of obtaining efficient emission from *A* by excitation in *S* is to make a phosphor with *SS*— and *SA*+. This is discussed below.

Phosphors with SS— and SA+

A good example of such a phosphor is $(Y,Tb)TaO_4$. The tantalate group here plays the role of *S*. From the fact that the tantalate emission of $YTaO_4$ has a very high efficiency [27], and thus shows no concentration quenching, it follows that the critical distance for the *SS* transfer is small (*SS*—).

Fig. 28 shows the relevant emission and excitation spectra. The TaO_4 emission lies far in the UV region and can only be excited with radiation whose wavelength is shorter than 225 nm. The figure also indicates the main reason for the low probability of the transfer $TaO_4 \rightarrow TaO_4$: the overlap of the emission and excitation (or absorption) bands is very slight, implying a small spectral-overlap integral.

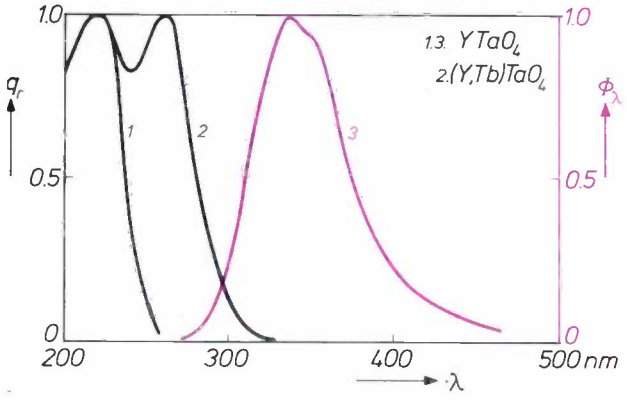


Fig. 28. Spectral overlap of the emission band of $YTaO_4$ and the excitation band of $(Y,Tb)TaO_4$ (*3* and *2*). *1* is the excitation band of $YTaO_4$. The spectral overlap of the emission and excitation bands of $YTaO_4$ is very slight.

In order to make efficient phosphors on the basis of this lattice we must try to create a case of *SA*+, that is to say we must ensure that the *SA* transfer takes place by Coulomb interaction. We are helped in this by the fact that the emission of $YTaO_4$ is of very short wavelength. There are many activators that show no or only forbidden absorption in the visible region but have allowed absorption bands in the UV region. An example is Tb^{3+}. In the visible part of the spectrum this ion shows only the very weak 4f-4f absorptions, but in the UV region we find the allowed 4f-5d absorption (see I). Transfer to the 4f levels over a large distance is impossible; it is, however, possible to the 5d level. Fig. 28 shows that the emission band of $YTaO_4$ does in fact overlap slightly the broad absorption and excitation bands of the Tb^{3+} ion in $YTaO_4$. Transfer by Coulomb interaction is therefore possible; the critical distance for this transfer is approximately 10 Å [27].

A similar situation is found for Eu^{3+} and Bi^{3+} in $YTaO_4$.

Another example of a phosphor with *SS*— and *SA*+ is $Eu(Al,Cr)_3B_4O_{12}$ [28]. The Eu^{3+} ion this time plays the role of *S* and the Cr^{3+} ion that of *A*. The $EuAl_3B_4O_{12}$ lattice shows efficient Eu^{3+} emission. The emission spectrum is given in *fig. 29*. Here the transfer from Eu^{3+} to Eu^{3+} ion is virtually impossible. The spectral overlapping of Eu^{3+} emission and absorption is very slight; moreover the distance between the Eu^{3+} ions is fairly large (about 6 Å). The substitution of Cr^{3+} for a small percentage of the Al^{3+} ions of $EuAl_3B_4O_{12}$ (e.g. 1 %) results in the almost complete disappearance of the Eu^{3+} emission, while now an efficient Cr^{3+} emission occurs. This lies in the red and the near infra-red part of the spectrum (fig. 29); both discrete lines and a broad band are found. We shall not here go any further into this remarkable emission spectrum. Fig. 29 also shows the absorption bands of the Cr^{3+} ion. The emission lines of the Eu^{3+} ion at about 600 nm are well covered by a Cr^{3+} absorption band. The critical distance for the transfer $Eu^{3+} \rightarrow Cr^{3+}$ is roughly 15 Å. The fact that this is not even higher, in spite of the good spectral overlapping, must be attributed to the relatively low absorption area of the Cr^{3+} absorption band (Q_A in equation 1). The energy transfer $Eu^{3+} \rightarrow Cr^{3+}$ also takes place by Coulomb interaction.

Phosphors with SS— and SA—

Finally we consider phosphors with *SS*— and *SA*—. This means that the quantum efficiency of the *A* emission for excitation of *S* will invariably be low, unless the *A* concentration can be raised to a very high value without the occurrence of concentration quenching of

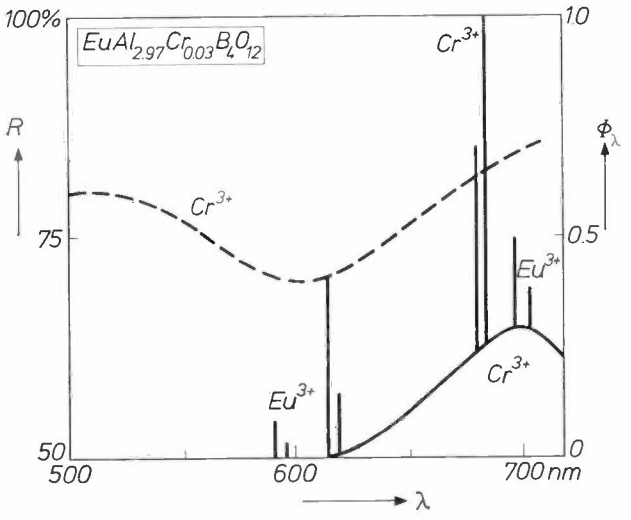


Fig. 29. Emission of $EuAl_{2.97}Cr_{0.03}B_4O_{12}$ (solid curve) for excitation in the Eu^{3+} ions. The emission consists of a broad band with a maximum at about 700 nm (Cr^{3+} emission) with, superimposed on it, the Cr^{3+} emission lines around 690 nm and the Eu^{3+} emission lines (around 600 and 700 nm). With this Cr^{3+} concentration the transfer $Eu^{3+} \rightarrow Cr^{3+}$ is not complete. The Cr^{3+} absorption band in the visible part of the spectrum appears as a valley in the reflectance spectrum (dashed curve). The Eu^{3+} emission around 600 nm is well overlapped by this Cr^{3+} absorption band.

the *A* emission, a condition which is difficult to fulfil.

Our first example in this section is (Y,Eu)NbO_4. The Eu^{3+} ion is *A* here, the niobate group is *S*. The emission and excitation spectra concerned are given in *fig. 30*. The compound $YNbO_4$ itself shows efficient luminescence, which is an indication that the *SS* (niobate-niobate) transfer does not take place over a large distance. The critical distance for this transfer is 4 Å, roughly equal to the shortest Nb-Nb distance in the crystal lattice [24]. This means, then, that the probability P_S^r of emission of an excited NbO_4 group is equal to that of the transfer P_{SS}. Fig. 30 shows the cause of the low probability of transfer, which is that the spectral overlapping of the niobate emission and absorption bands is very small. As regards the transfer $NbO_4 \rightarrow Eu^{3+}$ the same applies as stated above for the transfer $VO_4 \rightarrow Eu^{3+}$. This too takes place over a relatively short distance. In fact a phosphor with the composition $Y_{0.97}Eu_{0.03}NbO_4$ shows both red Eu^{3+} emission and blue NbO_4 emission if the excitation takes place in the niobate group.

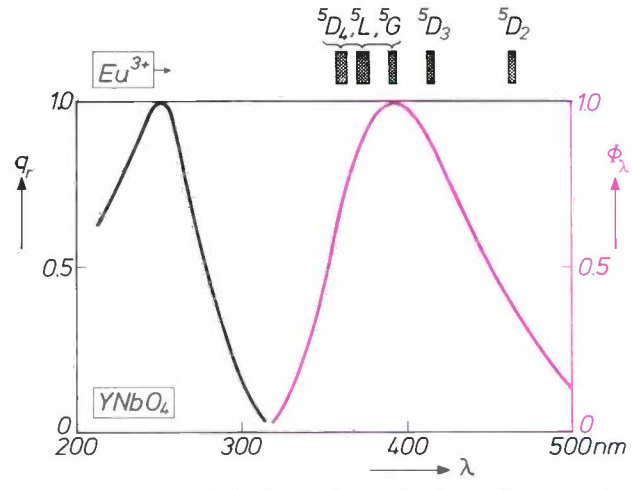


Fig. 30. Excitation (left) and emission band (right) of the niobate emission from $YNbO_4$. The spectral overlap is very slight. Above the figure the position of some absorption lines of the Eu^{3+} ion is shown. These correspond to transitions between the ground state and the indicated higher states. It can be seen that the absorption lines of the Eu^{3+} ion are well overlapped by the niobate emission.

The efficiency of the Eu^{3+} emission of (Y,Eu)NbO_4 can be raised by increasing the Eu^{3+} concentration. If this concentration is higher than 10%, then the efficiency is fairly high, in spite of the occurrence of some concentration quenching. The reason for this is that energy transfer over a large distance is not necessary with such an Eu concentration: nearly every excited NbO_4 group has an Eu^{3+} ion as a neighbour. Energy is readily transferred between these neighbours owing to the fact that the angle Nb-O-Eu^{3+} is roughly 135°.

Another example of a phosphor with *SS*— and *SA*— is (Ce,Tb)F_3. Unlike $CeBO_3$, the Ce^{3+} emission of CeF_3 exhibits no concentration quenching. On the contrary, CeF_3 shows UV emission with a high efficiency. The probability of *SS* transfer ($Ce^{3+} \rightarrow Ce^{3+}$) is therefore small. We have already seen above that the energy transfer $Ce^{3+} \rightarrow Tb^{3+}$ is also limited to a short distance. The low efficiency of the Tb^{3+} emission in (Ce,Tb)F_3 for excitation in the Ce^{3+} ions must therefore be attributed to the fact that both *SS*— and *SA*— are applicable. If the Tb^{3+} concentration is increased to 20%, an efficient phosphor is obtained even under these conditions.

Other cases

In the foregoing we have consistently assumed that, with only *S* present, the probability of emission from *S* far exceeds the probability of a non-radiative transition in *S* ($P_S^r \gg P_S^{nr}$). With *A* also present, the larger P_S^{nr}, the smaller will be the probability of energy transfer. The probability of energy transfer must be greater than the probabilities of emission and of non-radiative loss together ($P_{SA} > P_S^r + P_S^{nr}$) for significant energy transfer.

We have used this argument in order to explain, for example, the low efficiency of the Eu^{3+} emission in (Sc,Eu)NbO_4 for excitation in the niobate group. The factors that determine the transfer probabilities are identical with those in (Y,Eu)NbO_4. Nevertheless the

[26] G. Blasse and A. Bril, J. chem. Phys. **51**, 3252, 1969
[27] G. Blasse and A. Bril, J. Luminescence **3**, 109, 1970.
[28] G. Blasse and A. Bril, Phys. Stat. sol. **20**, 551, 1967.

efficiency of the Eu^{3+} emission is low compared with that in (Y,Eu)NbO_4. The reason for this is that quite substantial non-radiative losses occur in the niobate group in $ScNbO_4$. This appears from the fact that the efficiency of the luminescence of this compound is low (about 20%). A considerable part of the excitation energy is therefore immediately dissipated as heat after excitation.

Finally we should point out that not all cases of concentration quenching can be explained with the theory described above. The explanation of some cases of concentration quenching calls for an entirely different mechanism, based not on energy transfer but on another interaction between the centres [29].

Broadly speaking, our discussion in this third part of the article amounts to the following.

Energy transfer in phosphors with characteristic emission, and a large number of cases of concentration quenching of characteristic emission, can readily be explained with the Förster and Dexter theory. If we consider a phosphor where the excitation energy is absorbed by a centre S and emitted by a centre A, then the efficiency of the A emission can be understood and indeed often predicted by using the theory to estimate the probability of both the SS— and the SA— energy transfer. Efficient luminescence from A is only possible as a rule if at least one of the two transfer processes can take place over a distance which is large compared with the shortest distance between the relevant centres found in the lattice.

[29] For a short summary of the theories on concentration quenching of characteristic luminescence, see G. Blasse, J. Luminescence **1/2**, 766, 1970.

Résumé

If, to conclude this article, we review the present state of our knowledge about phosphors, we may distinguish three main themes.

1) The situation of the energy levels of ions in crystals is for the most part well known, and the influence of the host lattice on this situation can be explained in qualitative terms.

2) The energy transfer between two centres showing characteristic emission takes place as predicted by the theory, and this insight can successfully be used for explaining or predicting the efficiency of phosphors.

3) The non-radiative processes that may take place in a luminescent centre, and especially the influence of the host lattice on these processes, are not well known and constitute the major gap in our knowledge of phosphors showing characteristic emission. Nevertheless, even here it is possible to advance explanatory hypotheses and to make predictions on the grounds of empirical rules.

Summary III. This part of the article deals with the efficiency of the luminescence from phosphors showing characteristic emission, where the excitation energy is not absorbed by the emitting centre itself (the activator A) but by another centre (the sensitizer S). The efficiency is determined to a great extent by the probability of energy transfer from S to A. First the Förster and Dexter theories on energy transfer are discussed. The theories are then applied to a number of cases. Apart from transfer from S to A, it is found that transfer from S to S also plays a part. Phosphors in which energy transfer takes place can accordingly be subdivided into four groups, indicated by the combinations $SS+$ and $SA+$, $SS+$ and $SA-$, $SS-$ and $SA+$ and $SS-$ and $SA-$ (the symbol + refers to energy transfer over a distance larger than one atomic spacing, and the symbol − refers to transfer over one atomic spacing or less). Phosphors with $SS+$ and/or $SA+$ will have a high efficiency for A emission, even with a low A content.

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.

D. Blom & J. O. Voorman: Noise and dissipation of electronic gyrators.
Philips Res. Repts. **26**, 103-113, 1971 (No. 2). *E*

P. Branquart & J. Lewi: On the implementation of coercions in ALGOL 68.
Proc. Int. Computing Symp., Bonn 1970, Vol. II, pp. 322-345. *B*

J. C. Brice: An analysis of factors affecting dislocation densities in pulled crystals of gallium arsenide.
J. Crystal Growth **7**, 9-12, 1970 (No. 1). *M*

K. Carl & 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 (No. 10). *A*

P. J. Courtois & J. Georges: An evaluation of the stationary behavior of computations in multiprogramming computer systems.
Proc. Int. Computing Symp., Bonn 1970, Vol. I, pp. 98-115. *B*

P. Delsarte: Weights of p-ary Abelian codes.
Philips Res. Repts. **26**, 145-153, 1971 (No. 2). *B*

F. Desvignes: Pourquoi l'arséniure d'indium?
Acta Electronica **13**, 9-12, 1970 (No. 1). *L*

F. C. Eversteijn: Gas-phase decomposition of silane in a horizontal epitaxial reactor.
Philips Res. Repts. **26**, 134-144, 1971 (No. 2). *E*

R. C. French: Electronic arts of non-communication.
New Scientist **46**, 470-471, 1970 (4 June). *M*

A. Gowthorpe: Economical dual-polarity regulated power supplies.
Electronic Engng. **42**, No. 505, 33-35, March 1970. *M*

G. Groh: Holographie.
Röntgenstrahlen **22**, 23-27, 1970. *H*

F. N. Hooge & J. L. M. Gaal: Fluctuations with a $1/f$ spectrum in the conductance of ionic solutions and in the voltage of concentration cells.
Philips Res. Repts. **26**, 77-90, 1971 (No. 2). *E*

J. Hornstra: A program which, using a known part of the molecule, produces the parameters of all atoms.
Crystallographic Computing, Proc. 1969 Summer School, pp. 103-109; 1970. *E*

J. G. M. de Lau: Preparation of ceramic powders from sulfate solutions by spray drying and roasting.
Amer. Ceramic Soc. Bull. **49**, 572-574, 1970 (No. 6). *E*

J. Lewi & P. Branquart: Implementation of local name in ALGOL 68.
Proc. Int. Computing Symp., Bonn 1970, Vol. II, pp. 474-490. *B*

S. R. Longley: Multioctave tunable 3-port circulator using a y.i.g. sphere.
Electronics Letters **6**, 406-408, 1970 (No. 13). *M*

J. Michel & F. Desvignes: Simulateurs de rayonnement solaire destinés à la mesure du rendement des photopiles au silicium.
Techniques Philips 1970, No. 3, 15-37. *L*

J. A. Pals: On the noise of a transistor with d.c. current crowding.
Philips Res. Repts. **26**, 91-102, 1971 (No. 2). *E*

P. J. L. Reijnen: Nonstoichiometry and sintering in ionic solids.
Problems of Nonstoichiometry, editor A. Rabenau, North-Holland Publ. Co., Amsterdam 1970, pp. 219-238. *E*

G. Rinzema: A simple wavelength-independent modulator for linearly polarized light.
Appl. Optics **9**, 1934, 1970 (No. 8). *E*

T. E. Rozzi: Modal analysis for nonlinear processes in optical and quasi-optical waveguides.
IEEE J. Quantum Electronics **QE-6**, 539-546, 1970 (No. 9). *E*

E. Schwartz: Über eine schärfere Ungleichung als $dX/d\omega > 0$ bei Reaktanzzweipolen.
Archiv elektr. Übertr. **24**, 491-495, 1970 (No. 11). *A*

J. M. Shannon, R. A. Ford & G. A. Gard (A. E. R. E., Harwell, England): Annealing characteristics of highly doped ion implanted phosphorus layers in silicon.
Radiation Effects **6**, 217-221, 1970 (No. 3/4). *M*

J. M. Shannon, J. Stephen (A. E. R. E., Harwell, England) **& J. H. Freeman** (A. E. R. E., Harwell): Ion doping of M.O.S. structures.
Proc. Int. Conf. on properties and use of MIS structures, Grenoble 1969, pp. 593-604. *M*

A. M. Stark: Computer methods for electrostatic field determination and ray tracing in image intensifiers.
Computer Aided Design **1**, No. 3, 3-10, 1969. *M*

W. Tolksdorf & P. Holst: Gemeinsames Sintern von Ferriten mit unterschiedlicher Sättigungsmagnetisierung und Curie-Temperatur für integrierte Mikrowellensysteme.
Ber. Dtsch. Keram. Ges. **47**, 670-673, 1970 (No. 10). *H*

J.-C. Tranchart: Préparation de monocristaux d'arséniure d'indium.
Acta Electronica **13**, 13-21, 1970 (No. 1). *L*

J. F. Verwey: On the mechanism of h_{FE} degradation by emitter-base reverse current stress.
Microelectronics and Reliability **9**, 425-432, 1970 (No. 5). *E*

J. O. Voorman & D. Blom: Noise in gyrator-capacitor filters.
Philips Res. Repts. **26**, 114-133, 1971 (No. 2). *E*

H. P. J. Wijn: Werkstoffe der Elektrotechnik.
Vorträge Rhein.-Westfäl. Akad. Wissensch. N 204, pp. 37-62, 1970. *E*

R.-E. Zeida: Réalisation et étude des propriétés électriques et photoélectriques des jonctions *p-n* en arséniure d'indium.
Acta Electronica **13**, 23-101, 1970 (No. 1). *L*

H. Zijlstra: Coercivity and wall motion.
IEEE Trans. **MAG-6**, 179-181, 1970 (No. 2). *E*

H. Zijlstra: A vibrating reed magnetometer for microscopic particles.
Rev. sci. Instr. **41**, 1241-1243, 1970 (No. 8). *E*

Contents of Philips Telecommunication Review **29**, No. 4, 1971:

H. L. Bakker: Long-term stability of 12 MHz coaxial line equipment 8TR 317 (pp. 145-148).
H. L. Bakker & L. F. Dert: Modulation system, type 8TR 331, for the transmission of television signals over 12 MHz line equipment for coaxial cables (pp. 150-159).
H. L. Bakker & L. F. Dert: Measuring results of simultaneous television and telephony transmission over 12 MHz coaxial cable (pp. 160-164).
S. H. Liem, J. P. de Raaff & R. T. van der Schaaf: Type UV telephone system for automatic trunk and international calls (pp. 165-186).
B. J. Beukelman: The undetected-error probability of codes using two-coordinate parity check and of cyclical codes (pp. 188-204).

Contents of Electronic Applications **30**, No. 2, 1970:

B. J. Leenhouts: Graphical analysis of reflections on long lines (pp. 45-52).
J. M. Siemensma: Simulation of a READ-only memory with the MOS READ-WRITE memory FDQ106 (pp. 53-63).
A. H. Hilbers: High-frequency wideband power transformers (pp. 64-73).
D. J. G. Janssen: Dynamic drive circuits for multiple-decade indicator tubes (pp. 74-86).

Contents of Mullard Technical Communications **12**, No. 111, 1971:

E. C. Snelling: Design of power transformers having ferrite cores (pp. 2-26).
M. Keohane: Pulse-modulator thyristor BTW35 as a radar magnetron driver (pp. 27-29).
B. Goudswaard: Solid aluminium electrolytic capacitors, 121-series (pp. 30-36).

Volume 31, 1970, No. 10 pages 303-334 Published 8th October 1971

PHILIPS TECHNICAL REVIEW

VOLUME 31, 1970, No. 11/12

Code modulation with digitally controlled companding for speech transmission

J. A. Greefkes and K. Riemens

In telephone communication the aim is always to transmit the speech signals with acceptable quality using the smallest possible bandwidth. If code modulation is employed this means that the frequency of the pulses in which the transmitted signal is coded must be kept as low as practicable. In some cases, especially in mobile communications, efforts in this direction go so far that speech signals that are only just intelligible at the receiver are considered acceptable. In this article a code-modulated transmission system is described that largely eliminates the adverse effect that variations of level have on the intelligibility of a speech signal. This is done by "companding" the variations in level of the signals, i.e. subjecting them to compression followed by expansion. In conventional analogue-transmission systems with companding there has always been the difficulty that companding could only be applied to the signals to a limited extent. This difficulty does not apply to the digital system described here. With this system, an intelligible speech signal can be transmitted at a low bit rate even when the interference level in the transmission path is nearly as high as the signal level.

Introduction

For the transmission of speech, music and other "information" ever-increasing use is being made of digital systems, where the signal is quantized in amplitude and in time. *Quantization in time* means that the instantaneous value of the analogue signal to be transmitted is not sent out continuously but only at certain instants in time, when the signal is "sampled". *Quantization in amplitude* means that the continuous scale of instantaneous values is replaced by a finite number of discrete values. Every time a sample is taken, one of these discrete values is transmitted. This can be done by supplying the sampled and quantized signal to an analogue-digital converter (the coder) which forms a digital signal consisting of pulses representing "1" of "0" (bits) that contain coded information about the instantaneous value of the analogue signal. (Sampling and coding can also take place simultaneously in the same part of the circuit.) At the receiver the bits are fed to a digital-analogue converter (decoder), which reconstitutes the analogue signal.

Because of the quantization the decoded signal differs from the original signal. This effect is known as *quantization distortion*. The difference between the two signals covers a wide spectral region, and includes noise referred to as *quantization noise*. In this article we are only interested in the part of this noise that falls in the frequency band of the transmitted audio signal. The term quantization noise as used here will therefore refer to that part alone.

Compared with systems for the direct transmission of analogue signals, digital systems have various advantages. The digital pulse pattern can be regenerated both at repeater stations and at the terminal station, and if the spacing between stations is small enough for the pattern to be regenerated before the signal has become so far attenuated that interference such as noise distorts it beyond recognition, the original bit pattern is transmitted free from distortion. Noise in the transmission

J. A. Greefkes and K. Riemens are with Philips Research Laboratories, Eindhoven.

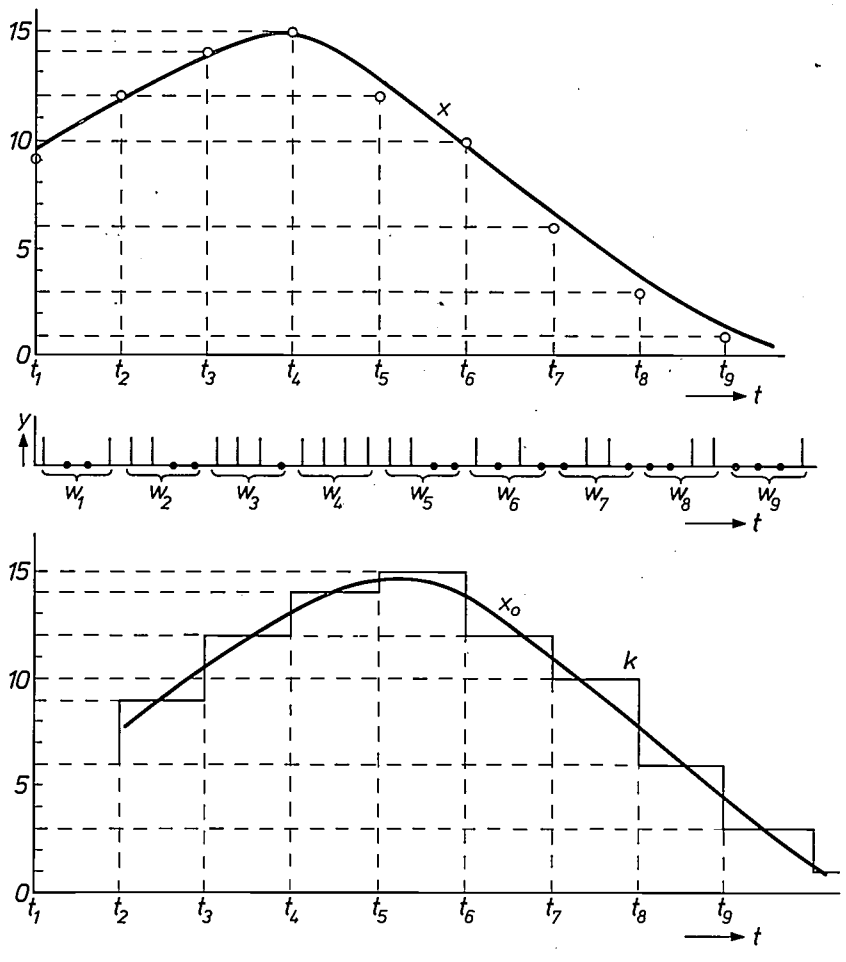


Fig. 1. Principle of pulse-code modulation (PCM) with code groups of 4 bits. The upper graph shows the signal for transmission x as a function of time t. The 16 discrete signal values (0 to 15) that can be coded with 4 bits are shown on the vertical axis. At the sampling instants ($t_1 \ldots t_9$, etc.) x is sampled and quantized at the values indicated by small circles. The bit pattern, y, in which the signal is coded is plotted along the second time scale (code groups $W_1 \ldots W_9$, etc.). Each code group is decoded at the receiver. The signal level remains constant until the next group is decoded (see lower graph). A stepped waveform k is thus received, from which x_o, a good approximation to x, is obtained by means of a low-pass filter. Since the decoding cannot begin until a complete code group has been received, there is a time lag in the signal received. The difference between x and x_o includes the quantization noise.

path does introduce a certain probability of error, but this is small. (For example, with a signal-to-noise ratio of 20 : 1 (13 dB) in the transmission path the probability of an error in the reception of a pulse is less than 10^{-5}.) Consequently, provided the signal level is sufficiently high, the bits are transmitted with a negligible number of errors. The length of the transmission path and the number of repeater stations have very little effect on this. The only noise then noticeable is the quantization noise. The fidelity of transmission of the analogue signals, after coding and decoding, is therefore virtually unaffected by the transmission path.

Another advantage of digital signal processing, which will probably become important in the future, is that it allows the time-division multiplex principle to be applied in a very simple and economic way to telephone-transmission links and switching devices for large numbers of telephone channels. The number of switching devices then needed is very much smaller than in conventional switching systems for analogue signals. The bit streams can be switched by means of coincidence circuits (AND gates) and digital memory devices can be used for temporary storage of the bits for the different channels. The bits can then be transmitted in any desired sequence and direction. Another important aspect of digital circuits is that integration techniques can readily be employed, giving very compact equipment.

The principal systems using quantized signals are *pulse-code modulation* and *delta modulation*. Pulse-code modulation is widely used for transmitting telephone signals. An article on these coding methods appeared in this journal as long ago as 1951 [1]. In the present article we shall discuss a number of new developments in speech transmission using digital signals. In particular we shall examine a number of methods designed to reduce the undesirable effect that variations of level in speech signals have on transmission quality. It will be shown that the use of a compander circuit which we have developed makes the application of delta modulation a particularly attractive proposition. First, however, we shall briefly recapitulate the principles of pulse-code modulation and delta modulation.

[1] J. F. Schouten, F. de Jager and J. A. Greefkes, Delta modulation, a new modulation system for telecommunication, Philips tech. Rev. 13, 237-245, 1951/52. See also J. F. Schouten, F. de Jager and J. A. Greefkes, Dutch patent No. 96166 (System for delta modulation, with associated transmitters and receivers), application made 22nd May 1948.

Code-modulation methods

Pulse-code modulation

In pulse-code modulation, PCM, the amplitude quantization takes place in such a way that the instantaneous value of the analogue signal is rounded off at every sampling instant to the nearest of the quantizing levels used. This value is then expressed as a number in the binary system. The number, referred to here as a *code group*, consists of a group of bits each of which may have the value 1 or 0. If, for example, code groups of four bits are used, $2^4 = 16$ different discrete values can be transmitted in this way. The method is illustrated in *fig. 1*. With code groups of five bits 32 discrete values can be coded and with seven bits 128. At every sampling instant the quantized signal thus goes up or down by as many quantizing intervals as are needed to approximate as closely as possible to the instantaneous value of the analogue signal. The interval between the quantization levels is referred to as the *quantizing unit*.

To transmit the complete information contained in the original analogue signal, the sampling rate must be at least twice the highest frequency in the signal spectrum. In the case of a speech signal, whose spectrum for telephony is usually limited to the frequency band between 200 and 3600 Hz, the sampling rate is therefore 8000 per second. The number of bits transmitted per second, the "bit rate", is equal to the product of the sampling rate and the number of bits in the code group. If there are seven bits in the code group, this gives a bit rate of 56 kilobits/second for telephony.

This only applies when the sampled series of bits follow each other continuously, as illustrated in fig. 1. As a rule, however, additional bits are needed for synchronizing the code groups, and this makes the bit rate somewhat higher. Moreover, in PCM a number of signals may often be transmitted simultaneously by time-division multiplex. The code groups transmitted in a sampling period then relate to different signals in turn, so that the bit rate increases in proportion to the number of signals to be transmitted.

Since the signal to be transmitted is approximated by a number of discrete values it follows that its amplitude is limited in two ways. If the highest discrete value is exceeded, exact coding is obviously not possible, and the same applies if the signal is smaller than the quantizing unit. These limits are of course independent of the frequency of the signal to be transmitted.

For each signal sample a certain number of quantizing units is used. The ratio of this number to the maximum number is called the *modulation index*. This is reflected in the bit pattern of each code group. If, for example, the most significant bit in a code group (the bit corresponding to the highest value in the binary number) is 1, the modulation index is greater than $\frac{1}{2}$.

The principle of a coder for PCM is illustrated in

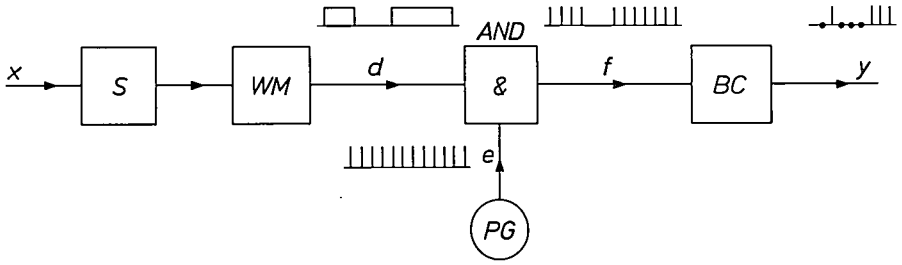


Fig. 2. Principle of a coder for pulse-code modulation. The signal to be coded x is sampled in the sampling circuit S and then fed to the pulse-width modulator WM. This delivers width-modulated pulses d to the AND gate, together with pulses e from the pulse generator PG. The output signal f from the AND gate goes to the binary counter BC, which delivers the PCM signal y.

fig. 2. The signal x to be coded is sampled and then converted into a signal d which consists of rectangular pulses with a constant height and a duration proportional to the instantaneous value of the signal at the instants of sampling (pulse-width modulation). These pulses are fed to an AND gate together with the continuous pulse trains e from a pulse generator. The number of the pulses let through from the train e is always proportional to the duration of the pulses d and thus proportional to the quantized signal. The pulse trains f thus obtained are fed to a binary counter (code converter), which reads out at the sampling rate to deliver the required PCM signal y.

A signal like that of fig. 1 is unipolar: the discrete values used in the quantizing process all have the same polarity. Usually, however, the signal will be an a.c. voltage. Positive and negative discrete values can then be used. It is also possible to make the signal unipolar by adding to it a d.c. voltage having such a value that the total signal always has the same polarity. Another way to make the signal unipolar is to pass it through a full-wave rectifier. In this case the polarity also has to be indicated; this is usually given by the first bit of each code group.

Differential pulse-code modulation

In another form of digital signal transmission, instead of quantizing the instantaneous value of the signal at the sampling instants, as in PCM, it is the difference between the instantaneous value and the quantized value for the previous sampling instant that is quantized. The magnitude and sign of this quantized difference are again converted into a code group of a number of bits [2]. This method is known as *differential pulse-code modulation* (*DiffPCM*). In fact what is now transmitted is not the signal itself but its time derivative. A DiffPCM system is therefore fully modulated (modulation index = 1) when the derivative of the signal has a certain maximum value. Consequently the maximum permissible amplitude of a sinusoidal signal to be coded is inversely proportional to the frequency. The significance of this for the transmission of speech will be dealt with presently.

by one unit. How this takes place is illustrated in *fig. 3*. Since no more than one pulse is sent out in every sampling interval, the term "code group" is not appropriate in DM; the bit rate here is equal to the sampling rate [3].

The principle of a circuit for generating DM signals is shown in *fig. 4*. A description of its operation is given in the caption. A coder of this type is essentially a negative-feedback circuit in which the feedback signal is quantized in both amplitude and time and is also integrated.

The integrator Int_1 consists in its simplest form of a resistor and a capacitor. (In this case an approximating signal like the one indicated by k in fig. 3 is produced.) A better approximation to the original signal can be obtained, however, by using an integrating circuit that gives double integration above a particular frequency. Because of the storage effect introduced by the double

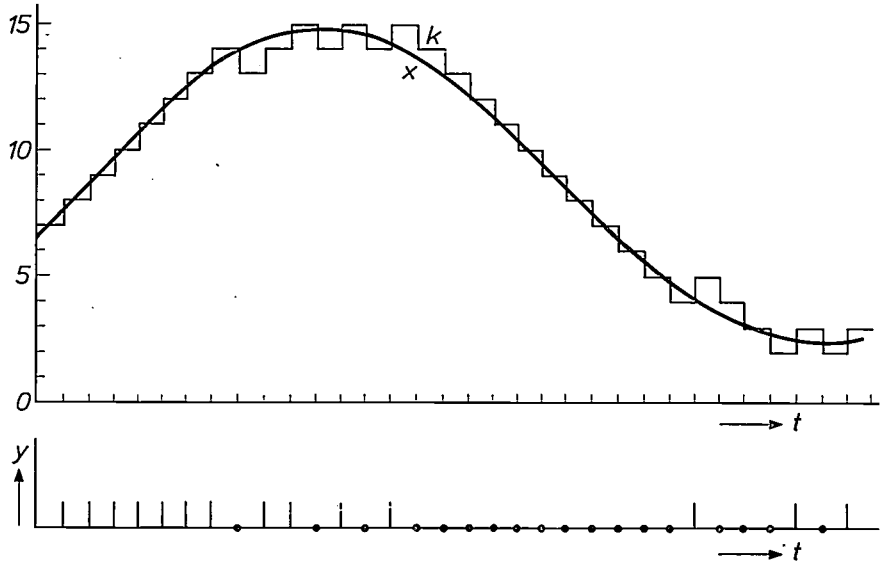


Fig. 3. Principle of delta modulation. The time scales, the bit rate and the discrete signal levels are the same as those of the PCM system in fig. 1. x is again the signal to be coded. k is the approximating signal, which goes up or down by one quantizing unit at each sample, depending on whether k is smaller or greater than x. The transmitted bit pattern y is plotted along the lower time scale.

Delta modulation

In delta modulation (DM) it is again the difference between the instantaneous value of the signal and the quantized value at the previous sampling instant that is quantized. Now, however, it is *not the magnitude* of this difference that is coded, as it is in DiffPCM, but *only the sign*. If the difference is positive a pulse is transmitted, causing the quantized value of the signal to rise by one quantizing unit in the receiver. If the difference is negative no pulse is sent out; the receiver reacts to this by making the quantized signal decrease

integration the input-signal value at more than one sampling plays a part in the approximation, and this reduces the quantization noise. In the transmission of speech an optimum signal-to-quantization-noise ratio can be achieved by designing the circuit Int_1 in such a way as to make the overload probability equal for all components of the average speech spectrum. The frequency above which double integration takes place is then about 2 kHz. A circuit of this type is shown with its attenuation characteristic in *fig. 5*.

In delta modulation the magnitude of the transmitted

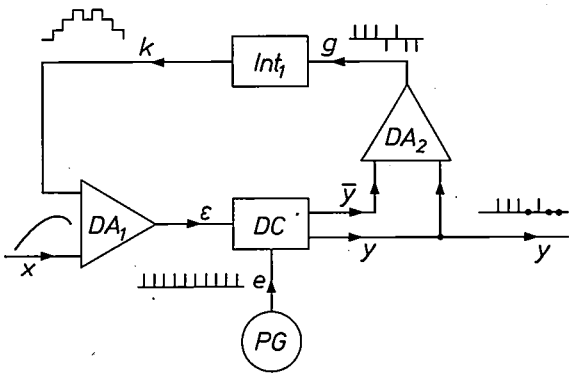


Fig. 4. Principle of a coder for delta modulation (DM). The differential amplifier DA_1 compares the instantaneous value of the signal x for transmission and the stepped approximating signal k. The sign of the output signal ε depends on the sign of the difference $x - k$. Depending on this sign the decision circuit DC passes the pulses e from the pulse generator PG as a pulse y or $\bar{y}$ on one of the outputs. The differential amplifier DA_2 now delivers a positive or a negative pulse (g) depending on whether a pulse y or $\bar{y}$ appears. Integration of these pulses in the integrator Int_1 results in the approximating signal k. One of the pulse trains y or $\bar{y}$ (here y) is sent out.

signal is limited in two ways. Since the quantized value always rises or falls one unit at each sampling, there is an upper limit for the time derivative just as there is in DiffPCM. With a sinusoidal signal the maximum permissible amplitude is again inversely proportional to the frequency of this signal. As with PCM, no coding is possible with DM for signals that are smaller than the quantizing unit.

In connection with this upper limit to the time derivative for proper coding of the DM signal, the modulation index is defined here as the ratio of the derivative of the signal to the maximum permissible value of this derivative. This ratio is again reflected in the bit pattern. If the pattern consists for example of a succession of three bits of value 1 followed by one bit of value 0 (or *vice versa*) then the modulation index is $\frac{1}{2}$. (Since three steps up and one down means two steps up in four samples.)

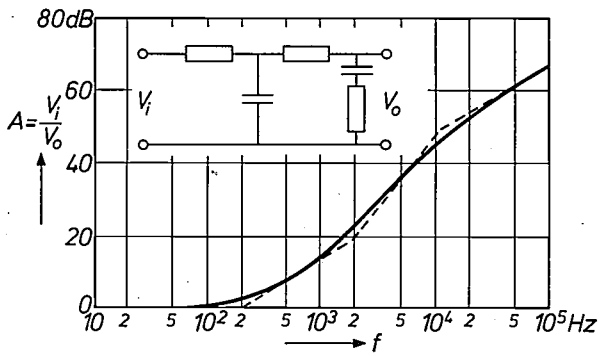


Fig. 5. Attenuation characteristic (attenuation A as a function of frequency f) of a network that can be used for speech transmission, at the position indicated in fig. 4 by Int_1.

In DM all bits of value 1 have the same significance: they indicate an increase in the height of the signal curve. Synchronization, which is necessary in PCM because the bits have differing significance, is therefore unnecessary in DM. (It does become necessary however if a number of signals have to be transmitted with DM in time-division multiplex.)

If the bit rate is already fixed (e.g. by the characteristics of the transmission path), then the sampling rate with DM is much higher than with PCM. This offers various advantages. In the first place it means that at a given signal-to-noise ratio a signal with a broader frequency band can be transmitted (for telephony, e.g. 0.2-5 kHz). With DM, furthermore, the requirements imposed on certain filters are not so severe. This applies in particular to the filter that is used in the transmitter to suppress components in the input signal with frequencies higher than half the sampling rate. The decoded signal at the receiver is also accompanied by unwanted components, whose frequencies increase with the sampling rate. The filter used for suppressing these components can also be simpler at a high sampling rate.

Other advantages of DM compared with PCM are the simplicity of the equipment and the fact that it does not have to meet such tight specifications as the equipment used for PCM. For example, the differential amplifiers in fig. 4 do not have to be linear, since it is only the *sign* and not the *magnitude* of the difference between instantaneous value and quantized value that is important. With PCM, on the other hand, linearity is important in certain parts of the circuit, in particular the sampling circuit S and the pulse-width modulator WM (see fig. 2).

The spectrum of the transmitted digital signal for DM differs considerably from that of a PCM signal. In the transmission of speech by DM the average number of pulses per second is constant (it is equal to half the bit rate), and therefore a DM signal contains a constant d.c. component. Furthermore, the spectrum contains no a.c. components at frequencies lower than those in the speech signal to be transmitted. This makes it possible to use capacitive coupling elements in amplifiers for DM signals. With PCM, on the other hand, there is no linear relationship between the instantaneous

[2] H. van de Weg, Quantizing noise of a single integration delta modulation system with an n-digit code, Philips Res. Repts. 8, 367-385, 1953. See also: R. W. Donaldson and R. J. Douville, Analysis, subjective evaluation, optimization, and comparison of the performance capabilities of PCM, DPCM, ΔM, AM, and PM voice communication systems, IEEE Trans. COM-17, 421-431, 1969.

[3] The system discussed here is sometimes referred to as "one-bit delta modulation". The term "multi-bit delta modulation" then refers to the differential pulse-code modulation discussed earlier.

value of the speech signal and the number of bits in the code groups. Because of this a PCM signal contains components at very low frequencies, which complicates amplifier design.

Quantization noise

The quantized signal is approximated by a stepped curve. In addition to other components, the difference between this curve and the original signal contains the quantization noise mentioned earlier. For PCM the ratio of the coded signal to this noise has been calculated by W. R. Bennett [4]. In the case of a sinusoidal signal whose magnitude is such that all the quantizing levels are utilized (maximum modulation) and using code groups of n bits, he found the following expression for the ratio of signal to quantization noise:

$$(S/N)_{max} = 6n + 3 \text{ (dB)}. \quad . \quad . \quad . \quad . \quad (1)$$

A speech signal, however, is not sinusoidal. A more practical equation, therefore, is one that relates to a signal of less regular waveform. For this we take a signal of 800 Hz, frequency-modulated by noise and limited in bandwidth (e.g. 750-850 kHz). Taking a sampling rate of 8 kHz, the signal-to-quantization-noise ratio at a bit rate f_p (expressed in kilobits/second) is then given by:

$$(S/N)_{max} = \tfrac{3}{4} f_p + 2 \text{ (dB)}. \quad . \quad . \quad . \quad (2)$$

The noise in this case has a flat spectrum. In *fig. 6* the points on curve *a* give the relation indicated by (2) between $(S/N)_{max}$ and f_p. (For n bits per code group, $f_p = 8n$. Of course, only the points relating to integral values of n are significant.)

For differential pulse-code modulation the signal-to-noise ratio has been calculated by H. van de Weg [2], E. N. Protonotarios [5], and others. Under the same conditions as quoted above for PCM, this ratio is:

$$(S/N)_{max} = \tfrac{3}{4} f_p + 6.5 \text{ (dB)}. \quad . \quad . \quad . \quad (3)$$

In fig. 6 the points on curve *b* give the values of $(S/N)_{max}$ corresponding to this equation.

Quantization noise for delta modulation was first calculated by F. de Jager [6]. Using a double-integrating circuit and with the coder fully modulated by a sinusoidal signal, he gives the signal-to-quantization-noise ratio as:

$$(S/N)_{max} = 10 \times \log_{10} \left(\frac{f_p^5}{f^2 f_m^2 f_0} \right) - 32 \text{ (dB)}. \quad . \quad (4)$$

Here f_p is again the bit rate, f is the signal frequency, f_m the frequency above which double integration takes place in the feedback loop, and f_0 is the upper cut-off frequency of the low-pass filter at the output of the decoder.

Equation (4) holds under the limiting condition that the bit rate f_p is high compared with the cut-off frequency f_0 (e.g. $f_p = 10 f_0$). For lower values of f_p the approximate equation:

$$(S/N)_{max} = 43 \times \log_{10} f_p - 28 \text{ (dB)} \quad . \quad . \quad (5)$$

has been experimentally established (see line *c* in fig. 6). In these experiments a double-integrating circuit with $f_m = 1.8$ kHz was used in the feedback loop, and the low-pass filter had a cut-off frequency f_0 of 3.4 kHz.

Equations (1) to (5) and fig. 6 are only valid for maximum modulation of the coder. Since the quantization noise is virtually independent of the signal level, the signal-to-noise ratio is nearly proportional to this level and therefore decreases as the signal level decreases. This is demonstrated for delta modulation on the gramophone record accompanying this article, and further particulars are given in the Appendix (demonstrations 1*a* and 1*b*). A demonstration on the same track illustrates the improvement in signal-to-noise ratio obtained with DM when a double-integrating circuit is used (demonstration 1*c*).

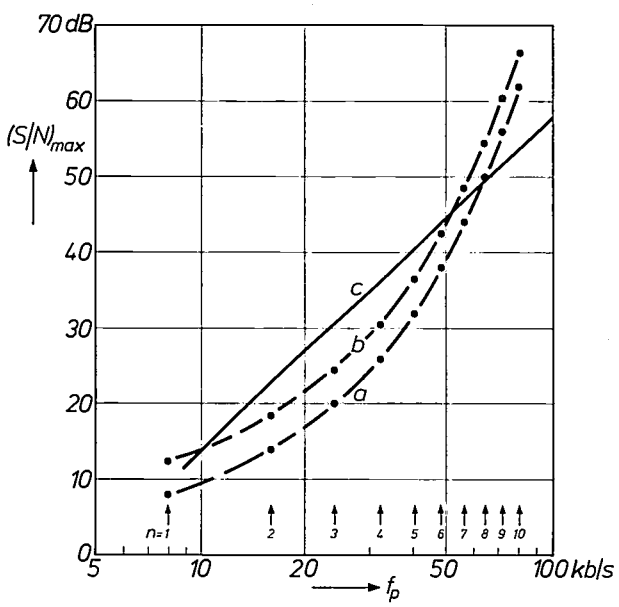


Fig. 6. Signal-to-noise ratio $(S/N)_{max}$ when the coder is fully modulated with a signal consisting of an a.c. voltage of 800 Hz, frequency-modulated by noise. The points on curve *a* relate to pulse-code modulation and those on curve *b* to differential pulse-code modulation (both at 8000 samples per second). Here n is the number of bits per code group. The line *c* relates to delta modulation, with a double integrating circuit as shown in fig. 5 in the feedback loop.

Fig. 6 enables us to make some comparisons between PCM and DM. For good speech transmission the signal-to-noise ratio must be at least 40 dB. From fig. 6 we see that for PCM the code groups must then consist

of 7 bits; for DiffPCM 6 bits are sufficient. At a sampling rate of 8 kHz this implies a bit rate of 56 kb/s in the one case and 48 kb/s in the other. As can be seen from curve *c*, a signal-to-noise ratio of 40 dB is reached at a bit rate of only 40 kHz for DM. This means that a smaller bandwidth is needed for the transmission path with DM. This advantage is even greater for communication where the transmitted speech does not have to be of very high quality, which is often the case with mobile links. If a signal-to-noise ratio of 20 dB is acceptable, fig. 6 shows that code groups of 3 bits are then required for PCM and DiffPCM, i.e. a bit rate of 24 kb/s. With DM, however, a bit rate of 16 kb/s is sufficient.

However, there is another important point that has a bearing on these figures. Since the signal-to-noise ratio decreases when the signal level decreases, a very low level of the input signal (or no input) is associated with a relatively high noise level, called "idle" noise. If nothing is done about this a speech-transmission system that meets the requirements given above would not be usable. In the following we shall consider measures that can be taken to reduce the noise with weak signals, using compression and expansion.

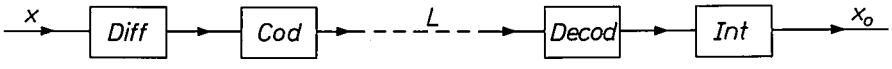


Fig. 7. PCM with equalization. By means of a differentiating circuit *Diff* at the input of the coder *Cod* (pre-emphasis) and an integrating circuit *Int* at the output of the decoder *Decod* (de-emphasis) the signal-to-noise ratio can be improved for the transmission of a speech signal. x is the speech signal to be coded, x_o the decoded output signal, L the transmission path.

Matching of the coding system to the signal spectrum

In most signal-transmission systems it is usual to keep the signal-to-noise ratio as high as possible and the bandwidth as small as possible. For digital-transmission systems a low bandwidth implies a low bit rate. To arrive at the optimum situation the systems under consideration should be studied in conjunction with the characteristics of the signal and the perceptional characteristics of the ear. This involves paying particular attention to the spectrum of the signal and to its dynamic range.

With regard to the *spectrum*, the bandwidth of the coding system must of course be sufficiently large to pass all the important components of the signal. We have already said that the sampling rate should be at least twice the highest frequency to be transmitted. We should also mention here that the bandwidth of the transmission channel ought to be equal to at least half the bit rate [7].

Apart from the width the shape of the spectrum of the signal to be coded is also important. If the spectrum is flat, i.e. if all components have the same average level [8], then the spectrum is matched to an overload limit which is independent of frequency, as is the case with PCM. In most cases, however, this is not so, and spectral matching is therefore necessary. In telephony *speech signals* have to be transmitted in the frequency band from 200 to 3600 Hz, and the average spectrum here has a maximum at about 500 Hz. As the frequency increases the components decrease in amplitude (by 8 to 10 dB per octave). In this case the full potential of PCM is more readily realized when the spectrum is *equalized*. The signal is then fed to the transmitter through a differentiating circuit that attenuates the components at higher frequencies less than those at lower frequencies (pre-emphasis). At the receiver an integrating circuit is used (de-emphasis, see *fig. 7*). The total transmission characteristic is then flat again. With suitably designed equalization networks the characteristics of this system correspond to those of differential pulse-code modulation. Therefore PCM with equalization can be considered as a form of DiffPCM.

The use of pre-emphasis and de-emphasis improves the signal-to-noise ratio, since both noise and signal pass through the de-emphasis network in the receiver, so that the higher-frequency components in the noise are also suppressed. Calculations have shown that with the system fully modulated the same equation applies as for DiffPCM (equation 3 and curve *b* in fig. 6).

In passing we should note that with PCM pre-emphasis can also be obtained using a circuit with a feedback loop containing a decoder and an *integrating circuit* (*fig. 8*). The signal-to-noise ratio here is about 1.5 dB greater than that given by equation (3) [2].

The situation is entirely different with DM where, as

[4] W. R. Bennett, Spectra of quantized signals, Bell Syst. tech. J. **27**, 446-472, 1948; B. Smith, Instantaneous companding of quantized signals, Bell Syst. tech. J. **36**, 653-709, 1957.

[5] E. N. Protonotarios, Slope overload noise in differential pulse code modulation systems, Bell Syst. tech. J. **46**, 2119-2161, 1967.

[6] F. de Jager, Delta modulation, a method of PCM transmission using the 1-unit code, Philips Res. Repts. **7**, 442-466, 1952.

[7] See for example E. Hölzler and H. Holzwarth, Theorie und Technik der Pulsmodulation, Springer, Berlin 1957, p. 114 *et seq.*

[8] This is the case with some types of electronic music.

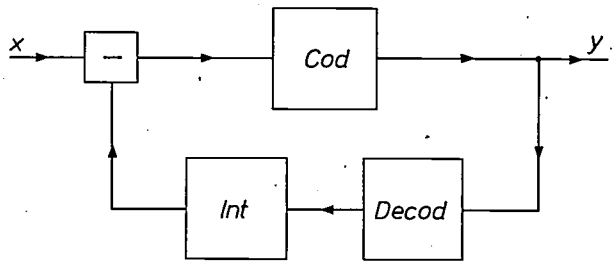


Fig. 8. In code modulation pre-emphasis can also be obtained by using a feedback loop incorporating a decoder *Decod* and an integrating circuit *Int*.

we have seen, the overload characteristic is not flat; at low frequencies a greater amplitude is permissible than at high frequencies. Because of this, DM is more or less naturally matched to the shape of the average speech spectrum. However, if signals are to be transmitted by DM in which all components of the spectrum are of the same average level, the coding system can be matched to these signals, i.e. the overload characteristic can be flattened, by inserting an integrating circuit in front of the coder. This means of course that a differentiating circuit has to be used in the receiver (*fig. 9*). This is known as a *sigma-delta modulation* system (ΣDM). Since the noise also passes through the differentiating circuit in the receiver, its higher-frequency components are emphasized, which causes a decrease of the signal-to-noise ratio. Psophometric measurements [9] have shown that the decrease is 7 dB. ΣDM is not therefore a very good system for the transmission of speech.

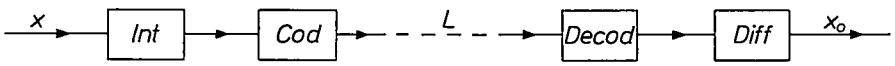


Fig. 9. With an integrating circuit *Int* at the input of the coder *Cod* and a differentiating circuit *Diff* at the output of the decoder *Decod* a delta-modulation transmission system can be matched to signals in which all components in the spectrum may be of equal amplitude (sigma-delta modulation).

Companding

Since the signal-to-noise ratio is proportional to the signal level, intelligibility can be very adversely affected by the kind of marked fluctuations in level of the transmitted signal that always occur in telephone communication. One of the devices used in analogue-signal transmission to minimize variations of level is the *compander*, a device that has formed the subject of an earlier article in this journal [10]. A *compressor* is connected to the input of the transmitter to reduce variations in level, and an *expander* is connected to the output of the receiver to restore these variations to their original values. This combination of compressing and expanding is referred to as *companding*.

The same principle can be applied to digital signal transmission, since a compressor can be connected to the input of the coder and an expander to the output of the decoder. In this case there is however another way of bringing about the compression and expansion. Up till now it has been tacitly assumed that the intervals between the quantizing levels are the same (*uniform quantizing*, UQ, also known as uniform coding), but this is in fact not necessary; levels of unequal intervals can be used instead (*non-uniform quantizing*, NUQ). These intervals then become greater when the instantaneous value of the modulation index increases, i.e. when the instantaneous value of the signal increases for PCM, and when the time derivative of the signal increases for DM.

There is another system that does use uniform quantizing, but with a variable quantizing unit. This unit depends on the *mean* value of the modulation index over a given time. In the system that we have developed the magnitude of the quantizing unit is controlled by the bit pattern of the coded signal. It is therefore referred to as *digitally controlled companding*. We shall show that this form of companding offers considerable advantage compared with NUQ, particulary for DM.

Pulse-code modulation with non-uniform quantizing

As we noted above, the spacing of the quantizing levels in PCM with NUQ depends on the instantaneous value of the signal to be coded. This system of companding is therefore known as *instantaneous companding*. If the signal is weak, the interval between the quantizing levels is small, if the signal values are high the levels are farther apart. In coder and decoder the intervals must obviously correspond exactly.

[9] i.e. measurements with a filter whose attenuation characteristic corresponds to the standard aural response characteristic.

[10] J. A. Greefkes, P. J. van Gerwen and F. de Jager, Companders with a high degree of compression of speech level variations, Philips tech. Rev. **26**, 215-225, 1965.

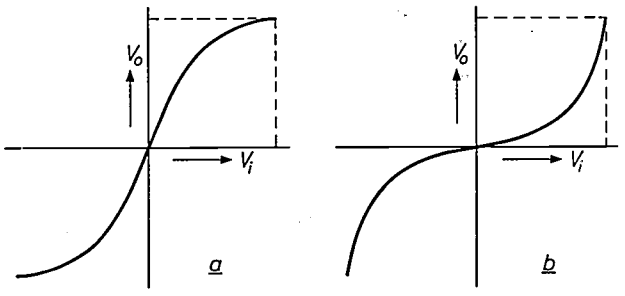


Fig. 10. Instantaneous companding in analogue-signal transmission can be achieved by passing the signal at the transmitter through a circuit with a characteristic like the one of (*a*). V_i input signal, V_o output signal. A circuit with the complementary characteristic (*b*) must then be used in the receiver. A drawback of this system is that it requires a larger bandwidth than is needed for transmitting the uncompanded signal.

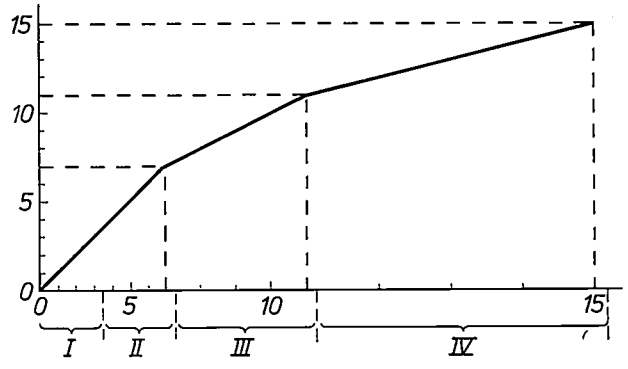


Fig. 11. Compression characteristic in a simple case of piecewise linear coding with code groups of 4 bits. The 16 discrete levels at which the signal is quantized are shown horizontally on an arbitrary scale. They are divided into four groups, I . . . IV, of 4 levels each. The intervals between these levels are equal to 1, 1, 2 and 4 quantizing units respectively. The code groups, numbered 0 . . . 15, in which the quantizing levels are coded, are shown vertically, on a linear scale.

It would also be possible to use instantaneous companding in analogue-signal transmission, by feeding the signal to the transmitter through an element with a characteristic curve like the one shown in *fig. 10a* (a compression characteristic). This would mean using a complementary characteristic (an expansion characteristic) in the receiver (fig. 10*b*). In practice, however, there are objections to this, because higher harmonics of the signal are formed in the transmitter. Although these are compensated in the receiver, they have to be transmitted faithfully, and this requires a larger bandwidth. This is not an objection when compression and expansion are employed with PCM and DM, since the sampling rate does not have to be increased. (The use of such a companding system does however decrease the accuracy with which the signals are transmitted.)

The intervals between the quantizing levels can be chosen in various ways. A frequently used system is illustrated in *fig. 11*. This simplified example relates to a unipolar signal transmitted with code groups of four bits each. There are thus $2^4 = 16$ discrete signal levels. These are divided into four groups, I to IV, each of which has four levels. In the first two groups there is no compression and the intervals between the levels are identical; we shall now call this interval the quantizing unit. In the third group the intervals are equal to two units and in the last group to four units.

Since the intervals in each of the four groups are constant, the compression characteristic (fig. 11) is composed of linear segments, and the system is referred to as *piecewise linear coding*. In each code group the first two bits now indicate in which of the four linear segments the related quantized voltage sample lies; the last two bits are a more exact indication of the position in each of the segments. *Fig. 12* shows an example of a signal quantized and coded in this manner.

In piecewise linear coding the smallest steps are

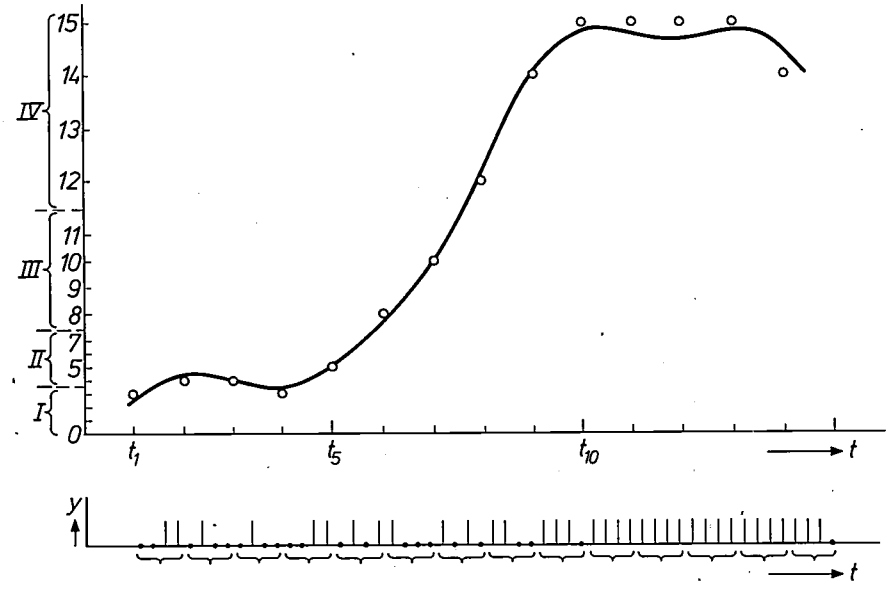


Fig. 12. Quantizing and coding of a signal for PCM with code groups of 4 bits, using piecewise linear coding with a compression characteristic as shown in fig. 11. The quantizing levels 0 . . . 15 are given on the vertical axis, the sampling instants t_1 etc. on the horizontal axis. Small circles again indicate the discrete instantaneous values of the signal at every sampling. At low signal levels greater accuracy and a better signal-to-noise ratio are now obtained than with uniform quantizing. At high signal levels, however, uniform quantizing is better. The bit pattern is shown along the lower time scale. The code groups are indicated by the brackets.

smaller than those which would be used if the same signal were to be uniformly quantized with the same number of discrete signal levels. Small signals are therefore more accurately reproduced than with UQ. To manage with the same number of levels, however, greater intervals must be used in piecewise linear coding for high instantaneous values of the signal than in UQ. The accuracy is then less than with UQ.

Since the intervals change with the instantaneous value of the signal, the quantization noise also varies with the signal. This is referred to as *modulation noise*. Also, because of the smaller steps with small signals the quantization noise with NUQ is less than with UQ. If, in the simple example given above, using code groups of 4 bits, the signals are so small that only the smallest steps are used, then the gain in signal-to-noise ratio is a factor of 4 (6 dB). Against this, however, there is a loss of about the same magnitude in signal-to-noise ratio when the coder is fully modulated.

Since when NUQ is used a low signal level is accompanied by greater accuracy and lower noise, there is an improvement in the intelligibility of the transmitted speech signals. This gain is greater than calculated above if the number of discrete signal levels is greater than 16 (i.e. if the code groups consist of more than 4 bits). Let us consider as an example the system recommended by the CCITT *(Comité Consultatif International de Télégraphie et Téléphonie)*. In this system an a.c. voltage is transmitted with code groups of 8 bits, of which 1 bit is used for indicating the polarity and 7 bits for quantizing the magnitude of the signal.

There are then $2^7 = 128$ levels available (and thus 127 intervals between the levels). The levels are now divided into 8 linear ranges each of them comprising 16 levels. In the first two ranges the intervals are again identical (one quantizing unit) and in the following ranges they are each increased by a factor of 2. The longest intervals are thus equal to 64 units.

It can easily be seen that in this case three bits of each code group indicate the linear range of the compression characteristic and four bits form a further subdivision of each range. The signal value at maximum modulation is then $31\times1 + 16\times2 + 16\times4 + \ldots + 16\times64 = 2047$ times the quantizing unit. This unit is therefore 127/2047 or about 16 times smaller than would be used with uniform quantizing in the same number of levels. With signals so small that only the smallest steps are used, the signal-to-noise ratio thus achieved is 16^2 times (24 dB) better than with UQ. When this coder is fully modulated the largest steps are used for a great deal of the time (for as long as the instantaneous value of the input signal is greater than half the maximum value). There is then a loss in signal-to-noise ratio compared with uniform quantizing. It can be shown that this loss is 10 dB in the transmission of an irregular signal (for which we can again take a signal frequency-modulated with noise in a narrow frequency band). In *fig. 13* curve *a* gives the signal-to-noise ratio as a function of signal strength, both on a relative scale, for this form of piecewise linear coding. Curve *c*, which relates to uniform quantizing, is included for comparison. The starting point (0 dB) is the signal-to-noise ratio $(S/N)_{max}$ given in fig. 6, as found for UQ with maximum modulation of the coder (at signal strength x_m). The improvement in signal-to-noise ratio that can be obtained with non-uniform quantizing in PCM at lower signal levels can be heard by comparing the demonstrations *a* and *b* of track 2 on the gramophone record.

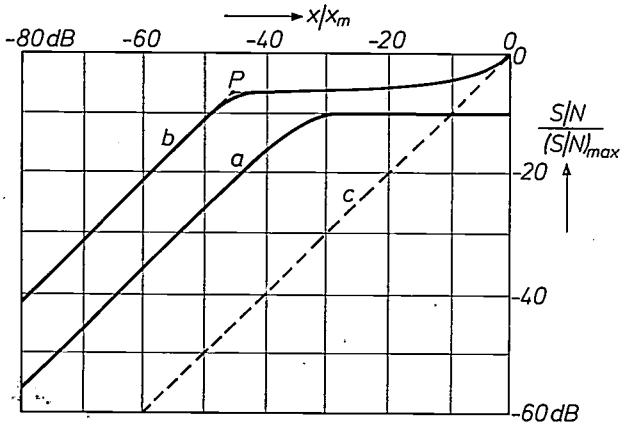


Fig. 13. Signal-to-noise ratio S/N as a function of the strength of a signal x (800 Hz, frequency-modulated by noise), both on a relative scale. The starting point (0 dB) is taken to be the signal-to-noise ratio $(S/N)_{max}$ when a signal x_m fully modulates a coder for uniform quantizing. Curve *a*: non-uniform quantizing (instantaneous companding); the ratio of the maximum to the minimum quantizing unit is 64. Curve *b*; digitally controlled companding; the ratio of the maximum to the minimum quantizing unit is 100. Curve *c*: uniform quantizing: there is a linear relationship between signal strength and signal-to-noise ratio.

Delta modulation with non-uniform quantizing

Instantaneous companding (NUQ) can also be used with delta modulation. Since the modulation level is determined here not by the magnitude of the signal but its *time derivative*, it is normal practice to use large steps if this derivative is large and small steps if its value is small. One companding system of this form has been proposed by M. R. Winkler [12]. In this system successive steps in the same direction always double in height, except that two steps in the same direction that follow a change in direction always have the same magnitude. If the steps alternate in direction, each step is half the previous one (*fig. 14*). Here again there is

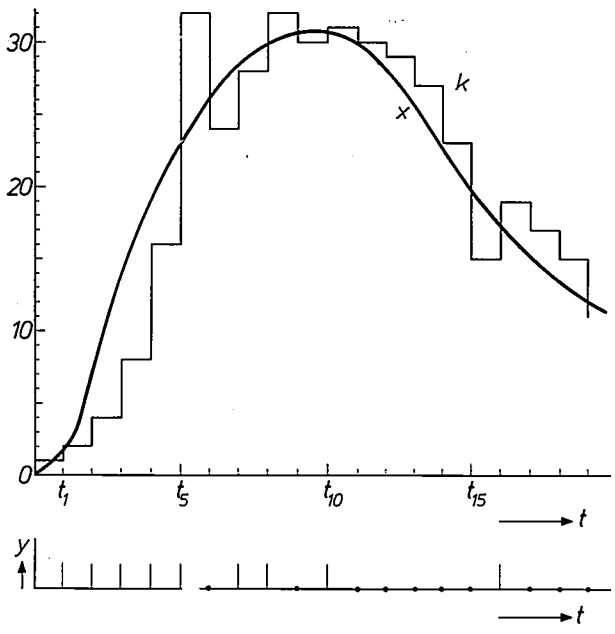


Fig. 14. Example of delta modulation with non-uniform quantizing, after M. R. Winkler [12]. Each successive step in the same direction is increased by a factor of 2. (Except for two steps in the same direction following a change in direction. These have the same magnitude.) If the steps alternate in direction, each step is half the previous one. The vertical axis gives the equidistant quantizing levels. x is the signal for transmission, k the approximating signal, both as a function of the time t; t_1 etc. sampling instants. The transmitted bit pattern is again shown along the lower time scale.

of course an upper and a lower limit to the magnitude of the steps. Clearly the large steps are used when the time derivative of the signal during a more or less "long" time is not very different from its maximum value; for a sinusoidal signal this means that the product of amplitude and frequency is large.

As in PCM with non-uniform quantizing, the smallest steps are smaller and the largest steps greater than in uniform quantizing. This again means a gain in signal-to-noise ratio at a low signal level, and a loss at a high signal level. If the ratio of the largest to the smallest steps is taken to be the same for both PCM and DM, the curves that give the signal-to-noise ratio as a function of signal strength on relative scales for PCM and DM are very nearly coincident.

Controlled companding

In the transmission of analogue signals it is common practice to use a companding system in which the degree of compression and expansion is controlled with a reaction time of about 30 milliseconds. This system cannot then follow the dynamic variations within a single syllable, and is therefore described as "syllabic companding". If the same principle is applied to digital signals, the compression and expansion can be obtained by varying the magnitude of the quantizing unit.

To restore the transmitted signal completely to its original form, it is necessary to control the quantizing unit simultaneously and in exactly the same way in the coder and decoder. This can be achieved by means of an *auxiliary signal*, as used for companders with a high degree of compression [10]. Examples of such systems combined with delta modulation are continuous delta modulation [13] and two-channel delta modulation [14]. The auxiliary signal must be transmitted from the transmitter to the receiver very faithfully, which in practice is a difficult operation. In the system we have developed, which we shall call *digitally controlled companding*, an auxiliary signal is not needed. In this system the magnitude of the quantizing unit is derived from the bit pattern in such a way that it depends upon the *mean value* of the modulation index during a certain time. We have chosen this time as 5 milliseconds, as it has been found by experiment that the quality of the transmitted speech signals is then better than with the 30 milliseconds mentioned above [13]. Since the bit pattern is exactly the same in transmitter and receiver, the variations in magnitude of the quantizing unit in transmitter and receiver can be made to follow each other accurately.

Digitally controlled companding

The basic diagram of a digitally controlled companding system (often called "digitally controlled code modulation") is shown in *fig. 15*. The diagram applies

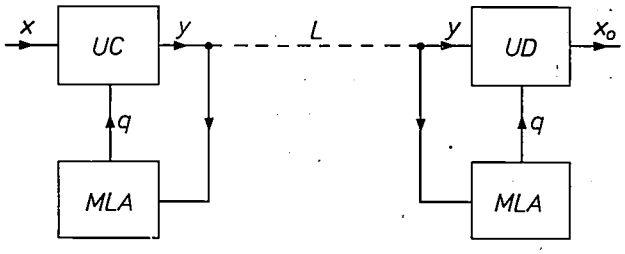


Fig. 15. Basic diagram of a transmitter and receiver for code modulation (PCM or DM) with digitally controlled companding. *UC* uniform coder, *UD* uniform decoder. *MLA* modulation-level analysers which determine the relationship between the digital signal y and the magnitude q of the quantizing unit of *UC* and of *UD*. *L* transmission path. x signal for transmission. x_0 received and decoded signal.

[11] Document WP 33/XV, No. 20 E of the CCITT report of the meeting of study group XV, Geneva, Feb. 7-11, 1966.
[12] M. R. Winkler, High information delta modulation, IEEE Int. Conv. Rec. **11**, No. 8, 260-265, 1963.
[13] J. A. Greefkes and F. de Jager, Continuous delta modulation, Philips Res. Repts. **23**, 233-246, 1968.
[14] J. A. Greefkes and K. Riemens, Dutch patent No. 6616294 (System for signal transmission by pulse delta modulation, with associated transmitters and receivers), application made 22nd November 1966.

both to PCM and DM. A uniform coder *UC* is used at the transmitter and a uniform decoder *UD* at the receiver. The quantizing unit q is varied by the output voltages from the modulation-level analysers *MLA*, which are controlled by the digital signal y. (The circuit will be dealt with in more detail presently.) The magnitude of q is now a function of y, and hence of the input signal x. If q were proportional to x, then y and hence the modulation index would be independent of x. This ideal state is not feasible in practice, but our system approximates to it.

If the two analysers *MLA* are identical, the magnitude of q is identical in coder and decoder, and the output signal x_o is then equal to the input signal x.

The modulation index, which is a measure of the degree of compression, can be measured by feeding the digital signal to a decoder operating with a constant quantizing unit. One way of doing this would be to make the quantizing unit of the decoder *UD* always equal to q_m, the maximum value of q. In this case x_o is a measure of the modulation index.

Modulation-level analyser

The relation between the quantizing unit and the digital signal can be chosen in many ways. Various types of modulation-level analyser can therefore be used. We shall discuss here a circuit that has given satisfactory results in our experiments. The principle is shown in *fig. 16*. It amounts to determining from the bit pattern how many times in a particular time interval the modulation index exceeds a particular value, which we have taken as half of the maximum value. The digital signal y (see fig. 15) is fed to the *threshold circuit Th*, which delivers a voltage signal V to the integrating circuit *Int*, whose time constant is 5 ms. As long as the modulation index α is greater than $\frac{1}{2}$, V has the value V_0. If α becomes smaller than $\frac{1}{2}$, then V goes to zero. The integrating circuit *Int* provides a control voltage V_c, which varies with the mean value of α between zero and a set maximum value, V_{cm}. The voltage V_c controls the amplitude modulator *AM*, which sets the quantizing unit q between zero and the maximum value q_m. If the input signal is very small, the value of α does not reach $\frac{1}{2}$ at any sampling, and therefore q would be zero. To prevent this, a small constant voltage δV_{cm} ($\delta \approx 0.01$) is added to the control voltage V_c. This added voltage thus determines the minimum value of q.

Just as in NUQ, the quantization noise varies with signal strength and therefore has the characteristics of modulation noise. There is an important advantage, however, in the use of controlled companding. In instantaneous companding, as we have seen, the gain in signal-to-noise ratio obtained with small signals is offset by a loss at high input levels, because the use of smaller quantizing intervals at small instantaneous signal values makes it necessary to use larger intervals at high instantaneous values. The higher quantization noise which this entails limits the choice of the ratio between maximum and minimum quantizing intervals.

In controlled companding, on the other hand, it is possible to use the same quantizing unit at maximum modulation of the coder ($\alpha = 1$) as for uniform coding. There is therefore no loss in signal-to-noise ratio at high signal levels. A drop in signal level results in a smaller quantizing unit, and the system can be designed in such a way that this causes only a slight decrease in signal-to-noise ratio. Consequently, the signal-to-noise ratio is greater *at all signal levels* than with uniform quantizing. To illustrate this in quantitative terms, let

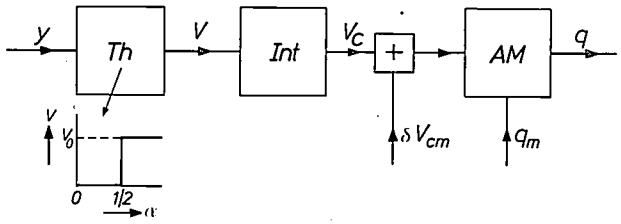


Fig. 16. Basic diagram of a modulation-level analyser. *Th* threshold circuit; the output voltage V is equal to a fixed value V_0 as long as the modulation index α is greater than $\frac{1}{2}$. If α is smaller than $\frac{1}{2}$, then V is zero. *Int* integrating circuit that delivers a control voltage V_c varying between 0 and V_{cm}. This control voltage is fed together with a constant voltage δV_{cm} to the amplitude modulator *AM*, which varies the magnitude q of the quantizing unit. q_m maximum value of this unit.

us consider a sinusoidal signal, $a = A \sin \omega t$, whose frequency is so low that the number of samples per period is fairly high. If this signal has the maximum permissible value, ($x/x_m = 1$, i.e. 0 dB), then a is greater than $\frac{1}{2}A$ during two-thirds of each period, which means that $\alpha > \frac{1}{2}$ for PCM. During the same fraction of the period $|da/dt|$ is also greater than $\frac{1}{2}A\omega$, which means that $\alpha > \frac{1}{2}$ for DM. The system is now arranged so that the quantizing unit has the magnitude it would have had for uniform quantizing. The signal-to-noise ratio S/N is then equal to $(S/N)_{max}$, the value that would be found with a fully modulated uniform coder.

If the signal level decreases, the number of samples for which $\alpha > \frac{1}{2}$ decreases. After a time of at least a few times 5 ms a new equilibrium sets in, with smaller values of q. Although the number of samples for which α is greater than $\frac{1}{2}$ has now increased again, the number is nevertheless smaller than in the previous equilibrium state. The decrease in the magnitude of the quantizing steps is therefore only slightly less than proportional to the signal level, so that there is not much decrease

in the signal-to-noise ratio. In fig. 13 curve *b* shows the signal-to-noise ratio for controlled companding in our system, where the minimum value of the quantizing unit is one hundredth of its maximum value. The minimum value is used when the signal level has dropped to a threshold below which samples with $\alpha > \frac{1}{2}$ no longer occur. If the magnitude of the quantizing unit q were constant, α would always remain smaller than $\frac{1}{2}$ if the amplitude of a sinusoidal signal were half that for maximum modulation. However, because of the reduction of q this situation does not arise until the signal has become a hundred times smaller, that is to say 200 times smaller than with maximum modulation. The signal level is then −46 dB, and the signal-to-noise ratio compared with that for maximum modulation has decreased by a factor of 4 (6 dB). This theoretical situation is indicated in fig. 13 by the point *P*.

If the signal decreases still further, q remains constant; the signal-to-noise ratio then remains larger by a factor of 100^2 (40 dB) than with uniform quantizing (see the left-hand part of curve *b*). The improvement in transmission quality that can be achieved with NUQ and with controlled companding can clearly be heard in tracks 2 and 3 on the gramophone record.

Operation of the threshold circuit

In the treatment of controlled companding we made no distinction between PCM and DM. In fact the principle described can be applied in both cases, using largely identical circuits. The only difference between PCM and DM is in the design of the threshold circuit (*Th* in fig. 16), which is required to determine from the digital pulse pattern at each sample whether or not the instantaneous value of the modulation index exceeds half the maximum value.

A PCM system using digitally controlled companding is often referred to as a DCPCM system (digitally controlled pulse-code modulation). As we saw earlier, the modulation index here exceeds the value $\frac{1}{2}$ if the input signal level is greater than half that at which the coder is fully modulated. This is reflected in the code group by the presence of the most significant bit. In this case the binary counter *BC* in fig. 2 could be used as the detector. Whenever this has received a complete code group, the most significant bit is contained in a fixed element. The state of the appropriate bistable circuit (flip-flop) can then be used for controlling a circuit which switches the voltage V on or off.

When digitally controlled companding is applied in delta modulation, the instantaneous value of the modulation level depends on the time derivative of the signal, and the threshold circuit has to detect whether this derivative is greater than half the maximum value that can be transmitted. This is also reflected in the pulse pattern, since the modulation index exceeds the value $\frac{1}{2}$ when more than three successive bits have the same value. The detector here can be a shift register consisting of four bistable circuits (flip-flops), as indicated by SR_1 in *fig. 17*. The outputs of the stages are connected to an AND gate, and the inverted outputs are connected to another AND gate. The two AND gates are connected to an OR gate, OR_1. When the digital signal y is applied to the input of SR_1, one of the AND gates, and hence also the OR gate, delivers a "1" voltage when all four bistable circuits are in the state "1" or "0". As soon as a bit of opposite sign enters the register, one of the AND gates changes state and OR_1 also changes state. The circuit therefore detects both positive and negative derivatives.

The shift register SR_1 causes a certain time lag in the response of the threshold circuit, because the equality of four or more successive bits is not indicated by the bistable circuits until the fourth has appeared at the input. The end of such a series of bits is observed, how-

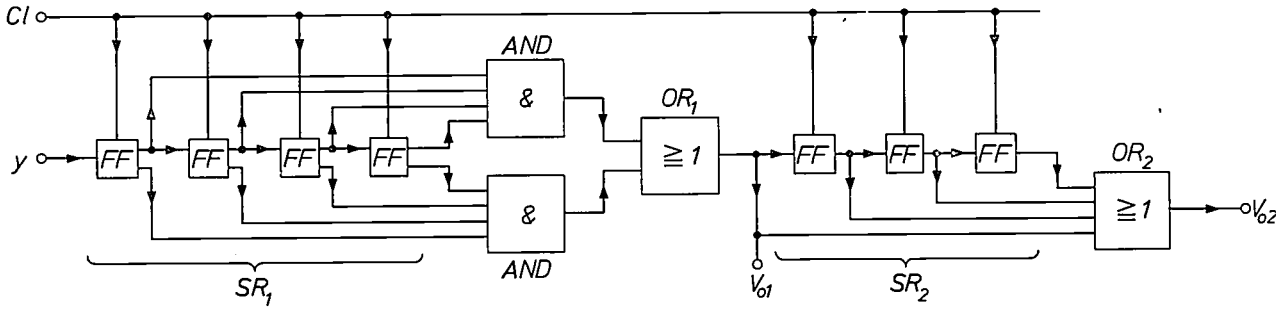


Fig. 17. Block diagram of a modulation-level analyser for delta modulation. SR_1 and SR_2 shift registers. *FF* stages of the shift registers; AND and OR gates are indicated in the figure. y digital input signal. *Cl* clock signal. The output voltage V_2 of OR_2 has the value "1" for times that are equal to the length of pulse trains of more than three identical bits in y, unlike the output signal V_1 of OR_1, in which each series of identical bits is followed by a dead time of 3 bits (see also fig. 18).

ever, as soon as a series of identical bits is interrupted by an opposite bit (in which case the shift register SR_1 no longer contains four identical bits). The time during which OR_1 indicates that the modulation index has at least one half of its maximum value will therefore invariably be 3 bit periods shorter than the corresponding series of identical bits. Each such series is consequently followed by a dead time of 3 bit periods. Thus, three bits are lost every time the derivative of the signal changes sign, which means that at high signal frequencies it would not be possible to determine the modulation index. To get around this difficulty the shift register SR_2 has been added [15]. The outputs of the bistable circuits of the shift register and also the output of OR_1 are connected to a second OR gate, OR_2. As can easily be seen the output voltage V_2 of OR_2 is equal to the output voltage V_1 of OR_1. The only difference being that V_2 takes over the change of state from "1" to "0" three bit periods later. The latter voltage V_2 is thus "1" for times that are equal in length to the incoming pulse trains of more than three identical bits. This is illustrated in *fig. 18*, which shows a signal to be coded x, the quantized signal k, the bit pattern y, and the inverted bit pattern $\bar{y}$. The output voltages V_1 and V_2 of the OR gates OR_1 and OR_2 are also shown.

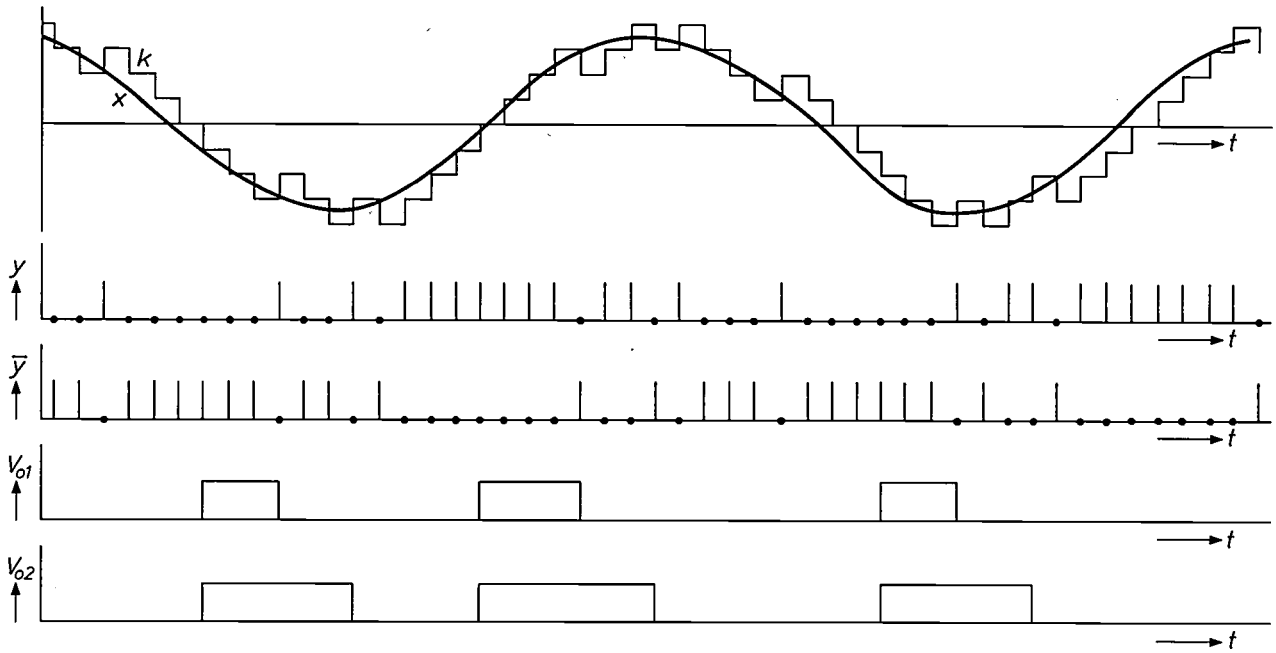


Fig. 18. Illustrating (for single integration) the operation of a modulation-level analyser for delta modulation corresponding to the block diagram of fig. 17. x signal for transmission. k stepped approximating signal, y bit pattern in which k is coded. y inverted bit pattern. V_1 output signal of the OR_1 gate. V_2 output signal of the OR_2 gate. This signal is a measure of the duration of pulse trains of more than three identical bits.

Fig. 19a shows the block diagram of a coder which we have developed for DM with digitally controlled companding (referred to in previous publications as digitally controlled delta modulation, DCDM). As a comparison with fig. 4 shows, in addition to the feedback loop used in uniform quantizing, the circuit now contains a second feedback loop formed by the modulation-level analyser (*Th*, Int_2 and *AM*; see fig. 16).

In DM with uniform quantizing (fig. 4) the feedback loop has a stabilizing effect on the circuit. Any asymmetry in the differential amplifier DA_1 gives rise to a difference between the average number of positive and negative pulses arriving at the input of the integrating circuit Int_1. The output from this circuit is then a d.c. voltage which restores the symmetry in the differential amplifier. If controlled companding with a high degree of compression and expansion is used, this stabilizing effect is almost non-existent for weak signals, since the pulses fed to Int_1 are then very small. In view of this the circuit was given a third feedback loop, which also consists of a differential amplifier and an integrating circuit (DA_3 and Int_3). Since the pulses received by DA_3 come direct from the decision circuit *DC*, their magnitude does not decrease when the signal level is low. The stabilizing effect has thus been restored.

Since the output voltage from Int_3 is fed to the differential amplifier DA_1 together with the signal for transmission, there is a danger that the signal will be affected by the third feedback loop. To avoid this, Int_3 should have a fairly large time constant. The atten-

uation characteristics of the circuits Int_1, Int_2 and Int_3 are given in *fig. 20* (the characteristic for Int_1 was also given earlier in fig. 5).

Fig. 19*b* shows the block diagram of a *decoder* for DM signals with controlled companding. The basic circuits are the same as those of the coder. Because of this the transmitter and receiver can be built up in a simple way with integrated circuits, only two types of chip being needed, one for the analogue part of the circuit and one for the digital part. The only circuits not included on the chips are the integrating circuits Int_1, Int_2 and Int_3. *Fig. 21* shows a coder and a decoder built up in this way. In each of these there is one chip of each type.

Comparison of the different companding systems

To get some idea of the improvement in signal-to-noise ratio that can be achieved with digital companding, both with PCM and with DM, we must compare the signal-to-noise ratios of the various systems in terms of their *absolute value*. The necessary data can be derived from figs. 6 and 13. Some results of such a comparison are presented in *figs. 22, 23* and *24*. In fig. 22 the signal-to-noise ratio is plotted as a function of signal level at a bit rate of 56 kb/s. Curve *a* relates to PCM with non-uniform quantizing (PCM-NUQ) and curve *b* to PCM with digitally controlled compand-

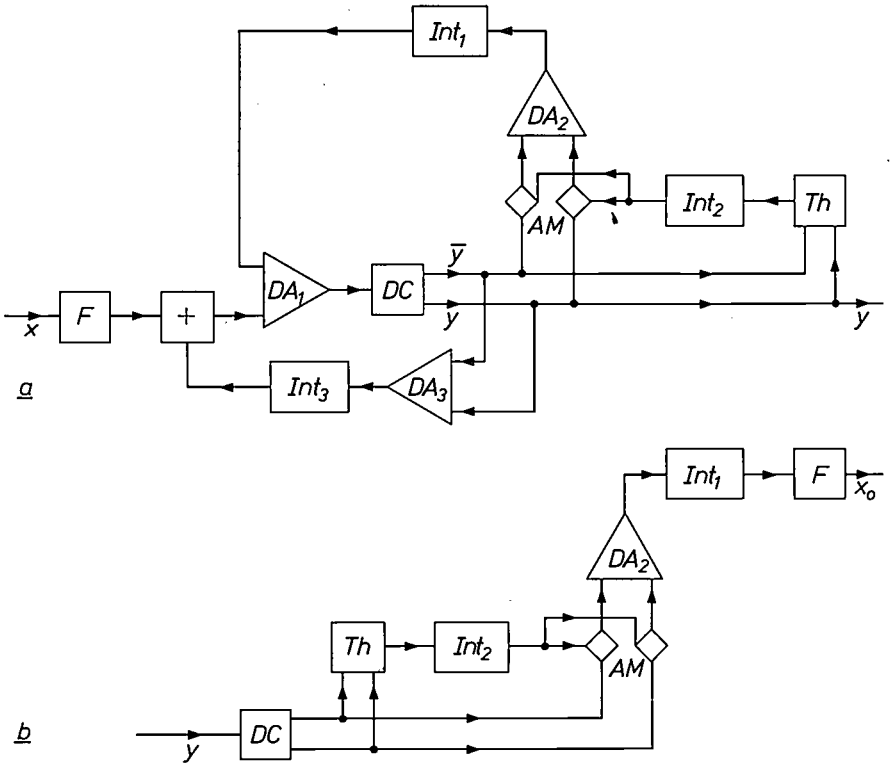


Fig. 19. *a*) Block diagram of a coder for delta modulation with digitally controlled companding (DCDM). The modulation-level analyser, formed by the threshold circuit *Th*, the integrating circuit Int_2 and the amplitude modulators *AM*, now constitute a second feedback loop, in addition to the one formed by DA_2 and Int_1 (see also fig. 4). A third loop, formed by the differential amplifier DA_3 and the integrating circuit Int_3, stabilizes any asymmetry in the differential amplifier DA_1. *F* low-pass filter. *b*) Block diagram of a decoder. This is built up from the same basic circuits as the coder.

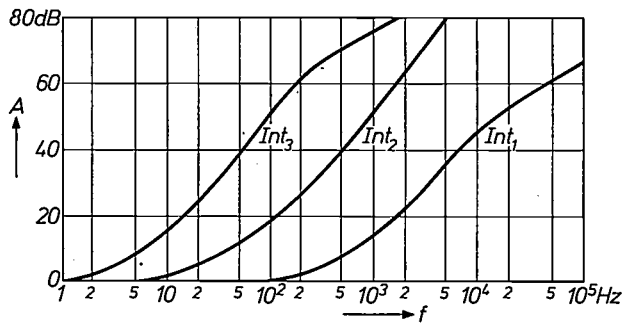


Fig. 20. Attenuation characteristics of the circuits Int_1, Int_2 and Int_3, as used in a coder and a decoder for DCDM corresponding to the block diagrams in fig. 19.

[15] J. W. Glasbergen, J. A. Greefkes, A. W. M. van den Enden and F. de Jager, Dutch patent No. 6803992 (System for speech transmission with pulse delta modulation, with associated transmitters and receivers), application made 21st March 1968.

ing (DCPCM). These curves in fact correspond to curves *a* and *b* in fig. 13 (where, however, a relative vertical scale was used). As we saw earlier from fig. 13, digitally controlled companding gives a better signal-to-noise ratio than NUQ. A further improvement of 4.5 dB is obtained with digitally controlled differential pulse-code modulation (DiffPCM, curve *c*). This system gives a slight improvement compared with delta modulation using digitally controlled companding (DCDM, curve *d*).

The results at an even lower bit rate, 16 kb/s, are presented in fig. 24, which gives the curves for the two best systems, Diff PCM and DCDM. It can be seen that none of these systems now meet the specification for satisfactory telephone communication. In practice, however, it has been found that with DCDM speech can be transmitted intelligibly at such a low bit rate.

The calculated curves in figs. 22, 23 and 24 agree very well with the measured results. As an illustration, *fig. 25* gives some theoretical curves for DCDM compared with the results of measurements. Curves *a*, *b* and *c* relate to bit rates of 40, 24 and 16 kb/s. The main discrepancies between theory and measurement are to be found at high signal levels. These discrepancies are not due to quantization noise but to signal distortion. In this connection it is important to note that signal distortion is much less troublesome than noise in telephone communications; distortion of a few per cent in a speech signal is of little significance.

Fig. 21. Practical circuit of a coder (*left*) and a decoder unit (*right*) for DCDM corresponding to the block diagrams in fig. 19. Except for the integrating circuits and one or two discrete components the units are made in integrated-circuit form; only two different types of chip are required.

Curve *e* in fig. 22 gives the CCITT specification for the signal-to-noise ratio required for good voice transmission by telephone. It can be seen that all the systems amply meet this specification at this high bit-rate. This is not always the case at lower bit rates, as can be seen from fig. 23, which relates to a bit rate of 32 kb/s. Curves *a* to *d* in this figure refer to the same systems as the corresponding curves in fig. 22. It can be seen that only DCDM now meets the CCITT specification.

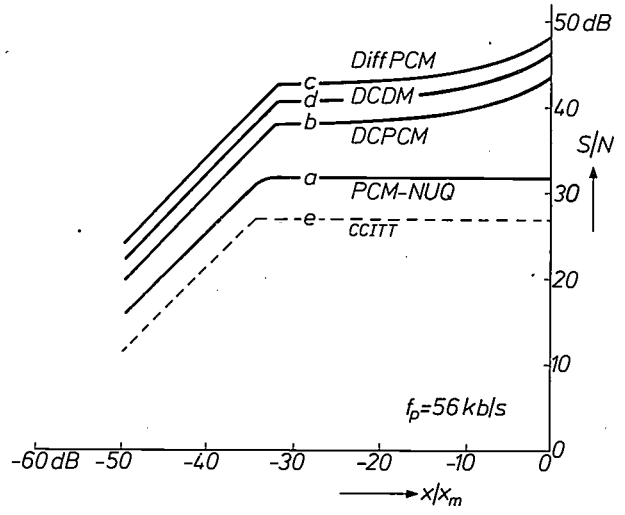


Fig. 22. Signal-to-noise ratio S/N at the output of the decoder as a function of the signal level x/x_m, at a bit rate of 56 kb/s. The curves relate to the following systems. Curve *a*: PCM with piecewise linear coding. Curve *b*: PCM with controlled companding. Curve *c*: differential PCM with controlled companding. Curve *d*: delta modulation with digitally controlled companding (DCDM). Curve *e* gives the CCITT specification for a good telephone link. At this high bit rate all the systems amply meet this specification.

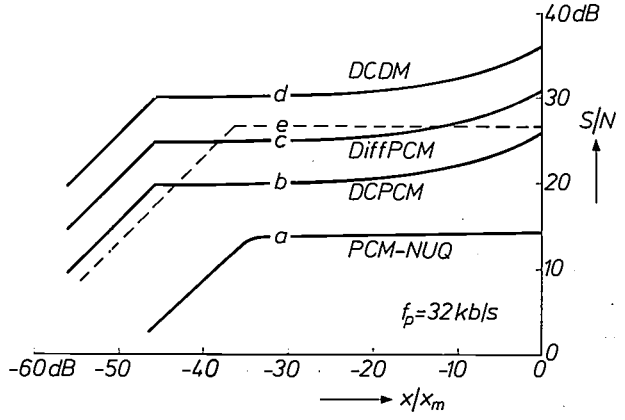


Fig. 23. Signal-to-noise ratio S/N as a function of the signal level x/x_m at a bit rate of 32 kb/s. Curves *a* . . . *d* relate to the same systems as the corresponding curves in fig. 22; *e* is again the CCITT specification. It can be seen that only DCDM now meets this specification.

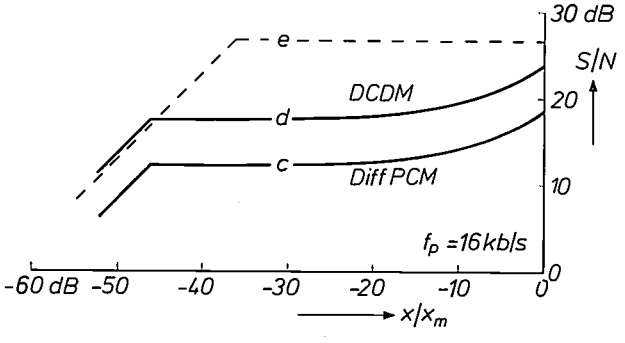


Fig. 24. Signal-to-noise ratio S/N as a function of the signal level x/x_m at a bit rate of 16 kb/s. Curve *c*: differential PCM with controlled companding. Curve *d*: DCDM. The line *e* is again the CCITT specification for a good telephone link, which none of the systems now satisfy. With DM, however, intelligible communication is still possible.

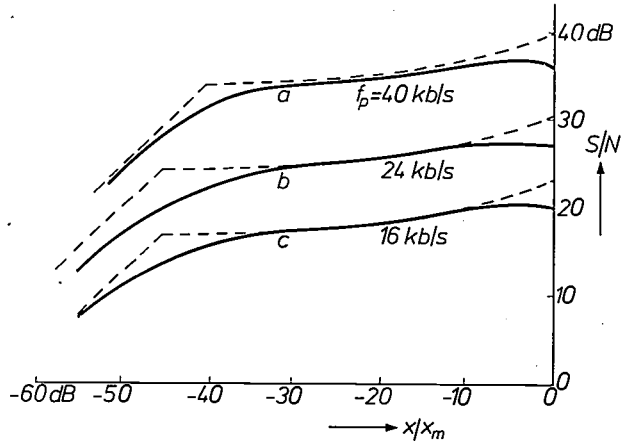


Fig. 25. Measured signal-to-noise ratio S/N as a function of the signal level x/x_m for DCDM at three different bit rates (solid curves: *a* bit rate 40 kb/s, *b* bit rate 24 kb/s, *c* bit rate 16 kb/s. The dashed lines give the calculated values. The discrepancy between the theoretical and practical results at high signal levels (on the right) is attributable to distortion of the signal.

The effect of interfering signals in the transmission path

As we said earlier, in digital signal transmission the distortion of the received signal and the noise it contains are practically independent of the length of the transmission path and of interfering signals introduced in the transmission path. This does not of course apply if the pulses received are so weak that interference in the transmission path causes errors in the regeneration of the pulses at repeater stations and terminal station. Radio links in particular can present such difficulties, caused for example by fading [16].

If a certain percentage of the bits received contains errors, a PCM signal is found to be more mutilated than a DM signal under otherwise identical conditions. If for example in PCM a significant bit in a code group corresponding to a small instantaneous value of the transmitted signal has the incorrect value 1, the decoded signal will show a large and undesirable voltage peak (a "spike"). An error of the same magnitude occurs when a significant bit wrongly has the value 0 in a code group corresponding to a high instantaneous value of the transmitted signal. (Errors of this type are not so conspicuous, however, at high signal levels.)

With DM, on the other hand, digital errors in the pulse pattern have much less effect on the decoded signal, for in this case the reception of an incorrect bit only causes the signal to make one step in the wrong direction.

A second reason why DM is less disturbed by a certain percentage of digital errors than PCM is that at a

[16] Particulars on the relation between the signal-to-noise ratio and the error probability in pulse reception will be found in: W. R. Bennett and J. R. Davey, Data transmission, McGraw-Hill, New York 1965.

particular bit rate a PCM signal has a much lower sampling rate then a DM signal. If for example we have code groups of 7 bits with an error percentage of 10%, then with PCM an average of 52% of the samples received will contain errors. With DM, where each bit is also a sample, errors are then found in only 10% of the samples, which is of course much less troublesome.

The use of non-uniform quantization gives an increase in the sensitivity to interference, both with PCM and with DM. In both cases interference may affect bits that correspond to greater steps in the signal than would have been the case with uniform coding.

Controlled companding is much better in this respect. This is because the *mean* value of the modulation index is little affected by the occurrence of a number of digital errors, particularly since the probability of receiving an erroneous bit is usually just as great as the probability of erroneously not receiving one. Where the errors are randomly distributed in time, as is the case with noise, there will therefore be little change in the control voltage and consequently the intervals between the quantizing levels will not be additionally increased by the presence of noise. Here again we find that DM has considerable advantages over PCM. With DM a sample with a modulation index greater than $\frac{1}{2}$ is not detected until at least four identical bits have been received. In periods when no such series of bits are received (quiet passages and intervals in the signal) any digital errors received have no effect on the control voltage, nor therefore on the magnitude of the quantizing steps. In such periods any interference is therefore practically inaudible. With PCM, on the other hand, every bit that wrongly causes the modulation index to exceed the value $\frac{1}{2}$ does affect the control voltage. Because of this the sensitivity to interference during intervals in the signal is much less with DM than with PCM.

Because of the lower sensitivity to interference a DM system with controlled companding for speech transmission still gives good intelligibility with digital error probabilities that are so high that intelligible transmission would be impossible with other systems. Some results of experiments that have demonstrated this are presented in *fig. 26*. These experiments were carried out with logatoms (special meaningless monosyllabic words). For DCDM transmission the percentage of logatoms understood is plotted as a function of the signal level. Curve *a* relates to a noise level in the transmission path that is low enough for all bits to be received without errors. Curves *b*, *c* and *d* were obtained at noise levels that were so high in relation to the level of the pulses that the average percentage of digital errors received was 1%, 5% and 10% respectively. Curve *e*, given for comparison, relates to uniform quantizing at a noise level low enough to permit good reception of all bits. We see that with DCDM even in the presence of 10% digital errors, the intelligibility is much better than with uniform quantizing.

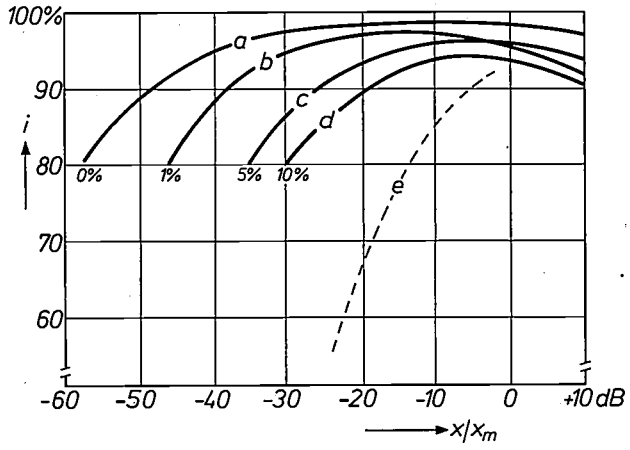


Fig. 26. Intelligibility of speech transmitted with DCDM at different interference levels and at a bit rate of 20 kb/s. The curves relate to experiments with logatoms (meaningless monosyllabic words). The graph shows the percentage of logatoms understood, i, as a function of the relative signal level x/x_m. Curve *a* relates to an interference level in the transmission path low enough for all bits to be received without errors. Curves *b*, *c* and *d* relate to higher interference levels. The interference is such that 1%, 5% and 10% of the bits are reception errors. Curve *e* relates to a system with uniform quantizing and pulse transmission without errors.

Even when the level of interference in the transmission path is so high that communication cannot be maintained with other systems DCDM can be used to give intelligible reception. This is demonstrated in track 4 on the gramophone record. DCDM is compared here with the code-modulation system nowadays most frequently used, PCM with piecewise linear coding.

Important contributions to the development of the transmission systems described in this article were made by J. W. Glasbergen of Philips Telecommunication Industries in Hilversum, and by Ir. F. de Jager, A. W. M. van den Enden and H. P. J. Boudewijns of Philips Research Laboratories, Eindhoven.

Gramophone record

The gramophone record [17] obtainable with this article presents a number of demonstrations that permit a comparison of the quality of the received and decoded signals transmitted by some of the systems discussed. The demonstrations all relate to the transmission of speech for telephony; the frequency band is therefore limited to between 200 and 3600 Hz.

The demonstrations, which are divided into four groups, are preceded by a code signal of dots and dashes.

Track 1. Demonstration of quantization noise at different signal levels with delta modulation and of the improvement obtained by using a double-integrating circuit.

Bit rate: 40 kb/s.

a — · Delta modulation with a single-integrating circuit. Signal level 0 dB [18].

b · · The signal level has been reduced to −20 dB. (The gain following the decoder has been increased by the same factor.) Note the disappearance of the weakest passages (the threshold effect).

c · · · The input level is again −20 dB. Now, however, the feedback loop of the coder incorporates a double-integrating circuit (see fig. 5). Note the improvement in quiet passages.

Track 2. The effect of companding with PCM with codewords of 5 bits.

Bit rate: again 40 kb/s.

a — — · PCM with uniform quantizing. Signal level −20 dB.

b · · PCM at the same signal level, now with non-uniform quantizing. The ratio of the maximum to the minimum intervals between the quantizing levels is 20.

c · · · Again at a signal level of −20 dB, now with digitally controlled companding. The ratio of the maximum to the minimum quantizing unit is 60.

Track 3. The effect on transmission quality of digitally controlled companding with DM.

Bit rate: 40 kb/s.

a — — — · DM with double integration; signal level −20 dB. This demonstration corresponds to 1*c*, and makes it easier to assess the next two demonstrations.

b · · Digitally controlled companding at the same signal level (DCDM). The quantizing unit q varies in magnitude by a factor of 50.

c · · · With the same variation of q the signal level can be reduced to −30 dB.

d · · · · For comparison: the same system with the coder fully modulated (0 dB).

[17] The record can be obtained by returning the form attached to the summary sheet.

[18] The 0 dB level is the one at which the coder is fully modulated; for speech and other signals of irregular waveform, a system is said to be fully modulated when this 0 dB level is exceeded 3 times in every 10 seconds.

Track 4. Comparison of PCM-NUQ with DCDM, relating to sensitivity to interfering signals in the transmission path.

To give a clearer demonstration of the effect of digital errors, particularly when they are few in number, the relatively high bit rate of 56 kb/s is now used for both systems. The signal level is 0 dB (coder fully modulated).

a — — — — · PCM with piecewise linear coding for a signal-to-transmission-path-noise ratio of 10 dB. This figure implies that an average of 0.1 % of the bits received are errors (bit-error rate 1 %).

b · · DCDM under the same conditions.

c · · · PCM with piecewise linear coding, at a signal-to-transmission-path-noise ratio of 7.5 dB. (Bit-error rate 1 %.)

d · · · · DCDM under the same conditions.

Summary. A review is first given of the principal code-modulation systems, pulse-code modulation (PCM) and delta modulation (DM). For speech transmission DM has certain advantages over PCM. The overload characteristic with DM corresponds very closely with the average speech spectrum, so that no equalization is required. With DM no code-group synchronization is necessary, and the equipment does not have to meet such difficult requirements as for PCM. Another advantage with DM is that simpler filters can be used in transmitter and receiver than with PCM. Quantization noise is given particular attention. The same signal-to-noise ratio can be achieved with DM for a lower bit rate than with PCM. In telephony the quality of the transmitted signals is adversely affected by the variations in level (dynamic range). A considerable improvement in this respect is obtained by compressing and subsequently expanding the signal (a process known as "companding"). With quantized signals the companding can be effected by non-uniform quantizing (NUQ), where the intervals between the discrete quantizing levels of the signal are not made equal. A system widely used at present is PCM with NUQ in which the compression characteristic consists of linear ranges (piecewise linear coding). Another method uses linear quantizing, but the quantizing unit varies with the mean value of the modulation index over a particular time interval. The authors have applied this principle to delta modulation, deriving the control voltages for varying the quantizing unit from the bit pattern (digitally controlled delta modulation, DCDM). With this system a much better signal-to-noise ratio can be obtained than with PCM using piecewise linear coding. A DCDM system is also much less sensitive to interfering signals in the transmission path. Even when the noise in the transmission path is so high that other systems are unusable, DCDM will maintain intelligible speech communication. Identical basic circuits are used in the transmitter and receiver for DCDM. The integrated circuits developed for this system require only two types of chip.

Digital tone generation for a transposing keyboard instrument

N.V. Franssen and C. J. van der Peet

On the 18th of November 1626 the Liège engineer Jean Gallé and the organ builder André Sévérin appeared before a notary in Liège to draw up a contract, in which Jean Gallé promised to teach Sévérin

*"la façon de faire orgues positives, régales espinettes et clavis, lesquelles par son invention se pourront haulser et abaisser s'accordantes à touts tons avec une harmonie meilleure qu'à l'ordinaire, pouvant commencer UT par tout l'octave, . . . " **

The requirement for a transposing keyboard instrument — one whose pitch can be varied — is thus quite long established. As to the precise details of Gallé's invention, we are now almost completely in the dark. The authors of the article below however have succeeded in using electronics to create a transposing keyboard instrument. The digital circuits employed in the instrument also make it possible to change the tuning of the instrument in a relatively simple way, so that the various systems of tuning that came into fashion in the course of history can be heard in rapid succession. This will certainly be of interest to institutions for musical education. These and other possibilities of the instrument built by the authors are demonstrated on a gramophone record which can be obtained with this article.

Introduction

In the traditional keyboard instruments — organ, harpsichord and piano — each key operates one or more pipes or strings, which have to be tuned to the correct pitch. Once the whole instrument has been tuned, two things have been established: the pitch of the instrument, and the system it is tuned in. Neither can easily be altered; the pitch cannot be adjusted to suit other instruments, and the system of tuning — nowadays always the scale of equal temperament — cannot be adjusted to the tuning practice of a former era for playing old music.

We have developed a digital process for generating audio frequencies that does offer these two facilities when it is applied in an electronic keyboard instrument. In this process all the audio frequencies are derived from a single oscillator, and when the frequency of this oscillator is altered the pitch of the whole instrument is changed. The oscillator frequency can be altered by just a small amount to tune up with another instrument, or by one or more semitones, so that the performer can play a piece of music in a different key from the one it is written in, without having to transpose on sight — something that few performers are really accustomed to. Moreover, the connections to the adding circuit, in which the audio frequencies arise through the summation of a number of pulse trains of different repetition rates, can easily be switched in and out. This allows an instrument to be made that can be switched between various systems of tuning. Such a facility is of interest because various systems of tuning came into fashion in the history of music, and these all had their effect on musical composition. Before going on to deal with the circuit for generating the audio frequencies, we shall now look a little more closely at the history of the musical scale.

Problems in tuning a keyboard instrument

The impression that two musical notes make on us when they are played together can either be one of consonance or of dissonance. The two notes form a

Dr. Ir. N.V. Franssen is with Philips Research Laboratories, Eindhoven. C. J. van der Peet was with Philips Research Laboratories until his death in January 1970.

* "the manner of making organs, positive organs, regals, spinets and harpsichords, which through his invention should permit of being transposed higher and lower and thus possess a better tuning than is usual, so that one may begin anywhere in the octave with DOH, . . . "

consonance if the ratio of their frequencies is the ratio of two small integers. The most perfect consonance is that between two notes of the same frequency: to the ear the two notes blend completely into one another. Again, when one of the notes has twice the frequency of the other, the two blend so well together that they are given the same letter as a musical name. The two notes can be thought of as being separated by a distance, which can be graphically represented by the spacing between the corresponding keys of the keyboard (*fig. 1*); in musical theory this is called an interval. The interval between two notes whose frequency ratio is 1 : 2 is called an *octave*. The octave on the keyboard (e.g. c to c^1 in fig. 1) is divided into twelve steps in a readily recognized pattern that is the result of historical development. This pattern repeats itself after every octave. The frequencies of the notes inside the octave are dependent on the system in which the instrument is tuned.

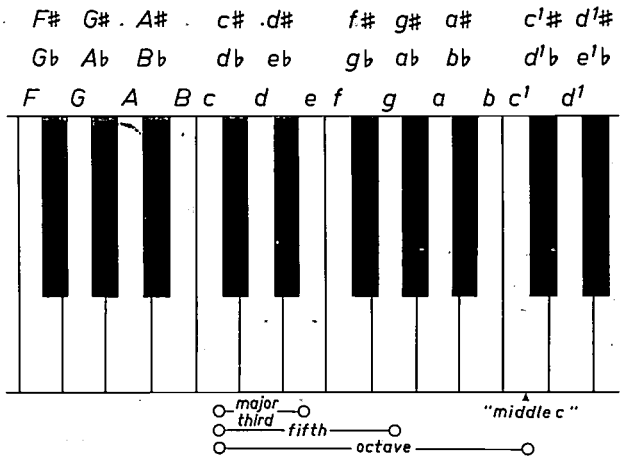


Fig. 1. Part of the piano keyboard with the names of the notes. The intervals of major third, fifth and octave from the C below middle C are indicated.

The Pythagorean scale

According to the classical writers it was the philosopher Pythagoras (6th century B.C.) who discovered that the unstopped lengths of a string vibrating at consonant intervals were related in simple ratios. His name has been given to a musical scale based upon the *fifth*. The fifth is the interval between two notes whose frequencies are in the ratio of 2 : 3, and is also a consonant interval. The fifth of the note F is C, the fifth of C is G, and so on; if we start at F and go up six fifths (see *fig. 2*) then all seven notes corresponding to the white keys of the keyboard will have been played. The notes fall in various octaves, but this does not matter, since we can assume that transpositions of one or more octaves are applied to bring them back to the same octave. If the series of fifths is then extended further upwards and downwards, then we obtain the sharps and flats (the black keys). Finally, whether going up or down, we arrive at the same black key, A flat/G sharp in fig. 2. It is found however that the G sharp in fig. 2 is slightly higher than the A flat. This can be shown directly from a calculation: the twelve fifths above the A flat represent a frequency ratio of $(\frac{3}{2})^{12} \approx 129.74$, which is slightly larger than the frequency ratio $2^7 = 128$ of seven complete octaves. The ratio 129.74 : 128 is called the *comma of Pythagoras*.

Clearly, there is a difficulty here: a keyboard instrument cannot be tuned in such a way that all the fifths have the proper simple ratio of frequencies — so that they are *true perfect fifths*, as the musician would say. One of the fifths — C sharp to A flat in fig. 2 — will have to be flattened to produce a closed system. This fifth is traditionally known as the "wolf fifth", since the dissonance of this interval put the hearer in mind of the howling of a wolf. Music in which these two notes are frequently sounded together sounds unacceptable when played on an instrument tuned in the Pythagorean scale.

The frequency ratios of the notes of the Pythagorean scale with respect to C are given in *Table I*, part 1. This and the following parts of the table also show the binary values of the ratios, which will be of interest in the discussion of the digital circuits later in the article.

The scale of just intonation

The most important consonant interval after the octave and the fifth is the *major third* with the frequency ratio 4 : 5. Major thirds are represented by the intervals such as F-A, C-E and G-B. In the Pythagorean system these intervals are built up from four intervals of a fifth (fig. 2), which means that on transposing back to the original octave they have the frequency ratio $(\frac{3}{2})^4 : 2^2 \approx 1.2656$ instead of the consonant ratio $\frac{5}{4} = 1.25$. The Pythagorean thirds consequently seem rather dissonant and "hard" in character, and the error is referred to as the *comma of Didymus*.

If it is desired to have a system in which there are true thirds, as well as true octaves and fifths, then a number of the intervals of a third that frequently occur in music can be made true. The scale that can be built up in this way is known as the scale of *just intonation*, the *natural scale*, or the *true scale*. Its thirds could for

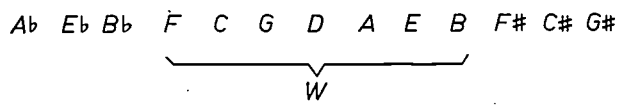


Fig. 2. All the notes of the keyboard appear in a series of twelve notes with an interval of a fifth between successive notes. *W* white keys.

example be the intervals noted above with the third D-F sharp. This is shown schematically in *fig. 3* by writing A, E, B and F sharp above F, C, G and D. The three notes in each L-shaped "box" of fig. 3, e.g. F-A-C give the "major common chord"; the notes of the chord are true and have the frequency ratios 4 : 5 : 6. The true major common chord is a perfect consonance; however, only a limited number of triads can be made true. Thus, although the common chords on F, C and G in fig. 3 can be made true, the one D-F sharp-A on D cannot. This is because the A is already determined as the major third of F and the fifth D-A is therefore a comma of Didymus smaller than the perfect fifth.

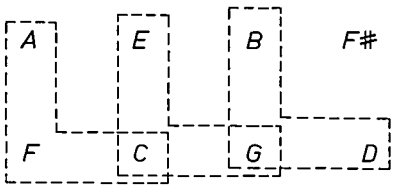


Fig. 3. When the intervals F-C, C-G and G-D are all made true fifths and the intervals F-A, C-E and G-B are all made true major thirds, then true common chords F-A-C, C-E-G and G-B-D are obtained. These are grouped in the L-shaped "boxes" of the figure.

In fig. 3 eight of the twelve notes have been established. The remaining four, all corresponding to black keys, can now be tuned as major thirds of notes already established. There are various possible ways of doing this; the G sharp could be located a major third above the E for example or it could be made a major third below the C (and then called A flat). The frequencies in the two cases are different and are in the ratio 1.5625 : 1.6000. *Fig. 4* shows one of the possible ways of completing the scale of just intonation; the five true major common chords are grouped in "boxes" as before. Part 2 of Table I shows the related frequency ratios. The seven imperfect major common chords are often very dissonant; in fact an instrument tuned in just intonation can only be played in one key. A special feature of the system of just intonation shown in fig. 4 is that the four intervals F-D sharp, C-A sharp, D flat-B and A flat-F sharp all approximate very closely to the frequency ratio 4 : 7 (the ratio is 4 : 7.03). This means that they produce an almost perfect 7th harmonic (which the musician knows as the true *diminished seventh*).

Mean-tone scale

Although the scale of just intonation does offer the advantage of a number of completely true intervals, the complete impossibility of a modulation in key limits its application considerably. The Pythagorean scale does allow key modulation to a limited extent by starting from as many true fifths as possible and making the best of the dissonant thirds. However, one can also start from true thirds and make the best of the imperfect fifths that then arise. The Pythagorean third is too large by a comma of Didymus. If this comma is divided over the four intervals of a fifth that together form a major third, then each fifth is only too small by a quarter of this comma. A system of tuning with as many true thirds as possible came into use in the 16th century; this was the system based on the mean-tone scale. In this scale the fifths were tuned slightly flat, to an extent such that the notes of the upper row in *fig. 5* formed a true third with the notes below. Neither was this system a closed one: the frequencies of the notes A flat-C-E-G sharp are in the ratios 3.2 : 4 : 5 : 6.25, so that the G sharp is clearly not the octave of the A flat. The G sharp was usually taken as the true third, above the E. Arnold Schlick (about 1500), one of the pioneers of mean-tone tuning, however recommended putting this note halfway between the G sharp and the A flat. With this scheme the consonance in thirds with the C, which could frequently occur in polyphonic music, is acceptable. On the other hand, the consonant third E-G sharp, which occurs in the closing cadence of a piece in the key of A, is not true. However, Schlick believed that a wrong third would be more acceptable here than anywhere else, and that a good organist should be able to add ornamentations and grace notes in such a way that the "harsh" G sharp is only heard in passing (*fig. 6*).

The relative frequencies for the mean-tone scale are shown in the 3rd part of Table I. The A flat is here

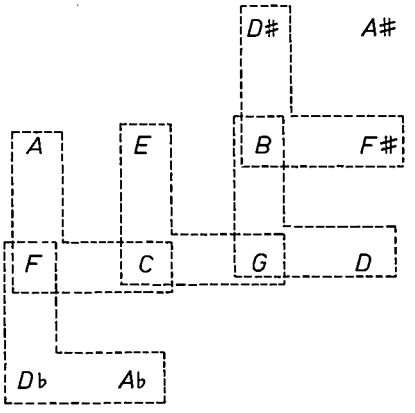


Fig. 4. One of the possible "true" systems of keyboard tuning. The interval between the notes in the horizontal rows is a true fifth, and the interval between the notes in vertical columns is a true major third. The five true common chords are grouped in the same way as in fig. 3.

made a true third below the C. The interval C sharp-A flat is much too large (frequency ratio 1.5312) and is also known as a "wolf fifth". The four intervals E-A flat, B-E flat, F sharp - B flat and C sharp - F are also all too large, with the frequency ratio 1 : 1.28.

Werckmeister's system of tuning

The flattened fifths are a disadvantage of the mean-tone scale; but its most important disadvantage is that here again the dissonant intervals make it impossible to play in certain keys. As the development of music for the organ and harpsichord progressed the need increased for a system of tuning in which music of any key could be played. There were many attempts to pro-

C G D A E B F# C# (G#)
Ab Eb Bb F C G D A E

Fig. 5. In the mean-tone scale the notes in the upper row are a true major third above the notes in the lower row. To make this possible the fifths (horizontal in the figure) are flattened slightly.

duce such a system; two of the pioneers in this field were Andreas Werckmeister (1645-1706) and Johann Philipp Kirnberger (1721-1783). Both tried to keep the thirds in some of the frequently used common chords as near true as possible. By way of example we shall look at one of Werckmeister's systems of tuning.

We saw that in the Pythagorean scale eleven true fifths are used, with the resulting error — the Pythagorean comma — located in a single fifth. In the mean-tone scale eleven flattened fifths were used and again the remaining error was located in a single fifth: In Werckmeister's system eight fifths are kept in perfect tune and the error is divided over the four remaining fifths. These fifths are then found to be about the same as the mean-tone fifths and are therefore acceptable. These flattened fifths are C-G, G-D, D-A and B-F sharp. The character of the tuning now changes with the key; in C major the tuning is almost mean-tone tuning, in F sharp major it is nearly Pythagorean. Consequently all keys are usable, but they have a different character. The relative frequencies are shown in part 4 of Table I.

The scale of equal temperament

The last step in the direction of a complete equivalence for all keys, that is to say a system of tuning in which the comma of Pythagoras is divided equally over all twelve fifths, had in theory a very early beginning. When Simon Stevin calculated this system very accurately in the early years of the 17th century and proposed its introduction he was by no means the

Table I. Ratios of the frequencies of the twelve notes of the octave to the frequency of the C.

1. Scale of Pythagoras			2. Scale of just intonation			3. Mean-tone scale			4. Werckmeister's scale			5. Scale of equal temperament		
	decimal	binary		decimal	binary		decimal	binary		decimal	binary		decimal	binary
C	2.0000	10.0000000000	C	2.0000	10.0000000000	C	2.0000	10.0000000000	C	2.0000	10.0000000000	C	2.0000	10.0000000000
B	1.8984	1.1110011000	B	1.8750	1.1110000000	B	1.8692	1.1101111010	B	1.8792	1.1110000100	B	1.8878	1.1110001101
B flat	1.7778	1.1100011100	A sharp	1.7578	1.1100001000	B flat	1.7889	1.1100101000	A sharp	1.7778	1.1100011100	A sharp	1.7818	1.1100100001
A	1.6875	1.1011000000	A	1.6667	1.1010101011	A	1.6719	1.1010110000	A	1.6704	1.1010101111	A	1.6818	1.1010111010
A flat	1.5802	1.1000111001	A flat	1.6000	1.1001100110	A flat	1.5000	1.1001100110	G sharp	1.5803	1.1001010010	G sharp	1.5874	1.1001011010
G	1.5000	1.1000000000	G	1.5000	1.1000000000	G	1.4954	1.0111111011	G	1.4949	1.0111111011	G	1.4983	1.0111111110
F sharp	1.4238	1.0110110010	F sharp	1.4063	1.0110100000	F sharp	1.3976	1.0110010111	F sharp	1.4047	1.0110011110	F sharp	1.4142	1.0110101000
F	1.3333	1.0101010101	F	1.3333	1.0101010101	F	1.3375	1.0101011010	F	1.3333	1.0101010101	F	1.3348	1.0101010111
E	1.2656	1.0100010000	E	1.2500	1.0100000000	E	1.2500	1.0100000000	E	1.2528	1.0100000011	E	1.2599	1.0100001010
E flat	1.1852	1.0010101101	D sharp	1.1719	1.0010110000	E flat	1.1963	1.0011001001	D sharp	1.1852	1.0010111110	D sharp	1.1892	1.0011000010
D	1.1250	1.0010000000	D	1.1250	1.0010000000	D	1.1181	1.0001111001	D	1.1174	1.0001111000	D	1.1225	1.0001111101
C sharp	1.0679	1.0001000110	D flat	1.0667	1.0001000100	C sharp	1.0450	1.0000101110	C sharp	1.0535	1.0000110111	C sharp	1.0595	1.0000111101
C	1.0000	1.0000000000	C	1.0000	1.0000000000	C	1.0000	1.0000000000	C	1.0000	1.0000000000	C	1.0000	1.0000000000

first to do so; but it was not until the middle of the 18th century that it began to come into use [1]. Nowadays this scale of equal temperament is in general use.

Fig. 6. Title page of the book "Spiegel der Orgelmacher und Organisten" by Arnold Schlick, organist to the Counts Palatinate, published at Speyer in 1511. Schlick advocated a mean-tone system in which the G sharp was not perfectly tuned with respect to E, but was tuned sufficiently sharp to give an acceptable consonance with the C. While the third E-G sharp does occur in a closing cadence on A — with the G sharp usually in the upper part — Schlick believed that a wrong third was more acceptable here than elsewhere. In his own words:

"It is however more to be tolerated at this place than at any other since it is a close and it is not needful that the G sharp of the descant should be held for as long as the other voices: instead in such a close the descant may be disguised or hidden with a little pause at its start or a sequence of short notes, a grace, run, shake or ornament — name it as thou wilt — so that the harshness of the close shall not be noticed, in such manner as a skilled organist knows."

(The title page is reproduced from the facsimile edition prepared by P. Smets, with the kind permission of the publishers, Rheingold Verlag, Mainz 1959.)

The system has been called the "system of proportional beats" (by A. D. Fokker) because all the intervals of a fifth have the same relative deviation in frequency, audible as slow beats in the consonance: the number of beats per second is proportional to the frequency of the notes that comprise the fifth. In the octave of the equally tempered scale all the twelve semitones are equal, each therefore corresponding to the frequency ratio $2^{1/12}$.

In the scale of equal temperament the only interval that is true is the octave; every other interval is imper-

fect in some way. The major third is fairly poor and lies halfway between the perfect major third and the Pythagorean third; this is really the main reason why the equally tempered scale did not gain earlier acceptance. The fifths do not differ greatly from the true fifth; the error is no more than a twelfth of the Pythagorean comma. Tuning an instrument to the system of equal temperament requires considerable training, for all the fifths have to be flattened by exactly the same small amount. The relative frequencies for the scale of equal temperament are shown in part 5 of Table I.

Tone generation by digital methods

The tone generation in electronic keyboard instruments usually takes place in a system of twelve oscillators, which generate the twelve tones for the notes of the highest octave. The tones for the lower octaves are derived from these tones by frequency dividers. The twelve master oscillators can be tuned independently of one another.

In such instruments there are no more than twelve oscillators that need tuning. The octave ratios are fixed once and for all time in the electronic circuits.

With digital circuits, the ratios between the frequencies of the twelve notes of the octave can also be fixed by the design of the circuit. A difficulty, however, is that the ratios corresponding to the equal-tempered scale are irrational ones. This means that they cannot be made exact, but can only be approximated. The approximation has to be extremely close to obtain a result acceptable to the ear; in our opinion it should be no more than 0.05% of the true frequency.

The reason for this is that the interval of the fifth in the equal-tempered scale is only 0.1% smaller than the true fifth (the ratio is 1.4983 instead of 1.5000). With a frequency error greater than 0.05%, two notes forming a fifth can both be wrong by 0.05% in different directions in such a way that the interval becomes equal to a true fifth or larger. Then not all of the fifths are smaller than the true fifth, and this degrades the character of the tuning.

To achieve such an accurate approximation a good deal of digital equipment is required and until recently the cost and size have been prohibitive. This situation is now changing, however, with the increasing variety of digital circuits now available.

A number of proposals for digital tone generation have consequently been put forward. Most designers approximate the twelve tone frequencies by dividing a single high master frequency by twelve carefully chosen large integers (between about 200 and 300) [2] [3]. In this way a train of pulses is obtained with a number of pulses per second that approximates to the required frequency. Our approach, however, has been rather different. Encouraged by related researches performed elsewhere [4], we took the apparently drastic step of dropping the requirement for the absolute periodicity of the pulses. Provided that the timing error for the separate pulses does not exceed a few per cent of their average spacing, trains of pulses that are not absolutely periodic are heard as a single tone. The pitch corresponds to the average repetition rate and the aperiodicity can only be sensed through the presence of a small noise-like signal that gives a certain husky quality to the note. If the aperiodicity is limited to a few tenths of one per cent this huskiness is barely perceptible.

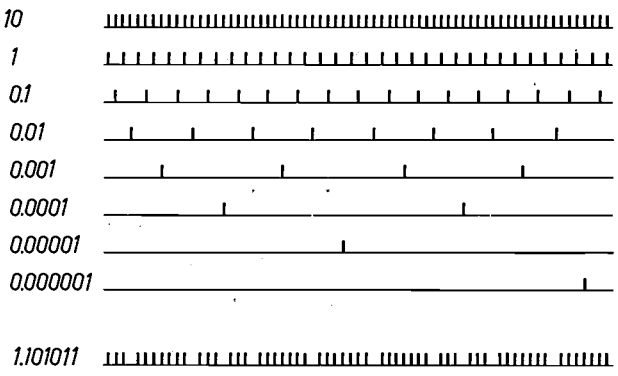


Fig. 7. The trains of pulses used for generating the tone frequencies, with pulses that do not coincide. The repetition rates are given in the binary system, with the frequency of the master oscillator set at 10 (2 in decimal notation). An example of the addition of a number of pulse trains can be seen at the bottom; the number of pulses per second is shown next to it in binary notation.

This tolerance of the human ear allows us to assemble the required number of pulses per second by adding trains of pulses at differing repetition rates. The pulse trains are derived from a master oscillator by a progressive frequency division in which each train of pulses has half the repetition rate of the one that preceded it (*fig. 7*). For simplicity we shall now express all frequencies in the binary system, normalizing half the master frequency to 1. The master frequency, which is its octave, then has the frequency 10 in binary notation and all twelve notes inside the octave are now given by a binary number between 1 and 10 (see Table I).

[1] The view that J. S. Bach published his "Wohltemperirtes Clavier", with its 48 preludes and fugues in all keys, to promote the use of the equal-tempered scale is now questioned by some authorities. In their view, Bach was attempting to gain support for a system proposed by Werckmeister. The question is discussed in: A. D. Fokker, De muzikale ontwikkeling op een tweesprong, Ned. Akoest. Genootschap Publ. No. 7, 3-7, 1965.

[2] D. Gossel, Generation of musical intervals by a digital method, Philips tech. Rev. **26**, 170-176, 1965.

[3] F. B. Maynard, Sr., Designing a digital organ tone generator, Motorola Application Note AN-424.

[4] The investigation we are referring to was carried out at the Institute for Perception Research, Eindhoven, and has been described in the article by B. Lopes Cardozo and R. J. Ritsma, On the perception of imperfect periodicity, IEEE Trans. AU-16, 159-164, 1968. See also I. Pollack, Discrimination of mean temporal interval within jittered auditory pulse trains, J. Acoust. Soc. Amer. **43**, 1107-1112, 1968.

The pulse trains obtained by further halving correspond successively with the binary numbers 0.1, 0.01, 0.001, etc. and provide the material for making up the binary numbers between 1 and 10 to the desired accuracy. If the frequency 1 is made to correspond to the note C, then the A, for example, of the equal-tempered scale is given by 1.1010111010 (see Table I, part 5), and is formed by adding the following pulse trains (see fig. 7):

```
1
0.1
0.001
0.00001
0.000001
0.0000001
0.000000001
————————— +
1.101011101
```

The other notes are formed in a similar way.

The number of binary digits required is determined by the degree of accuracy to which the frequencies produced by this binary synthesis have to match the frequencies required by the system of tuning. Our requirement that no deviation should be greater than 0.05% = 1/2000 indicates an accuracy of $1/2^{11} = 1/2048$ for the binary frequencies and hence eleven-digit binary numbers. This means that eleven frequency dividers are needed for the synthesis of the tone frequencies.

The aperiodicity of the pulse trains obtained is far too large to allow them to be used directly as tones for the notes of the instrument. We mentioned earlier a permissible aperiodicity of a few tenths of one per cent; the timing error Δt in the summation pulse trains (*fig. 8*) can however be nearly equal to one average period. If the summation pulse train of fig. 7 is halved, the pulses are used alternately; the situation is then improved by a factor of two, for the timing error remains the same while the average period becomes twice as long. This procedure is used to reduce the aperiodicity — the whole summation process is carried out at high frequencies and the sum frequencies obtained are divided by a large number (in the version to be described later, with a master frequency of 2 MHz, this large number is $2^9 = 512$) to obtain the pulse train for the highest octave of the instrument. The lower octaves are then obtained by further halving; the pulse trains give the raw tone, which is then made acceptable to the ear by means of tone-colour filters and other processing circuits. The pulse trains shown in fig. 7 have the special feature that the pulses in different rows never occur at the same time. This is an essential feature for the operation of our system, and cannot be obtained by using divide-by-two circuits alone. We have made use of logic circuits for this part of the process.

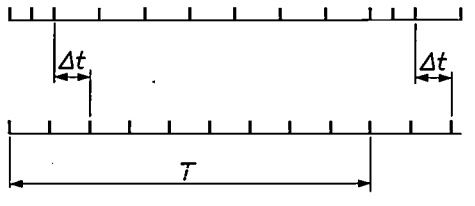


Fig. 8. For a situation in which the ideal regular pulse train has nine pulses in a time T (*below*), one of the pulses will have a timing error Δt nearly equal to the mean period when the nine pulses are produced by the binary synthesis 8 + 1 (*above*). The number 9 is just an example: this holds generally for all numbers $2n + 1$ and the effect becomes worse as n increases.

Circuits for deriving the pulse trains

In discussing a circuit that derives trains of non-coincident pulses from the master frequency we shall make use of *fig. 9*. In this figure F_0 is the symmetrical square-wave signal generated by the master oscillator. This signal is successively divided by two, using standard circuits, to give the symmetrical signals $F_1, F_2, \ldots, F_{11}$. Besides these signals the inverted signals $\bar{F}_0, \bar{F}_1, \bar{F}_2, \ldots, \bar{F}_{11}$ are also required. If, by using a NOR gate with two inputs, we now form the signal $I_1 = \overline{F_0 + \bar{F}_1}$, i.e. a signal that corresponds to a logical "0" when at least one of the input signals F_0 or $\bar{F}_1$ is "1", then this signal has the pulse width of F_0 but the repetition rate of $\bar{F}_1$. For the signal I_2, whose repetition rate should be half that of I_1, $\bar{F}_2$ is necessary to determine the periodicity and F_1 is required to prevent the pulses from coinciding with those of I_1; I_2 is produced by means of a NOR gate with three inputs: $I_2 = \overline{F_0 + F_1 + \bar{F}_2}$. The process is continued in this way to the last division; the last NOR gate (with 12 inputs) gives the pulse train $I_{11} = \overline{F_0 + F_1 + F_2 + \ldots + F_{10} + \bar{F}_{11}}$. In fig. 9 I_1, I_2 and I_3 are shown as examples, and so is the sum $\overline{I_1 + I_2 + I_3}$, also obtained with a NOR gate, which corresponds to the binary addition $1 + 0.1 + 0.01 = 1.11$ $(1 + 0.5 + 0.25 = 1.75$ in decimal notation).

The required number of pulses per second can be obtained not only by adding trains of pulses but also by suppressing a number of the pulses from the master oscillator, i.e. by subtraction instead of addition. With this method the wiring can be rather simpler, and we have therefore chosen it for our experimental version of a digital tone-generation system. Pulse trains that are non-coincident are again required with this approach, now for the synthesis of the binary number that has to be subtracted from the binary 10 of the primary frequency to obtain the desired tone frequency. The frequency of the note A for example, is shown by part 5 of

Table I to be equal to $1.1010111010 = 10 - 0.0101000110$. The pulses in the train can now have double width; some of these pulse trains, I_2' and I_3' are shown in fig. 9; these are also formed by using a NOR gate. The signal $F_0 + I_3'$ is shown as an example; this takes the same form as $\overline{I_1 + I_2 + I_3}$ but is produced here by subtraction. The binary subtraction is $10 - 0.01 = 1.11$ ($2 - 0.25 = 1.75$ in decimal notation). The signal $F_0 + I_3'$ is obtained with the aid of an OR gate. The output of an OR gate is "1" when at least one of the inputs is "1".

A disadvantage of the addition or subtraction method described here is that it requires NOR circuits with up to twelve inputs for making the I pulse trains. This can be countered by generating a subsignal that effectively contains all the information from the preceding signals at various points on the divider chain. These subsignals then serve as a starting point for the following stages. Here, however the unavoidable delays introduced by dividers and other logic circuits are something of a nuisance. The last of the sections into which the divider chain is now split is dependent for its time reference on a signal that has acquired a certain delay in its passage through the previous sections. It can be shown from fig. 9 that this delay should be kept small enough to prevent the pulses of the partial pulse trains I_{11} or I_{11}' from being delayed by more than half a period of the master signal F_0, which in our case means that it should not be delayed by more than 250 ns with respect to F_0. The advantage of NOR gates with fewer inputs therefore has to be weighed against the disadvantage that the circuits must be very fast-acting.

In the simple addition or subtraction method described in the previous paragraphs the delays do not play any part, since each signal I or I' is derived directly from the pulses of F_0; the signals $F_1, F_2, \ldots$ merely create the windows w (fig. 9) through which the correct pulses of F_0 appear, and these windows have sufficient "play" to cover any delays in the F pulse trains.

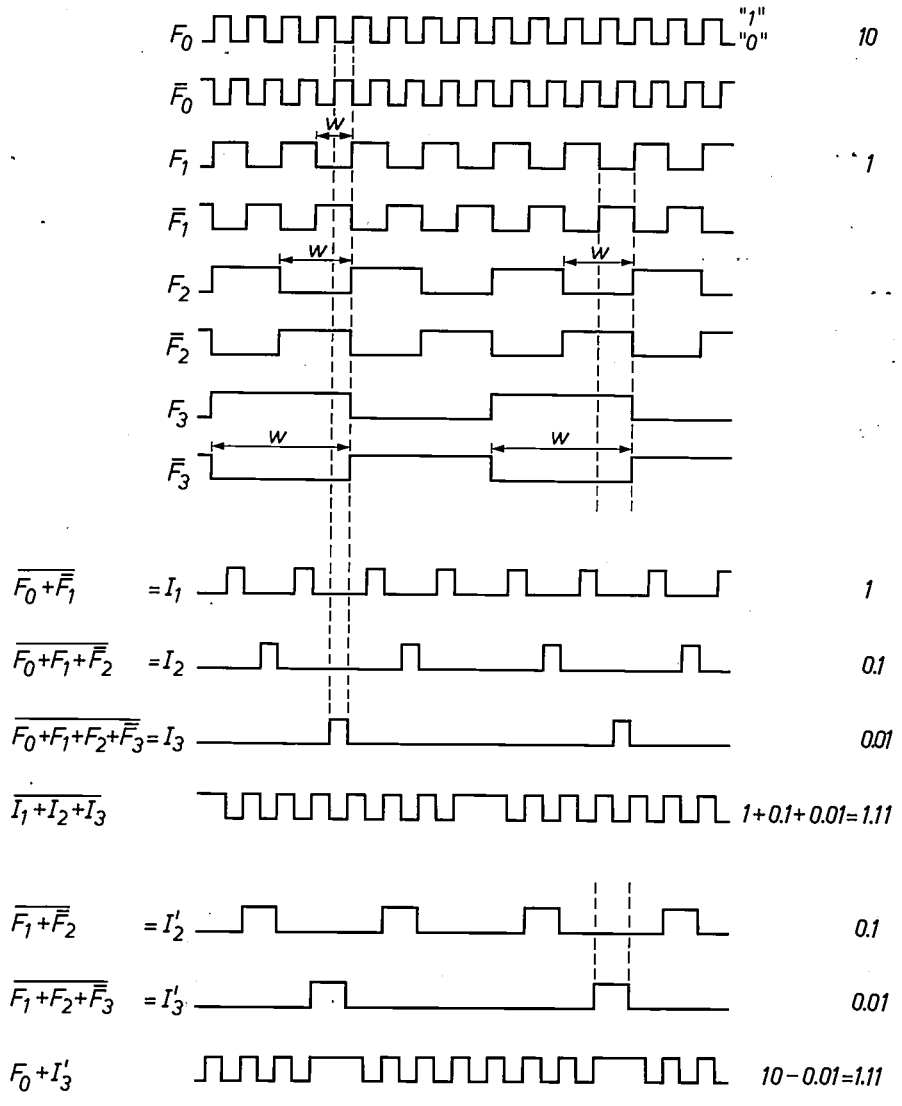


Fig. 9. The various signals for digital tone generation.

F_0 symmetrical square-wave master oscillator.

$F_1, F_2, \ldots$ output signals of successive divider stages.

$\bar{F}_0, \bar{F}_1, \bar{F}_2, \ldots$ inverted signals.

$I_1, I_2, \ldots$ trains of non-coincident pulses, which are added together to obtain pulse trains with the required number of pulses per second.

$I_2', I_3', \ldots$ trains of non-coincident pulses for generating the required number of pulses per second by suppressing the pulses from the master oscillator, i.e. by subtraction. The pulse trains F form windows w through which the pulse trains I can now and again see a half-period of F_0, or the pulse trains I' can now and again see a half-period of F_1.

The signal $\overline{I_1 + I_2 + I_3}$ is an example of an addition, and the signal $F_0 + I_3'$ is an example of a subtraction; both give the same result. The repetition rates of the signals are shown at the right in binary notation, with the frequency F_1 made equal to 1.

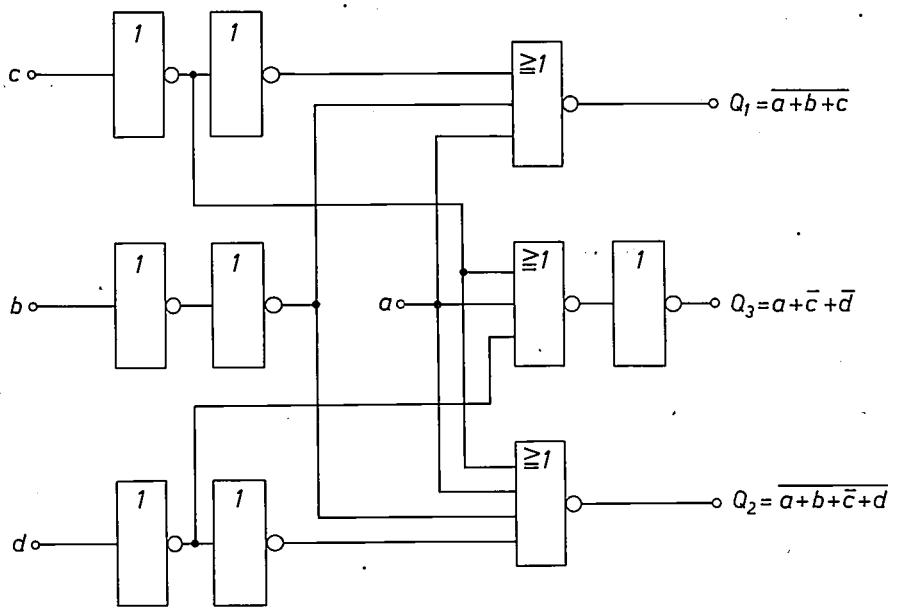


Fig. 10. Schematic diagram of the logic of circuit A, one of the two integrated circuits specially developed for the instrument. The truth table is shown on the right.

In the development of the integrated circuits for the digital tone generators it was found that the advantage of fewer inputs to the NOR gates was an important one, and the required speed of action of the circuits was achieved.

Integrated circuits

We have succeeded in making the complete circuit for generating the tones, apart from the master oscillator, from only two types of integrated circuit, which we developed specially for the purpose. Both types are made up entirely from NOR-gate circuits, some of which have only one input and serve as polarity inverters. One of the two types of circuit, which we shall call circuit A, is shown schematically in *fig. 10*. Five circuit A modules are required for producing the asymmetrical signals I from the symmetrical signals F obtained by halving the master signal. A subsignal is used here to reduce the number of inputs for I signals of higher order.

The other integrated circuit, circuit B, is used in two versions, which differ only in the external connections. Circuit B contains an OR gate with eight inputs and a chain of eight divide-by-two circuits; the OR gate and the first divider are shown schematically in *fig. 11*. Circuit B is used both for the divide-by-two circuits for the signals F and for the connecting chain of divide-by-two circuits that give the tone frequencies, giving a total of 24 units in the complete instrument. *Fig. 12* shows a block diagram of the complete system built up from integrated circuits A and B; as noted above the operation depends on subtraction of pulses. The system can

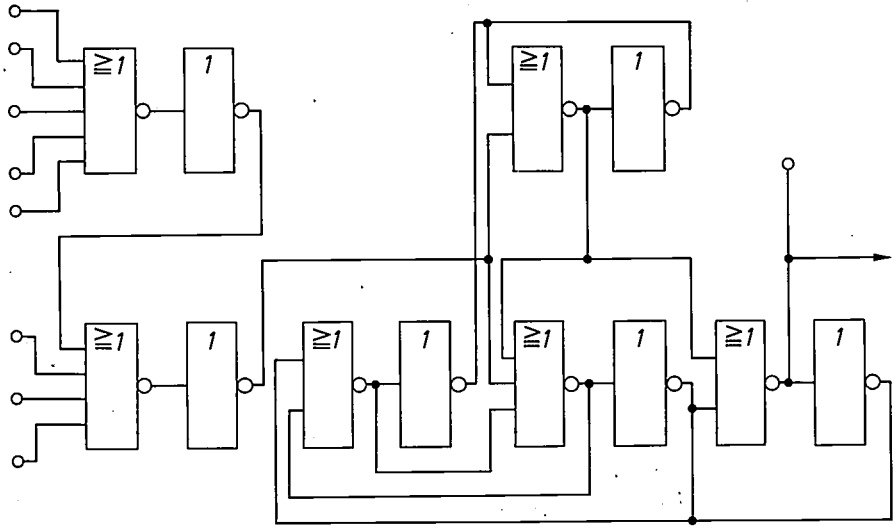


Fig. 11. Schematic diagram of the logic of circuit B. This contains an OR circuit for eight inputs (*left*) and a chain of eight divide-by-two circuits, of which only the first one is shown (*right*). There are two versions, which differ only in the final assembly; in version B_1 there are external connections to all eight inputs of the OR circuit but not to all of the divider outputs, in version B_2 there are external connections to only two of the OR-circuit inputs but to all of the divider outputs.

be produced on a single printed-wiring board (*fig. 13*).

Both integrated circuits operate with bipolar transistors. The logic circuits combine short switching times and low energy consumption. The delay times are 20 to 30 ns per logic function or divide-by-two circuit; the dissipated power is 14 mW for circuit A and 55 mW for circuit B; the voltage swing between logic levels "0" and "1" is about 200 mV [5].

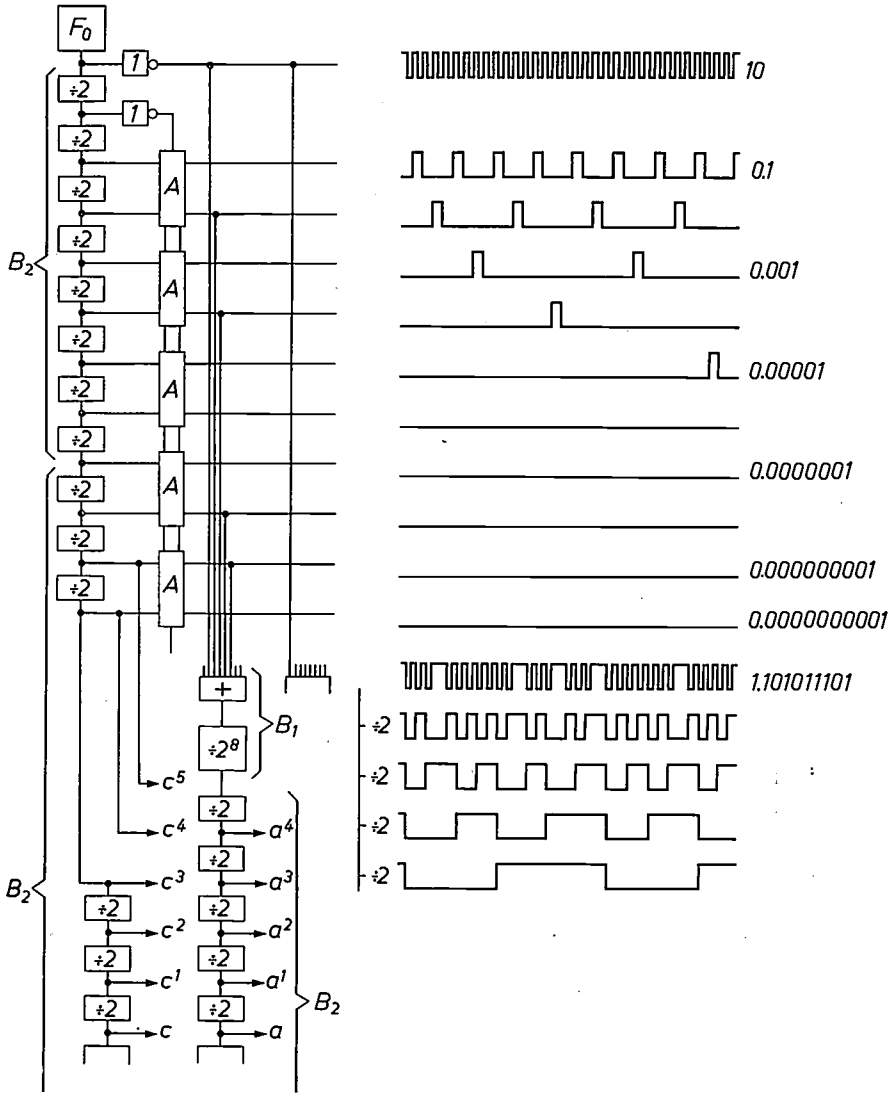


Fig. 12. Block diagram of digital tone generation with integrated circuits A and B. The circuit operates by pulse subtraction. The production of the note A is shown by way of example; the subtraction takes place in a B_1 circuit and the pulse train obtained is divided by 2^8. Further division-by-two in a B_2 circuit gives the As for the various octaves of the instrument. Every note is produced in this way with the aid of two B circuits, with the exception of the various Cs, which are produced by direct division-by-two from the master oscillator.

Further possibilities

In the introduction it was pointed out that the complete instrument could be detuned by detuning the master oscillator, and that this permitted the pitch to be altered to tune up with other instruments of an ensemble or to give an easy way of transposing music to a different key from the one it is written in. Glissandi can also be produced, like those encountered in the "Hawaiian" style of light music.

If an extra divide-by-two circuit is periodically switched in and out after the master oscillator, at say several times a second, the pitch of the instrument jumps an octave several times a second. We call this

[5] A. Slob, Fast logic circuits with low energy consumption, Philips tech. Rev. 29, 363-367, 1968.

effect the "*octave tremolo*"; it sounds rather like the *tremolo* of a street organ and is also of interest for light music. It can therefore be seen that digital methods of tone generation offer various facilities that conventional instruments do not have and which arise quite naturally from the use of electronics.

tions per second by international agreement since 1939, it was 870 between 1885 and 1939, and there were various fluctuations in concert pitch in the preceding centuries. It is therefore desirable that a device that can be used for tuning to a historical system should also be tunable to the appropriate standard pitch.

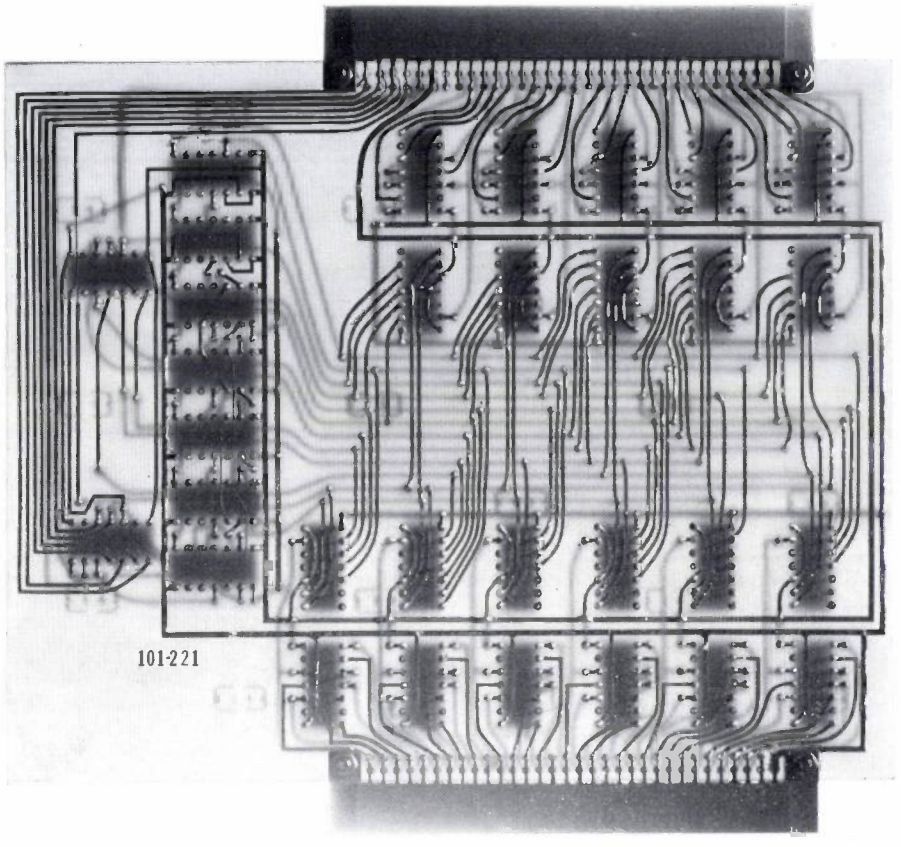


Fig. 13. The complete circuit for generating the tone frequencies on a single panel with printed wiring on both sides. The field of lines that carry the master signal and the partial pulse trains appears centrally (see the block diagram in fig. 12; there are twelve lines here, since two are provided for the master signal). Above and below this field can be seen the eleven B_1 circuits in which the pulse trains for the notes C sharp to B are synthesized from the master signal and the partial pulse trains.

Finally, we should like to point out that the electronic tone generator is not only of use for instruments but can also be used for an electronic device for tuning musical instruments. In this device the partial pulse trains *I* are connected via switches to the inputs of a gate circuit. A table indicates which switches should be closed for a particular note and in this way any note in the octave can be produced for whatever kind of musical scale is desired. These notes can be compared by ear or electronically [2] with the instrument to be tuned. For this application the master oscillator should have a fixed frequency or, preferably, it should be possible to switch it between several standards of concert pitch. Although a^2 has been standardized at 880 vibra-

Gramophone record

A gramophone record [6] is available to give a better illustration of various points that have been discussed here in connection with historical systems of tuning and with the particular capabilities of this system of digital tone generation. Each example is preceded by a code signal of dots and dashes.

Track 1. Consonances in different kinds of scale

The true fifth C sharp - G sharp alternating with the imperfect "wolf" fifth C sharp - A flat in the Pythagorean scale. The difference in pitch between the G sharp and the A flat is the Pythagorean comma.

The Pythagorean third C-E, followed by the true third C-E. The Pythagorean third is larger than the true third by the "comma of Didymus".

[6] A request coupon is attached to the summary sheet.

... The common chord C-E-G in just intonation. The tetrad C-E-G-A sharp in just intonation, followed by the same chord in the modern equal-tempered scale. In the true tuning used here the A sharp is almost equal to the true diminished seventh of the C, and therefore the tetrad in this system is an almost perfect consonance (frequency ratio 4 : 5 : 6 : 7).

.... The common chord C-E-G in just intonation, followed by the same chord in the mean-tone scale. In this scale the third C-E is true, but the fifth C-G is too small.
Closing cadence on A in the mean-tone scale, first in its basic form, and next played with an ornamented upper part to hide the dissonant third E-G sharp (*fig. 14*).

Track 2. A musical example

The first line of the chorale "Herzlich tut mich verlangen" in a setting by Johann Pachelbel (1653-1706).

— With the keynote C, successively in the scale of just intonation, the mean-tone scale, Werckmeister's system as described in the article, and the scale of equal temperament.

—·· With the keynote C sharp, successively in the same four kinds of scale. The appearance of a number of very impure concords in the just-intonation and mean-tone scales shows that music of all keys cannot be played in these systems.

Track 3. Technical capabilities

——· Continuous detuning of the pitch.
Glissando in a piece of "Hawaiian" music.

——·· Octave tremolo.

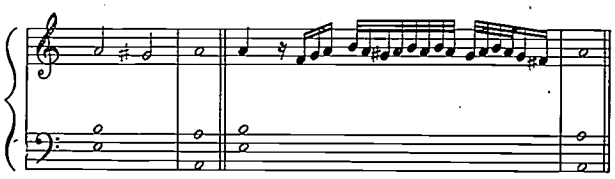
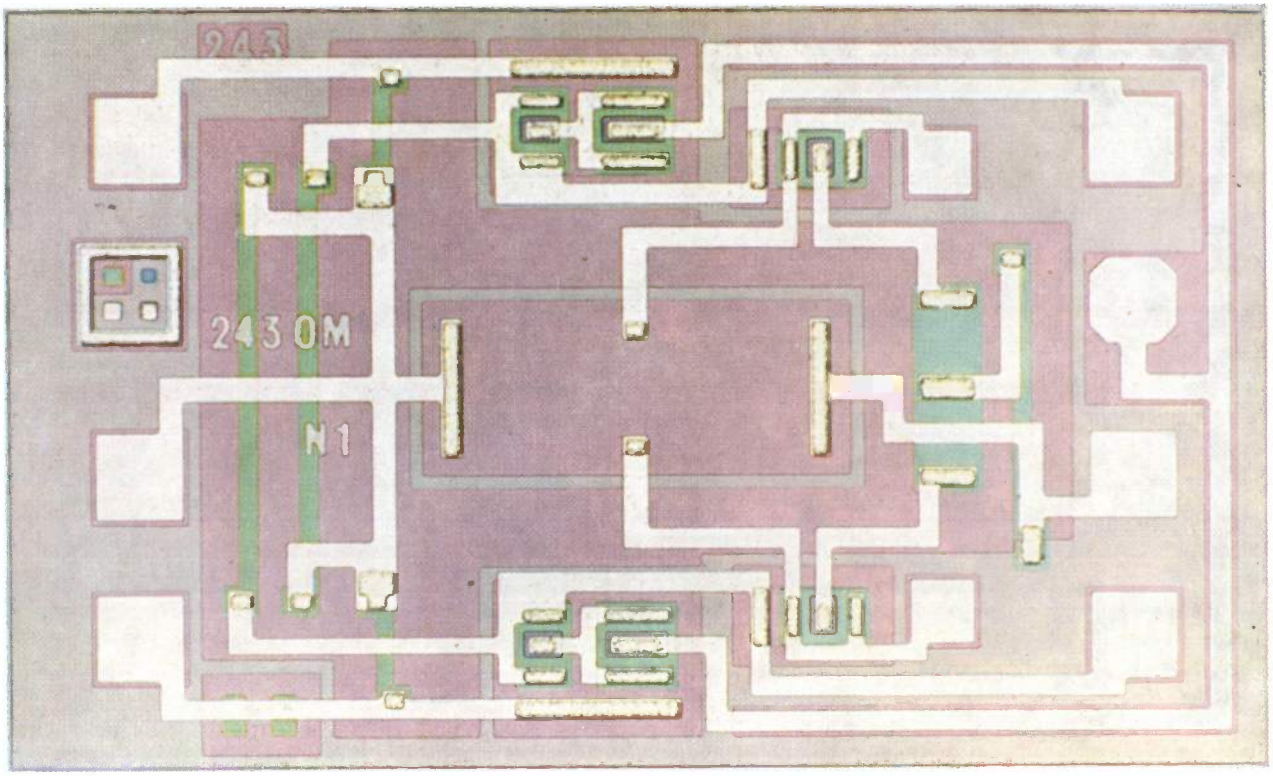


Fig. 14. The closing cadence in A in mean-tone tuning as on the gramophone record. The basic version is shown on the left; in the version shown on the right the upper part is ornamented in such a way that the dissonant third E-G sharp is only heard in passing.

Summary. The twelve tone frequencies in an electronic keyboard instrument are generated from a set of eleven pulse trains in which each pulse train has half the repetition rate of the previous one. Pulse trains are chosen and added together to give the desired number of pulses per second. The frequencies are then accurate to 1 part in $2^{11} \approx 0.05\%$. The pulse trains are derived from a master oscillator at a frequency of about 2 MHz by logic circuits; pulses from different trains never coincide. It is also possible to suppress a number of pulses from the master oscillator in such a way that the correct number of pulses per second remain. The resulting pulse trains contain the desired numbers of pulses but not in strict periodicity; dividing by 2^9 reduces the aperiodicity to below the level of perception and provides the pulse frequencies for the highest octave of the instrument. When the master oscillator is detuned the pitch of the whole instrument is altered: this can be employed to give glissandi and electronic key transposition. The instrument can be tuned to different kinds of scale by switching between different summations of pulse trains.



Integrated circuit with Hall device for brushless d.c. motors

Magnetic field strengths are often measured with Hall devices made from the semiconducting materials indium arsenide or indium antimonide. We have now made Hall devices from silicon, using the usual technology for producing integrated circuits with bipolar transistors [1] [2]. Although a silicon Hall device has a lower sensitivity than devices made with the materials mentioned above, it offers the great advantage that it can be manufactured by an existing method of quantity production. In any case the loss of sensitivity can be amply compensated by incorporating the Hall device in an integrated circuit which amplifies the Hall voltage. A further advantage of using silicon technology is that extremely small devices can be made (with a sensitive area down to 10×10 μm).

We have developed Hall devices with amplifier for use as magnetic sensors in a brushless d.c. motor. In small d.c. motors of the conventional type the stator is generally a permanent magnet which provides the required magnetic flux, and the rotor is fitted with coils to which the current is supplied through the brushes and the rotating commutator. The disadvantages of this type of motor are commutator wear, undesirable noise (particularly undesirable in sound recording and reproduction equipment) and high-frequency inter-

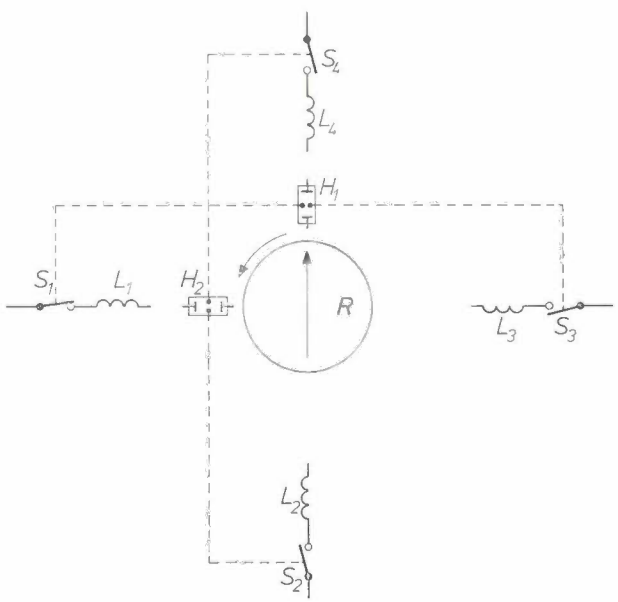


Fig. 1. Controlling a brushless d.c. motor by means of Hall devices. When the north pole of the permanent magnetic rotor R is close to the Hall device H_1, one of the electrodes of the Hall device supplies a positive voltage which keeps the switch S_1 closed (in the actual circuit, switches S_1 to S_4 are transistors). The stator coil L_1 is then energized. A quarter of a revolution later, coil L_2 is energized by Hall device H_2, and a further quarter of a revolution later coil L_3 is energized by Hall device H_1, which is now close to the south pole of the rotor.

ference. These disadvantages can be avoided by using a permanent magnet for the rotor, which is diametrically magnetized, and by incorporating the coils in the stator. They can then be supplied with current at the right moments by means of an electronic switching system. A device is then needed, however, for sensing the position of the rotor. Hall devices are particularly suitable for this purpose, and have already been used for this application [3]. A possible basic circuit is shown in *fig. 1*.

Usually the Hall devices are mounted in recesses cut into the stator. We have avoided the necessity for such recesses by mounting the devices at the head of the rotor.

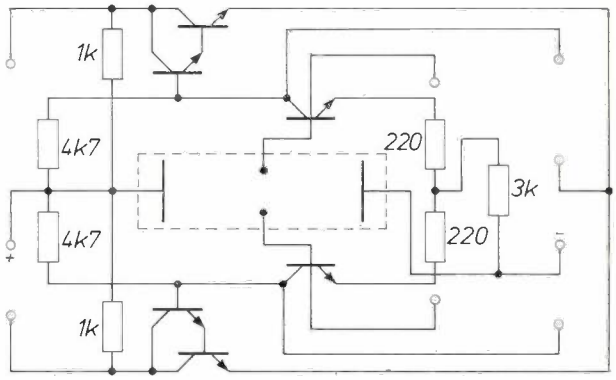


Fig. 2. Diagram of the Hall circuit, drawn to correspond with the pattern of the integrated circuit. The value of the resistors is given in ohms.

Fig. 3. The integrated Hall circuit in a plastic encapsulation. Tapered recesses at both ends contain pole pieces for increasing the magnetic flux density in the device. In the motor that we have developed this is about 0.6 T (1 T = 10^4 gauss). The devices are made in strip form and later punched out.

The Hall device which we have developed for this application is shown in the title photograph; the dimensions are about 1.5×1 mm. There are six *NPN* transistors and seven resistors in the circuit. The actual Hall device in the middle is surrounded by a clearly visible P^+ isolation diffusion. Two of these chips switch four discrete transistors which energize the stator coils of the motor. The diagram of the Hall circuit is given in *fig. 2*.

The integrated Hall devices are used in d.c. motors intended for semi-professional tape recorders. The devices are contained in a plastic standard encapsulation which in this case, however, has tapered recesses on both sides into which pole pieces are fitted to increase the flux density (*fig. 3*). A cut-away photograph of the motor is shown in *fig. 4*; the two Hall devices (see the arrows) are mounted on a printed wiring board together with their output leads.

Another possible application of a Hall device with integrated amplifier is a clip-on probe; other applications might be found in cases where a mechanical displacement has to be converted into an electrical signal, such as the keyboard of an electric typewriter [4], or a tachometer.

In addition to the Hall circuit shown in the title photograph, a small Hall probe made from silicon has been developed (*fig. 5*) for magnetic-field measurements at locations where access is difficult. This Hall

[1] G. Bosch, A Hall device in an integrated circuit, Solid-State Electronics **11**, 712-714, 1968.
[2] A. Schmitz, Solid circuits, Philips tech. Rev. **27**, 192-199, 1966.
[3] W. Dittrich and E. Rainer, Siemens-Z. **40**, 690, 1966. Another method of position sensing, using ferrite cores, was described by W. Radziwill in Philips tech. Rev. **30**, 7, 1969.
[4] R. H. Cushman (Ed.), Hall effect put in IC, Electronic Design News, 11 Nov. 1968, p. 87.

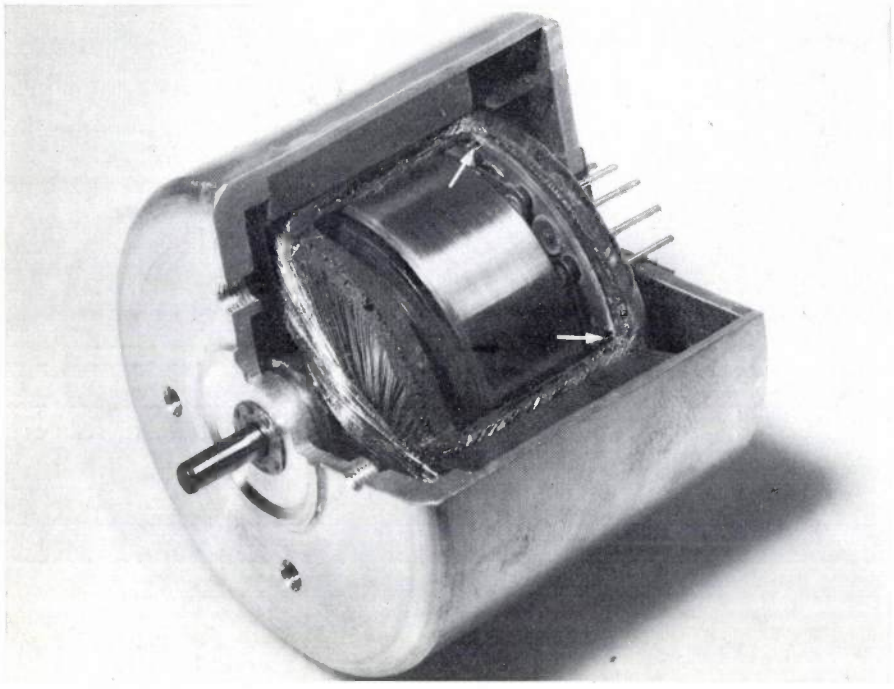


Fig. 4. Cut-away view of our "Hall motor". The Hall devices (white arrows) are mounted on a printed wiring board, to which the lead-out pins are also attached. The motor takes about 7 watts at a speed of 2000 rev/min. Speed control is effected by a transistor outside the driving circuit.

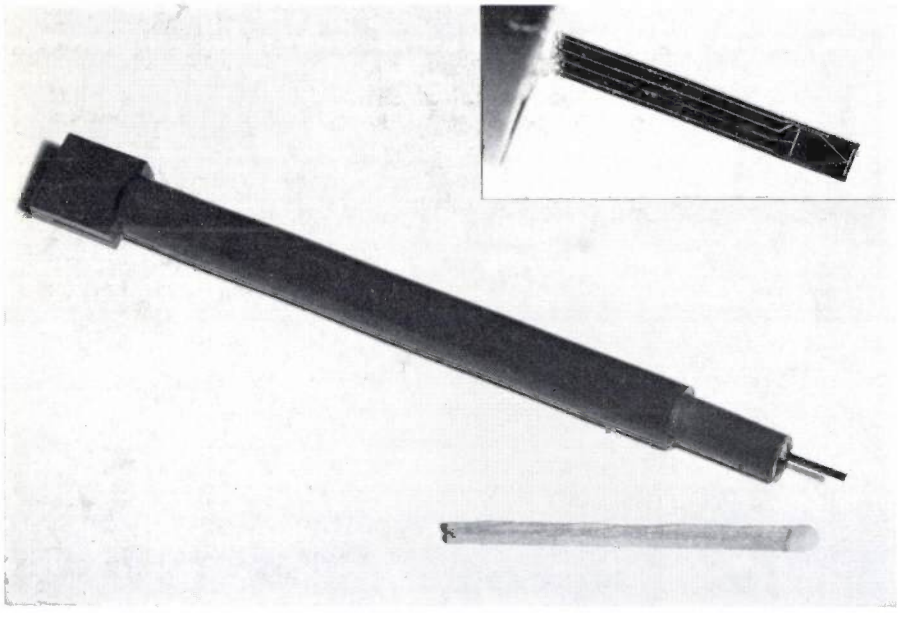


Fig. 5. Hall probe. The rectangular Hall device (1.5×1 mm) can be seen at the end of the silicon chip.

probe is not combined with an amplifier. Apart from the advantages of silicon Hall devices mentioned earlier, such as small dimensions, this probe has the additional advantage that for the small thickness required in this case (0.1 mm) a silicon single crystal is not so fragile as the insulating substrate (e.g. of aluminium oxide) which a Hall device of indium antimonide requires. Moreover the silicon device is linear up to higher field strengths.

The dimensions of the silicon chip are $12\times1.5\times0.1$ mm; the actual Hall device on the chip (dimensions 1.5×1 mm) is surrounded by a P^+ isolation diffusion in the same way as the "islands" in an integrated circuit. The contacts, like the collector contacts in an integrated circuit, are N^+. The sensitivity at a supply voltage of 10 V is about 0.75 V/T (1T (tesla) = 10^4 gauss).

G. Bosch
J. H. H. Janssen

Ir. G. Bosch and Ir. J. H. H. Janssen are with the Philips Radio, Television and Record-playing Equipment Division, Eindhoven.

Objects of glassy carbon

A new solid modification of carbon that is very different in some ways from diamond and graphite, the known modifications, has been prepared by means of a synthetic method [1]. This "glassy" carbon is obtained by decomposition (pyrolysis) of appropriate polymers, such as phenol-formaldehyde resins. In its physical properties glassy carbon is completely isotropic. The new material combines the useful features of graphite (high thermal and electrical conductivities and ability to withstand high temperatures and temperature fluctuations) with very high strength. Its hardness (Mohs 6-7) is much greater than that of graphite, and even though its density is low (1.4-1.5 g/cm^3) it is impermeable to gases. The material is exceptionally resistant to corrosion, and many experiments have demonstrated that glassy carbon is a very suitable material for crucibles for chemical and metallurgical work. The photograph shows a number of crucibles, bowls and discs of different shapes and sizes made in the Philips Aachen laboratories.

The piece at the lower right consists of *carbon foam* strengthened with glassy carbon. This material differs from the commercially available carbon foam in its very much greater strength and greater resistance to oxidation (it will not burn in air below 600 °C). This strengthened carbon foam is obtained by carbonization of a suitably impregnated plastic polymer foam, like the phenol-formaldehyde resins mentioned above. The density can be varied from 0.1 g/cm^3 to say 1.0 g/cm^3; the corresponding bulk modulus varying from 10 to 25 000 N/cm^2 but the thermal conductivity only from 8×10^{-4} to 14×10^{-4} J/°C s cm. This is a most advantageous feature for heat insulation, which is one of the principal applications of this material. Strengthened carbon foam can also be used as a structural material. It can be machined either before or after carbonization, although a diamond tool has to be used with the carbonized material.

[1] B. Lersmacher, H. Lydtin and W. F. Knippenberg, Glasartiger Kohlenstoff, Chemie-Ing.-Technik **42**, 659-669, 1970.

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.

R. P. Adriaanse & P. van der Laan (Philips Information Systems and Automation Division, Eindhoven): Some remarks on a general class of Markov processes with discrete time parameter and dependent increments.
Technometrics **12**, 851-866, 1970 (No. 4).

C. S. Aitchison & R. Davies: Large-signal varactor measurements.
Electronics Letters **6**, 780-781, 1970 (No. 24). *M*

C. S. Aitchison & B. H. Newton: Varactor-tuned X band Gunn oscillator using lumped thin-film circuits.
Electronics Letters **7**, 93-94, 1971 (No. 4). *M*

G. A. Allen: The performance of negative electron affinity photocathodes.
J. Physics D **4**, 308-317, 1971 (No. 2). *M*

J. A. Appels & M. M. Paffen: Local oxidation of silicon; new technological aspects.
Philips Res. Repts. **26**, 157-165, 1971 (No. 3). *E*

G. Arlt, W. Puschert & P. Quadflieg: Dielectric behaviour of lithium iodate.
Phys. Stat. sol. (a) **3**, K 243-246, 1970 (No. 4). *A*

V. Belevitch: Relationships between design parameters of very selective filters.
Electronics Letters **6**, 585, 1970 (No. 18). *B*

F. Berz & C. G. Prior: Test of McWhorter's model of low-frequency noise in Si m.o.s.t.s.
Electronics Letters **6**, 595-597, 1970 (No. 19). *M*

F. Berz & E. Pyrah: Numerical-integration procedures in transistor-analysis computer programs.
Electronics Letters **7**, 94-97, 1971 (No. 4). *M*

G. Blasse, A. Bril & J. A. de Poorter: Radiationless transitions in the Eu^{3+} center in $LaAlO_3$.
J. chem. Phys. **53**, 4450-4453, 1970 (No. 12). *E*

G. Blasse & A. D. M. de Pauw: Crystal structure of some $LiMe^{5+}Me^{6+}O_6$ compounds.
J. inorg. nucl. Chem. **32**, 3960-3961, 1970 (No. 12). *E*

J. Bloem (Philips Semiconductor Development Laboratory, Nijmegen): Silicon epitaxy from mixtures of SiH_4 and HCl.
J. Electrochem. Soc. **117**, 1397-1401, 1970 (No. 11).

P. A. Boter, M. D. Wijnen & H. L. C. Nuyens (Philips Electroplating Laboratories, Eindhoven): Some new refill cell systems.
Power Sources 1968, editor D. H. Collins, Pergamon Press, Oxford 1970, pp. 359-371. *E*

A. Boucher & L. Hollan: Thermodynamic and experimental aspects of gallium arsenide vapor growth.
J. Electrochem. Soc. **117**, 932-936, 1970 (No. 7). *L*

J. Bouma, A. J. J. Franken & J. D. B. Veldkamp: A tensile testing machine for thin filaments of high strength and stiffness.
J. Physics E **3**, 1006-1008, 1970 (No. 12). *E*

J.-P. Boutot & G. Piétri: Ultrahigh-speed microchannel photomultiplier.
IEEE Trans. **ED-17**, 493-495, 1970 (No. 7). *L*

M. J. G. Braken: Het lassen van dunne materialen.
Lastechniek **36**, 243-247, 1970 (No. 11). *E*

P. C. Brandon: Thiocyanato-indoles as energy-transfer inhibitors in photophosphorylation.
Arch. Biochem. Biophys. **138**, 566-573, 1970 (No. 2). *E*

P. Branquart & J. Lewi: Structure d'un compilateur ALGOL 68.
Congrès AFCET, Paris 1970, pp. 5.4.138-152. *B*

P. B. Braun, J. Hornstra & J. I. Leenhouts: The crystal structure of tricyclo[4,3,1,$1^{3,8}$]undecane-4,5-dione.
Acta cryst. **B 26**, 1802-1806, 1970 (No. 11). *E*

C. H. J. van den Brekel: Easy controllable injection of vapours in carrier gases.
J. Physics E **3**, 878-880, 1970 (No. 11). *E*

J. C. Brice: Pulling crystals for optical devices.
Opt. Laser Technol. **2**, 206-208, 1970 (No. 4). *M*

E. Bruninx: The measurement of fast neutron flux densities above 0.5 MeV by a simple moderation method.
Int. J. appl. Rad. Isot. **21**, 657-666, 1970 (No. 11). *E*

K. H. J. Buschow: Rare-earth cobalt intermetallic compounds.
Les éléments des terres rares, Coll. Int. C.N.R.S. No. 180, Paris-Grenoble 1969, tome I, pp. 101-112; 1970. *E*

K. H. J. Buschow: The crystal structure of Th_2Co_7.
Acta cryst. **B 26**, 1389-1392, 1970 (No. 10). *E*

K. H. J. Buschow, A. M. van Diepen (Natuurkundig Laboratorium der Universiteit van Amsterdam) **& H. W. de Wijn** (Natuurk. Lab. Univ. Amst.): Magnetic properties and nuclear magnetic resonance of cubic RCu_5 intermetallic compounds.
J. appl. Phys. **41**, 4609-4612, 1970 (No. 11). *E*

K. H. J. Buschow & A. S. van der Goot: The crystal structure of rare-earth nickel compounds of the type R_2Ni_7.
J. less-common Met. **22**, 419-428, 1970 (No. 4). *E*

K. H. J. Buschow & R. P. van Stapele: Magnetic properties of some cubic rare-earth-iron compounds of the type RFe_2 and $R_xY_{1-x}Fe_2$.
J. appl. Phys. **41**, 4066-4069, 1970 (No. 10). *E*

K. H. J. Buschow & J. S. van Wieringen: Crystal structure and magnetic properties of cerium-iron compounds.
Phys. Stat. sol. **42**, 231-239, 1970 (No. 1). *E*

H. B. G. Casimir: Microwave and optical generation and amplification.
Proc. MOGA Conf., Amsterdam 1970, pp. 0.1-0.6. *E*

Ph. Chevalier, J.-P. Boutot & G. Piétri: A PM of new design for high speed physics.
IEEE Trans. **NS-17**, No. 3, 75-78, June 1970. *L*

T. Chisholm & A. M. Stark: A technique for the computation of charged-particle trajectories in radio-frequency quadrupole devices.
J. Physics D **3**, 1717-1726, 1970 (No. 11). *M*

R. W. Cooper: Optical communication.
Rev. HF **8**, 67-72, 1970 (No. 3). *M*

C. D. Corbey & R. Davies: Amplitude stabilisation of a varactor frequency multiplier using self-bias resistance compensation.
Electronics Letters **7**, 151-152, 1971 (No. 7). *M*

R. Cosier, A. Wise, A. Tressaud (Faculté des Sciences de Bordeaux), **J. Grannec** (Fac. Sci. Bordeaux), **R. Olazcuaga** (Fac. Sci. Bordeaux) **& J. Portier** (Fac. Sci. Bordeaux): Sur de nouveaux composés fluorés ferrimagnétiques à structure wébérite.
C.R. Acad. Sci. Paris **271C**, 142-145, 1970 (No. 2). *M*

J. B. Coughlin & R. W. Lindop: Subnanosecond delays in circuit complexes measured by slice probing.
Solid State Technol. **14**, No. 3, pp. 12 & 70, March 1971. *M*

H. J. van Daal & K. H. J. Buschow: Kondo effect in some intermetallic compounds of Ce.
Phys. Stat. sol. (a) **3**, 853-871, 1970 (No. 4). *E*

P. A. van Dalen & C. A. A. J. Greebe: Generation of Bleustein-Gulyaev waves in piezoelectric plates.
Physics Letters **33A**, 93-94, 1970 (No. 2). *E*

P. Delsarte: BCH bounds for a class of cyclic codes.
SIAM J. appl. Math. **19**, 420-429, 1970 (No. 2). *B*

P. Delsarte, J. M. Goethals & F. J. MacWilliams (Bell Telephone Laboratories, Murray Hill, N.J.): On generalized Reed-Muller codes and their relatives.
Information and Control **16**, 403-442, 1970 (No. 5). *B*

F. Desvignes, V. Duchenois & R. Polaert: Equipement pour la mesure continue du diamètre de fibres.
Techniques Philips 1970, No. 5, 2-10. *L*

F. Dintelmann (II. Physikalisches Institut der Technischen Hochschule Darmstadt), **E. Dormann** (II. Phys. Inst. T.H. Darmstadt) **& K. H. J. Buschow:** NMR-investigation on ferromagnetic, yttrium-diluted $GdAl_2$.
Solid State Comm. **8**, 1911-1913, 1970 (No. 22). *E*

J. A. W. van der Does de Bye, A. T. Vink, A. J. Bosman & R. C. Peters: Kinetics of green and red-orange pair luminescence in GaP.
J. Luminescence **3**, 185-197, 1970 (No. 3). *E*

F. Doittau, H. Bourcier, G. Verschoore, M. Hébert (all with SODERN, Paris) **& J. C. Pauwels:** Caméras ultra-rapides. Problèmes relatifs au dispositif d'ouverture du tube obturateur (générateur d'impulsions très haute tension).
Techniques Philips 1970, No. 5, 11-22. *L*

H. Duifhuis (Institute for Perception Research, Eindhoven): Audibility of high harmonics in a periodic pulse.
J. Acoust. Soc. Amer. **48**, 888-893, 1970 (No. 4, Part 2).

A. J. Fox & P. W. Whipps: Longitudinal quadratic electro-optical effect in KTN.
Electronics Letters **7**, 139-140, 1971 (No. 5/6). *M*

K. G. Freeman: The physics of colour television.
Physics Education **5**, 326-331, 1970 (No. 6). *M*

K. G. Freeman & R. E. Ford (University of Surrey, Guildford): Variable gamma corrector improves television video signals.
Electronic Engng. **42**, No. 511, 90-93, Sept. 1970. *M*

K. L. Fuller: AVOID — short-range high-definition radar.
Wireless World **77**, 110-113, 1971 (No. 1425). *M*

S. Garbe & G. Frank: Photo-emission from silicon-doped *p*-type gallium arsenide.
Gallium Arsenide, Proc. 3rd Symp., 1970, pp. 208-211; 1971. *A*

R. Genève: Some complementary observations as conclusion *(to the series of papers on medical thermography in Acta Electronica).*
Acta Electronica **13**, 179-183 (in French), 185-189 (in English), 1970 (No. 2). *L*

T. G. Gijsbers & W. J. Hoogenboezem (Philips Mechanical Engineering Works, Eindhoven): Een numeriek bestuurde coördinatograaf biedt vele mogelijkheden.
Metaalbewerking **36**, 347-352, 1970 (No. 13). *E*

J. J. Goedbloed: On the up-converted noise of IMPATT diode oscillators.
Proc. MOGA Conf., Amsterdam 1970, pp. 12.36-12.40. *E*

J. M. Goethals & J. J. Seidel (Technical University of Eindhoven): Strongly regular graphs derived from combinatorial designs.
Canad. J. Math. **22**, 597-614, 1970 (No. 3). *B*

H. C. de Graaff: A modified charge-control theory for saturated transistors.
Philips Res. Repts. **26**, 191-215, 1971 (No. 3). *E*

G. Groh: Holographische Methoden in der IC-Technologie.
Mikroelektronik 4 (Vortr. Kongreß INEA, München 1970), 162-179, 1971. *H*

J. de Groot & M. T. Vlaardingerbroek: Some numerical results on modes of oscillation in a transferred-electron device.
Proc. MOGA Conf., Amsterdam 1970, pp. 20.34-20.39. *E*

S. H. Hagen: Leitungsmechanismus in spannungsabhängigen SiC-Widerständen.
Ber. Dtsch. Keram. Ges. **47**, 630-634, 1970 (No. 10). *E*

P. Hansen: Contribution of some 4*d* and 5*d* transition-metal ions on octahedral sites to the anisotropy of ferrites and garnets.
Phys. Rev. B **3**, 862-870, 1971 (No. 3). *H*

P. A. H. Hart, J. A. Pals, J. Prins & C. A. G. Stricker: Noise-measuring set-up for devices under low-gain conditions.
Philips Res. Repts. **26**, 216-228, 1971 (No. 3). *E*

H. Haug & K. Weiss: A new derivation of the equations of motion for the complex order parameter of helium II.
Physics Letters **33A**, 263-264, 1970 (No. 4). *E*

E. E. Havinga, H. Damsma & M. H. van Maaren: Oscillatory dependence of superconductive critical temperature on number of valency electrons in Cu_3Au-type alloys.
J. Phys. Chem. Solids **31**, 2653-2662, 1970 (No. 12). *E*

D. Hennings & K. H. Härdtl: The distribution of vacancies in lanthana-doped lead titanate.
Phys. Stat. sol. (a) **3**, 465-474, 1970 (No. 2). *A*

B. Hill: Spatial noise in optical data-storage systems using amplitude Fourier-transform holograms.
J. Opt. Soc. Amer. **61**, 386-398, 1971 (No. 3). *H*

F. N. Hooge: $1/f$ noise in the conductance of ions in aqueous solutions.
Physics Letters **33A**, 169-170, 1970 (No. 3). *E*

B. B. van Iperen & H. Tjassens: Measurement of large-signal impedance, optimum ac voltage and efficiency of Si pnn$^+$, Ge npp$^+$ and GaAs Schottky-barrier avalanche transit time diodes.
Proc. MOGA Conf., Amsterdam 1970, pp. 7.27-7.32. *E*

R. E. Jesse & H. F. J. I. Giller: Cellular growth: the relation between growth velocity and cell size of some alloys of cadmium and zinc.
J. Crystal Growth **7**, 348-352, 1970 (No. 3). *E*

H. Kalter, J. J. H. Schatorjé & E. Kooi: Electric double layers in MIS structures with multilayered dielectrics.
Philips Res. Repts. **26**, 181-190, 1971 (No. 3). *E*

M. A. Karsmakers: Het verbinden van glas aan metaal.
Glastechn. Meded. **8**, 116-125, 1970 (No. 4). *E*

D. Kasperkovitz & J. G. van Santen: Integration density and power dissipation of MOS and bipolar shift registers — A comparison.
Microelectronics and Reliability **9**, 497-501, 1970 (No. 6). *E*

J. T. Klomp: Festkörperbindung zwischen Metall und Keramik.
Ber. Dtsch. Keram. Ges. **47**, 627-629, 1970 (No. 10). *E*

A. Klopfer: Wasserstoffdesorption von Wolfram durch Elektronenstoß.
Vakuum-Technik **19**, 167-170, 1970 (No. 7). *A*

J. E. Knowles & P. Rankin: Disaccommodation of permeability in manganese-zinc-titanium ferrites.
J. Physique **32**, C1/845-846, 1971 (Colloque No. 1, Vol. II). *M*

H. G. Kock & D. de Nobel: Technology of silicon Schottky barrier IMPATT diodes.
Proc. MOGA Conf., Amsterdam 1970, pp. 7.1-7.3. *E*

H. Koeman (Philips Research Labs. Amsterdam): A controlled analogue pulse height store.
Nucl. Instr. Meth. **86**, 301-309, 1970 (No. 2).

E. Kooi, J. G. van Lierop, W. H. C. G. Verkuijlen & R. de Werdt: LOCOS devices.
Philips Res. Repts. **26**, 166-180, 1971 (No. 3). *E*

L. J. Koppens: Beziehung zwischen magnetischen Eigenschaften und Mikrostruktur von Speicherkernen, die mit einer neuen Technik hergestellt wurden.
Ber. Dtsch. Keram. Ges. **47**, 654-657, 1970 (No. 10). *E*

E. Kraetzig: Investigation of superconducting proximity effects with acoustic surface waves.
Physics Letters **33A**, 343-344, 1970 (No. 6). *H*

D. E. Lacklison, J. Chadwick & J. L. Page: Photomagnetic effect in ferric borate.
J. appl. Phys. **42**, 1445-1446, 1971 (No. 4). *M*

P. R. Locher & R. P. van Stapele: Supertransferred hyperfine fields on tetrahedral sites in some chromium sulpho- and selenospinels.
J. Phys. Chem. Solids **31**, 2643-2652, 1970 (No. 12). *E*

J. Loeckx: The parsing for general phrase-structure grammars.
Information and Control **16**, 443-464, 1970 (No. 5). *B*

M. H. van Maaren, H. B. Harland & E. E. Havinga: Critical carrier concentration for superconductivity in mixed metal-semiconductor systems.
Solid State Comm. **8**, 1933-1935, 1970 (No. 22). *E*

F. E. Maranzana: Kondo sidebands.
Phys. Rev. Letters **25**, 239-242, 1970 (No. 4). *E*

R. J. Meijer: De Philips stirlingmotor (edited by **H. van Ginkel**).
Chem. Weekblad **66**, No. 44, 58-62, 30 okt. 1970. *E*

R. Memming & H. Tributsch (Physikalisch Chemisches Institut, Technische Hochschule, Munich): Electrochemical investigations on the spectral sensitization of gallium phosphide electrodes.
J. phys. Chem. **75**, 562-570, 1971 (No. 4). *H*

R. Metselaar: Een nieuwe wisselwerking: het fotomagnetisch effect.
Chem. Weekblad **66**, No. 41, 31-34, 9 okt. 1970. *E*

R. Metselaar, P. J. Rijnierse & U. Enz: Lichtinduzierte Änderungen der magnetischen Eigenschaften von polykristallinem Yttrium-Eisen-Granat.
Ber. Dtsch. Keram. Ges. **47**, 663-665, 1970 (No. 10). *E*

D. Meyer-Ebrecht: Entwurfsprinzipien für Präzisions-Relaxationsoszillatoren.
Mikroelektronik 4 (Vortr. Kongreß INEA, München 1970), 391-403, 1971. *H*

M. Monneraye: Le scellement métal céramique. Un renouveau du scellement non métallique.
Techniques Philips 1970, No. 5, 23-35. *L*

B. J. Mulder: Preparation of $BaTiO_3$ and other ceramic powders by coprecipitation of citrates in an alcohol.
Amer. Ceramic Soc. Bull. **49**, 990-993, 1970 (No. 11). *E*

G. T. M. Neelen: Vacuum brazing of complex heat exchangers for the Stirling engine.
Welding J. **49**, 381-386, 1970 (No. 5). *E*

D. de Nobel & R. P. Tijburg: Improved method for the fabrication of GaAs microwave devices.
Proc. MOGA Conf., Amsterdam 1970, pp. 9.1-9.3. *E*

R. C. Oldfield: Stratification in evaporated nickel-iron films.
J. Physics D **3**, 1495-1496, 1970 (No. 10). *M*

W. J. Oosterkamp: Die Entwicklung der Röntgenröhre.
Röntgenpraxis **23**, 252-260, 1970 (No. 11). *E*

Ph. Piret: Les codes de convolution utilisés pour la correction d'erreurs en paquets.
Rev. HF **8**, 49-59, 1970 (No. 2). *B*

R. J. van de Plassche: A new high speed operational amplifier.
Mikroelektronik 4 (Vortr. Kongreß INEA, München 1970), 336-350, 1971. *E*

E. Roeder: Extrusion of glass.
J. non-cryst. Solids **5**, 377-388, 1971 (No. 5). *A*

F. L. J. Sangster: Integrated MOS and bipolar analog delay lines using bucket-brigade capacitor storage.
1970 IEEE Int. Solid-State Circuits Conf. Digest tech. Papers, pp. 74-75 & 185. *E*

P. Schagen: Electronic aids to night vision.
Phil. Trans. Roy. Soc. London A **269**, 233-263, 1971 (No. 1196). *M*

C. Schiller: Interface analysis by X-ray diffraction topography.
Solid-State Electronics **13**, 1163-1166, 1970 (No. 8). *L*

U. J. Schmidt & W. Thust: Korrektur der Ablenkfehler doppelbrechender Prismen in digitalen Laserstrahlablenkern und Bildvervielfachern.
Optik **32**, 570-584, 1971 (No. 6). *H*

J. F. Schouten (Institute for Perception Research, Eindhoven): Technologische en maatschappelijke evolutie.
Natuurk. Voordr. Diligentia Nieuwe Reeks No. 48, 51-60, 1970.

J. F. Schouten (Institute for Perception Research, Eindhoven): Tijd voor toonhoogte.
Versl. gew. Verg. Afd. Natuurk. Kon. Ned. Akad. Wetensch. **79**, 150-153, 1970 (No. 9).

E. Schwartz: Obere Grenzen für die Empfindlichkeit der Resonanzfrequenzen von Reaktanzzweipolen.
Nachrichtentechn. Z. **23**, 553-558, 1970 (No. 11). *A*

P. J. Severin: On the infrared thickness measurement of epitaxially grown silicon layers.
Appl. Optics **9**, 2381-2387, 1970 (No. 10). *E*

A. M. Stark & B. Singer (Philips Laboratories Briarcliff Manor, N.Y., U.S.A.): Beam discharge lag in the silicon diode array camera tube.
1971 IEEE Int. Solid-State Circuits Conf. Digest tech. Papers, pp. 136-137. *M*

A. Thayse: Boolean differential calculus.
Philips Res. Repts. **26**, 229-246, 1971 (No. 3). *B*

J. P. Thiran, Ph. van Bastelaer & C. Wellekens: Description of a filter synthesis and analysis program with a special study of the numerical accuracy.
Rev. MBLE **13**, 39-58, 1970 (No. 2). *B*

D. R. Tilley: Superradiance in arrays of superconducting weak links.
Physics Letters **33A**, 205-206, 1970 (No. 4). *E*

R. J. Tree, M. J. Josh & C. T. Foxon: On a failure mechanism in indium phosphide microwave oscillators.
Solid-State Electronics **14**, 519-520, 1971 (No. 6). *M*

J. Ubbink: Optimization of the rotor surface resistance of the asynchronous electrostatic motor.
Appl. sci. Res. **22**, 442-448, 1970 (No. 6). *E*

J. Ungelenk: Herstellung dünner epitaktischer Galliumarsenidschichten durch Ionenzerstäubung.
Vakuum-Technik **19**, 231-234, 1970 (No. 9). *A*

J. G. Verhagen, G. den Ouden, A. Liefkens & G. W. Tichelaar: Nitrogen absorption by ferritic weld metal during arc welding.
Metal Constr. Brit. Welding J. **2**, 135-143, 1970 (No. 4). *E*

A. G. van Vijfeijken: Technological forecasting: een bruikbaar instrument voor wetenschapsbeleid?
Ned. T. Natuurk. **37**, 5-11, 1971 (No. 1). *E*

A. T. Vink & R. C. Peters: Absorption and luminescence due to excitons bound to neutral acceptors in GaP.
J. Luminescence **3**, 209-229, 1970 (No. 3). *E*

J. Vlietstra: FOURBAR, een interactief systeem voor het samenstellen van vierstangenmechanismes.
Informatie **12**, 508-516, 1970 (No. 12). *E*

K. Weiss: Untersuchungen an kubischem Kupfer(I)sulfid (Digenit), III. Zum elastischen Entmischungseffekt in Digenit.
Berichte Bunsen-Ges. phys. Chemie **74**, 1257-1261, 1970 (No. 12). *E*

F. F. Westendorp & A. G. Rijnbeek: A transducer to measure forces under high pressure.
Rev. sci. Instr. **41**, 1881-1882, 1970 (No. 12). *E*

H. J. de Wit & C. Crevecoeur: Deviation from Ohm's law in As_2Se_3 glass.
Physics Letters **33A**, 25-26, 1970 (No. 1). *E*

J. P. Woerdman: Formation of a transient free carrier hologram in Si.
Optics Comm. **2**, 212-214, 1970 (No. 5). *E*

H. Zijlstra: Domain-wall processes in $SmCo_5$ powders.
J. appl. Phys. **41**, 4881-4885, 1970 (No. 12). *E*

Contents of Electronic Applications **30**, No. 3, 1970:

Contents of Valvo Berichte **16**, No. 4, 1971: