PHILIPS TECHNICAL REVIEW

VOLUME 35
1975

Subject index, Volume 35, 1975

Aircraft guidance equipment, MADGE 271
Avalanche photodiodes, fast scanning microscope used in the development of 23
Bearing systems, spiral-groove, with grease . . . 137
Boron filament: a light, stiff, strong material . . 125
Brushless d.c. machines 106
Bubble memories, garnet films for 1
Cathodoluminescence for investigation of crystal defects 239
Ceramics, pore-free, isostatic hot pressing of, without metallic encapsulating layer 65
Ceramic-to-metal bonding with sputtering as metallization technique 209
Colour rendering of light sources 347, 361
Computer-aided design of logic circuits 261
Crystal defects and their transport in solids; industrial applications 181
Crystal defects, investigation by cathodoluminescence 239
Diffusion of donors and acceptors in silicon . . 181
Digital circuits, structural testing of 261
Digital ICs with low dissipation 212
Discharge lamps:
survey 308
high-pressure sodium lamp 334
with metal halides 347
getter for metal-iodide high-pressure mercury-vapour lamps 354
electrodes in 356
Electric motors:
early history of the electric motor 77
designing a small d.c. motor 96
applications of stepping motors 104
steady-state performance of a class of electronically commutated d.c. machines 106
scaling laws for electric motors 116
Electrical conduction in ferromagnetic metals . . 29
Electron microscope, scanning, type PSEM 500 . 153
Electron resists for the manufacture of integrated circuits 41
Endoscope, experimental, with miniature TV camera 166
Etching with ions 199
Extractor, simple and flexible automatic 196
Ferromagnetic metals, electrical conduction in 29
Filament composites 125
Fluorescent lamps 321, 361
Garnet bubble materials, perfection of 1
Gas discharges, investigation of, by light scattering 344
Gas discharges, low-pressure 321
Holographic strain analysis 53
HPI lamps 347
Incandescent lamps:
research on, survey 295
tungsten transport and burn-out 296
regenerative chemical cycles in tungsten-halogen lamps 302
carbon-filament lamps with a chemical transport cycle 316
temperature profiling of tungsten filaments by a chemical transport reaction 332
Infrared detection with pyroelectric materials based on triglycine sulphate 247
Integrated circuits, digital, with low dissipation 212
Interconnection patterns for hybrid ICs 144
Ion etching, investigation of 199
Isostatic hot pressing of ceramics without metallic encapsulating layer 65
Light:
demonstration and test vehicle for exterior lighting 307
open-air laboratory for road lighting 320
low-pressure gas discharges 321
investigation of gas discharges by light scattering 344
optimum spectra for light sources 361
see also **Discharge lamps** and **Incandescent lamps**
Logic circuits, testing of 261
MADGE, microwave aircraft digital guidance equipment 271
Magnetic anisotropy and magnetostriction of Ru and Ir ions in yttrium iron garnet 225
Magnetic bubbles in iron garnet 1
Medical diagnosis: automatic measurement of X-ray photographs 170
Metals, ferromagnetic, electrical conduction in 29
Metal-halide lamps 347
Motors, see **Electric motors**
PINPOINT radio system for locating and monitoring vehicles 15
'Push-pull' spiral-groove bearing 11
Pyroelectric materials based on triglycine sulphate (TGS) for infrared detection 247
Refractometer, fast, for evaporated thin films . . 142
Resists, electron, for the manufacture of integrated circuits 41
Scanning electron microscope PSEM 500 . . . 153
Scanning microscope, fast, used in the development of avalanche photodiodes 23
Semiconductor technology with electron resists 41
Sintering 188
Sodium lamps 321, 334
Spectra for light sources, optimum 361
Spiral-groove bearing, 'push-pull' 11
Spiral-groove bearing systems with grease . . . 137
Sputtering as metallization technique for ceramic-to-metal bonding 209
Stepping motors, applications of 104
Structural testing of digital circuits 261
Television camera, miniature, in experimental endoscope 166
Thick-film technology 144
Thin evaporated films, fast refractometer for . . 142
Tin-halide lamp 347
Transport cycle, chemical, in carbon-filament lamps 316
Transport cycle, chemical, in tungsten-halogen lamps 302
Triglycine sulphate (TGS) as a pyroelectric material for infrared detection 247
Vehicle location with PINPOINT radio system 15
X-ray photographs, medical, automatic measurement of 170
YIG, magnetic anisotropy and magnetostriction of Ru and Ir ions in 225

Author index, Volume 35, 1975

Page

Adriaansz, M., see Vriens, L.
Bleekrode, R., M. Koedam and L. Rehder
Discharge lamps 308
Boers, P. M. and L. J. M. Bollen
A fast scanning microscope used in the development of avalanche photodiodes 23
Bollen, L. J. M., see Boers, P. M.
Bos, J. G. G.
Spiral-groove bearing systems with grease 137
Boulou, M., see Schiller, C.
Bouwknegt, A., H. Nienhuis, D. J. Schipper and P. A. W. Tielemans
Electrodes in discharge lamps 356
Bowers, Brian
The early history of the electric motor 77
Bronnes, R. L., R. C. Hughes and R. C. Sweet
Ceramic-to-metal bonding with sputtering as a metallization technique 209
Corbett, B. D.
'MADGE', a microwave aircraft digital guidance equipment, II. The data link: data transmission and distance measurement 271
Dettingmeijer, J.H., G. Dittmer, A. Klopfer and J. Schröder
Research on incandescent lamps, III. Regenerative chemical cycles in tungsten-halogen lamps 302
Dijken, R. H.
Designing a small d.c. motor 96
Dimigen, H. and H. Lüthje
An investigation of ion etching 199
Dittmer, G., see Dettingmeijer, J. H.
Dorleijn, J. W. F., see Miedema, A. R.
Drop, P. C., E. Fischer, F. Oostvogels and G. A. Wesselink
Metal-halide discharge lamps 347
Fischer, E., J. Fitzgerald, W. Lechner and W. Lems
Research on incandescent lamps, II. Transport and burn-out in incandescent lamps 296
Fischer, E., see Drop, P. C.
Fitzgerald, J., see Fischer, E.
Funk, W.
Thick-film technology 144
Gibson, R. W.
PINPOINT — a radio system for locating and monitoring vehicles 15
Groot, J. J. de, J. A. J. M. van Vliet and J. H. Waszink
The high-pressure sodium lamp 334
Hansen, P.
Magnetic anisotropy and magnetostriction of Ru and Ir ions in yttrium iron garnet 225
Härdtl, K. H.
A simplified method for the isostatic hot pressing of ceramics 65
Hoek, W. J. van den and W. A. Klessens
Carbon-filament lamps with a chemical transport cycle 316
Hughes, R. C., see Bronnes, R. L.
Kamerbeek, E. M. H.
Scaling laws for electric motors 116
Keve, E. T.
Pyroelectric materials based on triglycine sulphate (TGS) for infrared detection 247
Klessens, W. A., see Hoek, W. J. van den
Klopfer, A., see Dettingmeijer, J. H.
Koedam, M., see Bleekrode, R.
Kreuwels, W. G. J.
Structural testing of digital circuits 261
Kuus, G.
A getter for metal-iodide high-pressure mercury-vapour lamps 354

Pa

Kuypers, W. and J. C. Tiemeijer
The Philips PSEM 500 scanning electron microscope . . 1
Lechner, W., see Fischer, E.
Lems, W., see Fischer, E.
Lüthje, H., see Dimigen, H.
Maaren, A. C. van, O. Schob and W. Westerveld
Boron filament: a light, stiff and strong material . . . 12
Miedema, A. R. and J. W. F. Dorleijn
Electrical conduction in ferromagnetic metals 2
Nienhuis, H., see Bouwknegt, A.
Oostvogels, F., see Drop, P. C.
Opstelten, J. J., D. Radielović and J. M. P. J. Verstegen
Optimum spectra for light sources 36
Périlhou, J.
An experimental endoscope with miniature television camera 16
Pistorius, J.-A., J. M. Robertson and W. T. Stacy
The perfection of garnet bubble materials . . . . . .
Polman, J., H. van Tongeren and T. G. Verbeek
Low-pressure gas discharges 32
Radielović, D., see Opstelten, J. J.
Radziwill, W.
Steady-state performance of a class of electronically commutated d.c. machines 10
Rehder, L., see Bleekrode, R.
Roberts, E. D.
Electron resists for the manufacture of integrated circuits 4
Robertson, J. M., see Pistorius, J. A.
Schiller, C. and M. Boulou
Investigation of crystal defects by cathodoluminescence 23
Schipper, D. J., see Bouwknegt, A.
Schob, O., see Maaren, A. C. van
Schröder, J.
Temperature profiling of tungsten filaments in incandescent lamps by a chemical transport reaction 33
Schröder, J., see Dettingmeijer, J. H.
Spiesberger, W. and M. Tasto
The automatic measurement of medical X-ray photographs 17
Stacy, W. T., see Pistorius, J. A.
Sweet, R. C., see Bronnes, R. L.
Tasto, M., see Spiesberger, W.
Tielemans, P. A. W., see Bouwknegt, A.
Tiemeijer, J. C., see Kuypers, W.
Tongeren, H. van, see Polman, J.
Tooren, A. van
A simple and flexible automatic extractor 19
Troye, N. C. de
Digital integrated circuits with low dissipation 21
Velzel, C. H. F.
Holographic strain analysis 5
Verbeek, T. G., see Polman, J.
Verstegen, J. M. P. J., see Opstelten, J. J.
Vink, H. J.
Crystal defects and their transport in solids; industrial applications 18
Vliet, J. A. J. M. van, see Groot, J. J. de
Volman, H. J. W. M.
The 'push-pull' spiral-groove bearing — a thrust bearing with self-adjusting internal preloading 1
Vriens, L. and M. Adriaansz
Investigation of gas discharges by light scattering . . . 34
Wal, J. van der
A fast refractometer for evaporated thin films 14
Waszink, J. H., see Groot, J. J. de
Wesselink, G. A., see Drop, P. C.
Westerveld, W., see Maaren, A. C. van

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 1

The perfection of garnet bubble materials

J. A. Pistorius, J. M. Robertson and W. T. Stacy

Some fifteen years ago, when U. Enz and C. Kooij were studying the magnetic domain configurations in thin plates of $BaFe_{12}O_{19}$ in a perpendicular magnetic field, they found that there was a narrow field range in which a plate could be magnetized parallel to the field except for a number of tiny cylindrical domains that were magnetized in the opposite sense. These cylindrical domains, now known as magnetic bubbles, have been recognized as ideal carriers of information, and the preparation of magnetic films in which such bubbles can move around freely remains the subject of a great deal of research. The films are usually grown epitaxially on a nonmagnetic single-crystal substrate. Two Philips teams are working on this subject. One is at Philips Research Laboratories, Eindhoven, and is doing general research on the growth and perfection of such magnetic films. The other is at Philips Forschungslaboratorium Hamburg, and is concentrating on the problems that arise when films have to be made reproducibly within the operational margins of practical device applications. The present article, by members of the Eindhoven team, deals with this subject, concentrating on the materials that appear most suitable: iron garnets, and on the method that looks most promising for making the films: liquid-phase epitaxy.

Introduction

In certain ferrimagnetic materials the magnetization has a strong preference for one crystallographic direction, owing to a uniaxial anisotropy of the material. A thin plate of such a material, cut from a single crystal and perpendicular to this direction, separates into snake-like domains of opposite magnetization (*fig. 1a*). In a slowly increasing magnetic field in the preferred direction (i.e. perpendicular to the plate), the domains of antiparallel magnetization shrink, the others grow (fig. 1*b*). At a certain stage only separate, small, cylindrical domains magnetized antiparallel are left (fig. 1*c*). These are the well known 'magnetic bubbles'. At a further stage these also collapse, and the slice is completely magnetized in the direction of the field [1].

Recently magnetic bubbles have been recognized as ideal carriers of information [2]. Under favourable conditions they are extremely mobile, stable entities of only a few microns in diameter that can easily be generated, moved around and annihilated. Magnetic-

J. A. Pistorius, Dr J. M. Robertson and Dr W. T. Stacy are with Philips Research Laboratories, Eindhoven.

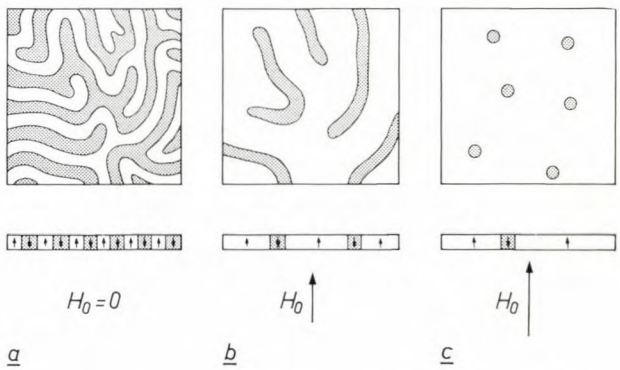


Fig. 1. Magnetic domains in a thin plate of a magnetic material with a strong preference for magnetization perpendicular to the plane of the plate. *Above*: top view, *below*: cross-section of the plate. In the absence of an external magnetic field the plate is equally divided between domains of opposite magnetization (*a*). In an increasing field H_0 perpendicular to the plate the domains of antiparallel magnetization shrink (*b*), until only small, cylindrical domains ('magnetic bubbles') are left (*c*).

[1] C. Kooy and U. Enz, Philips Res. Repts. **15**, 7, 1960.

[2] A. H. Bobeck, R. F. Fischer, A. J. Perneski, J. P. Remeika and L. G. Van Uitert, IEEE Trans. **MAG-5**, 544, 1969. See also A. H. Bobeck and H. E. D. Scovil, Sci. Amer. **224**, No. 6, 78, June 1971.

bubble memories, in which each bubble would carry one bit of information, promise a combination of memory capacity and access rate that compares favourably with memories available now. *Fig. 2* shows one way of moving bubbles around in a well defined manner, and in *fig. 3* a laboratory model of a magnetic-bubble memory is shown.

The primary problem in this field of research is that of preparing the 'plates' that carry the bubbles. A stable bubble of one or a few microns in diameter can only exist in a plate a few microns thick. Many imperfections that are not much smaller than the bubble itself tend to pin it, impeding its motion. The plate should therefore be a perfect single crystal and its faces should be perfectly flat and parallel. At present the most promising answer to this problem is to grow a single-crystal magnetic film epitaxially on a nonmagnetic single-crystal substrate, and the most promising materials to be used are iron garnets for the magnetic film combined with nonmagnetic garnets for the substrate.

Of the various techniques for growing epitaxial garnet films, *liquid-phase epitaxy* [4] has been the most successful to date. It consists in dipping a polished substrate slice in a melt that is *supersaturated* to the appropriate degree with the required oxides. With due precautions the iron garnet will nucleate on the substrate surface where it will grow as a single-crystal thin film. Growth rates are of the order of 1 μm per minute. In *fig. 4* a substrate in its holder is shown as it is lowered into a furnace containing the melt.

In this article we shall discuss some of the problems connected with this method of making perfect garnet bubble materials. We shall deal first with the substrate and its possible imperfections, then with the liquid-phase epitaxy itself. Many considerations and parameters such as the specifications drawn up by the device engineer, the composition of the film, its degree of perfection, the composition of the solution from which the film is to grow, its degree of supersaturation, and the growth temperature are all closely interlinked in liquid-phase epitaxy. Important precautions are a tight

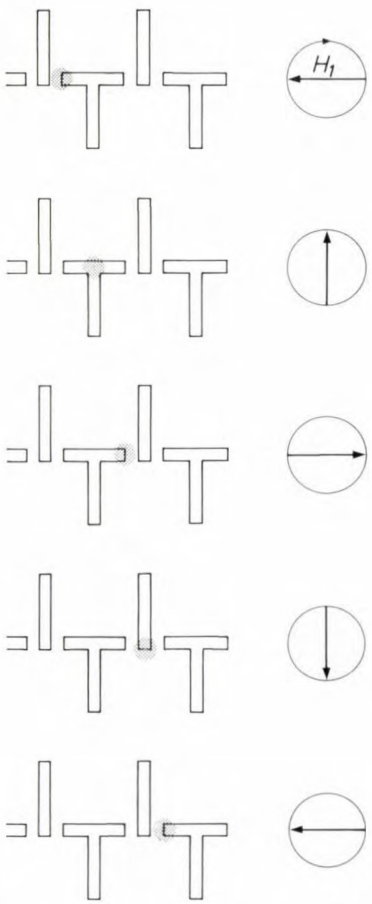


Fig. 2. To move a bubble in a well defined manner through a magnetic plate, strips of permalloy, in the form of tees alternating with bars, are applied to the plate. In a circularly polarized in-plane field H_1 the bubble progresses steadily, always seeking the closest north pole.

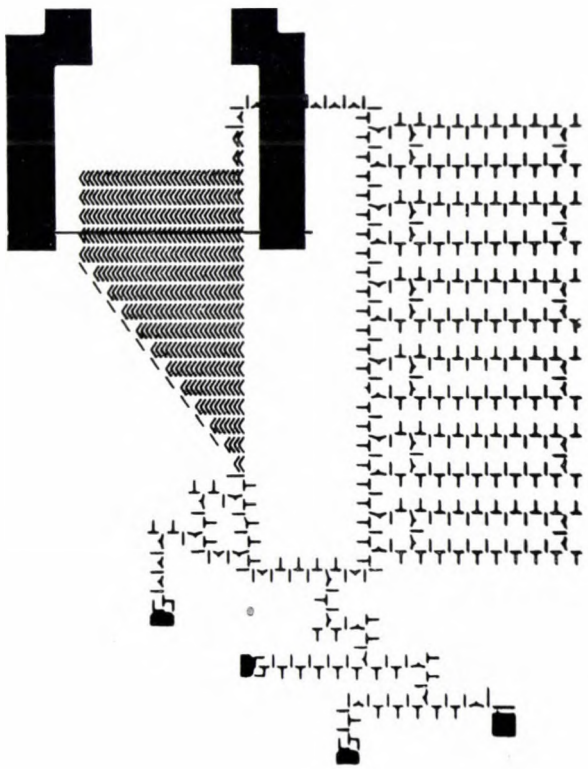


Fig. 3. A laboratory model of a magnetic-bubble memory [3]. The circuit is about 1 × 1 mm. Bubbles are generated from under the permalloy square at lower right, and annihilated at the other 'specks'. They can be manoeuvred along the tracks and into and out of the loops by the alternate use of clockwise and counter-clockwise rotating in-plane fields. The actual memory is in the six loops on the right hand side. The pattern at the upper left is for detecting the bubbles. They are stretched into strip domains by the pattern of 'chevrons' and give a change of resistance in the wire when they pass it.

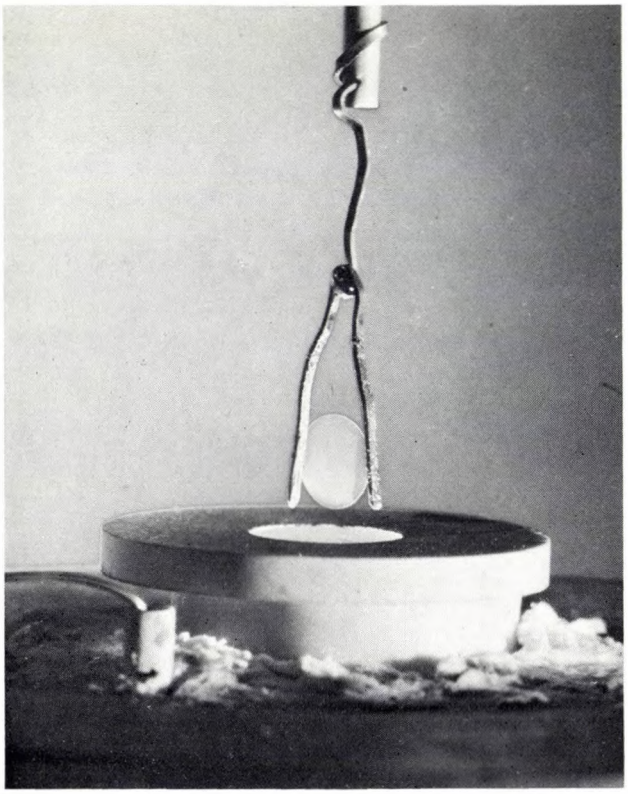


Fig. 4. A platelet of a gadolinium-gallium-garnet single crystal is lowered into a melt, to be covered epitaxially with a magnetic garnet film ('liquid-phase epitaxy').

control of the growth temperature and the convection currents in the melt. One kind of imperfection, that of inhomogeneous incorporation of lead from the solvent in the film, will receive special attention.

Substrate

In general, a perfect film can only be grown on a perfect substrate. Physical defects in the magnetic film such as mechanical strain and dislocations originate primarily in the substrate. Therefore we shall consider the substrate first.

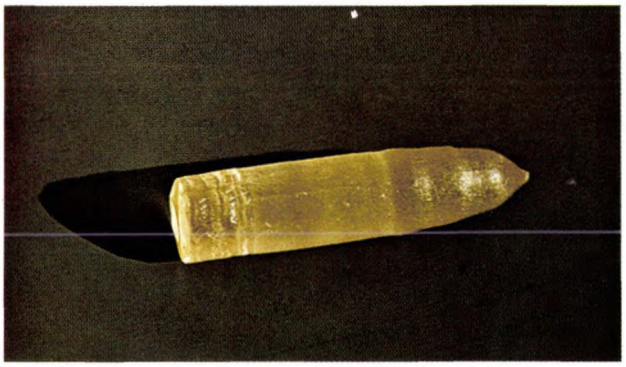


Fig. 5. A boule of gadolinium gallium garnet (GGG), grown by the Czochralski technique.

One widely used substrate is gadolinium gallium garnet, $Gd_3Ga_5O_{12}$. usually called GGG. Single crystals of this material can be grown by the well known Czochralski technique, in which the crystal, starting from a seed, is slowly pulled from the molten material [5]. Boules 2 cm in diameter and several centimetres long are regularly produced in this way (*fig. 5*). Substrate platelets 0.5 mm thick are sawn from the boules in the appropriate orientation, usually (111). The platelets are mechanically polished, etched and cleaned. Some common defects in these crystals are dislocations, facet regions and growth striations. These defects, which we shall now shortly discuss in turn, can all interfere with the epitaxial layer.

Dislocations

A dislocation that intersects the surface is replicated by the epitaxial layer, and may become a pinning point for the bubbles. Dislocations in garnets, however, are no great problem, since dislocations do not readily occur naturally in garnets. The elastic energy associated with a dislocation in a given crystal lattice is proportional to the square of the lattice constant. Now in garnets the lattice constant is 1.24 nm, which is very large compared with that of, say, silicon (0.542 nm). Thus, the nucleation energy of a dislocation in garnet is relatively large. At present, Czochralski-grown GGG is readily prepared with a dislocation density less than 5 cm^{-2}.

Dislocations in garnets are of three types: isolated line defects, closed loops and helices. The helices [6] in particular are of such an intricate structure that we will consider them here briefly. *Fig. 6* is a photograph of an X-ray reflection topograph [7] of a GGG platelet showing a longitudinal section of a giant dislocation helix. The diameter of the helix is 250 μm. The platelet is a longitudinal section of a boule cut parallel to the [111] axis. *Fig. 7a* is an X-ray transmission topograph of several of the helices viewed end-on, along their axes. The platelet thickness is just sufficient to contain one or two turns of the helix. A radial spoke structure is also visible. In the reflection topograph of fig. 7*b* this structure only appears in a triangular sector inside the helices. This implies that the spoke structure lies on the helicoidal surface of the helix; the reflection topograph reveals only the portion of the helicoid that rotates down through the X-ray penetration depth (about 5 μm in this particular case). Thus the total defect is rather like a helical staircase with the dislocation as the handrail and the radial spokes corresponding to the steps.

[3] W. F. Druyvesteyn, F. A. Kuijpers, E. de Niet and A. W. M. van den Enden, IEEE Trans. **MAG-10**, 757, 1974 (paper presented at Intermag Conference, Toronto 1974).
[4] H. J. Levinstein, S. Licht, R. W. Landorf and S. L. Blank, Appl. Phys. Letters **19**, 486, 1971.
[5] See for example J. Goorissen, Philips tech. Rev. **21**, 185, 1959/60.
[6] J. W. Matthews, E. Klokholm and T. S. Plaskett, AIP Conf. Proc. **10**, Part 1, 271, 1973.
W. T. Stacy, J. A. Pistorius and M. M. Janssen, J. Crystal Growth **22**, 37, 1974.
[7] A. E. Jenkinson, Philips tech. Rev. **23**, 82, 1961/62.

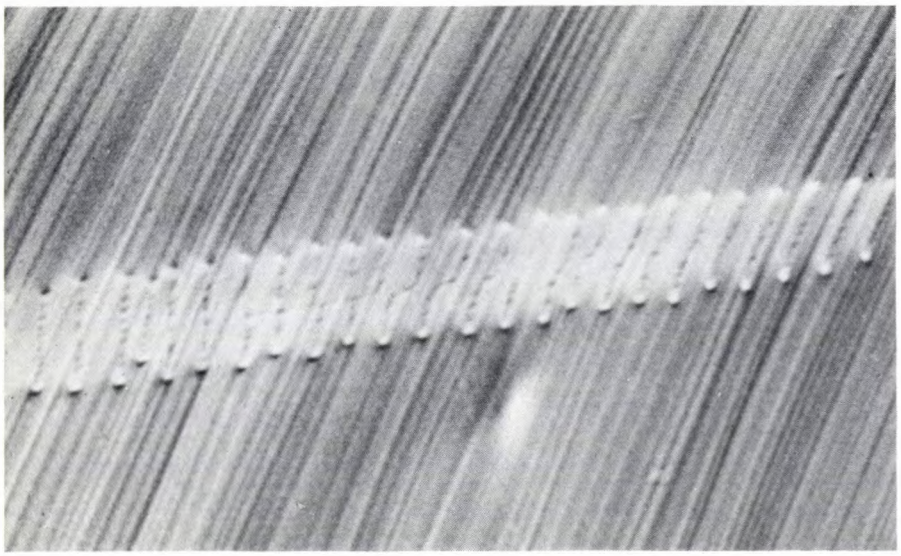


Fig. 6. X-ray reflection topograph of a GGG platelet showing a dislocation helix in longitudinal section. The helix has a diameter of 250 μm.

The mechanism responsible for the large helices is believed to be the precipitation of a second chemical phase that occurs preferentially at the dislocation [8]. A colony of precipitate particles that form inside the helix as the dislocation grows is considered to be the source of the radial-spoke contrast.

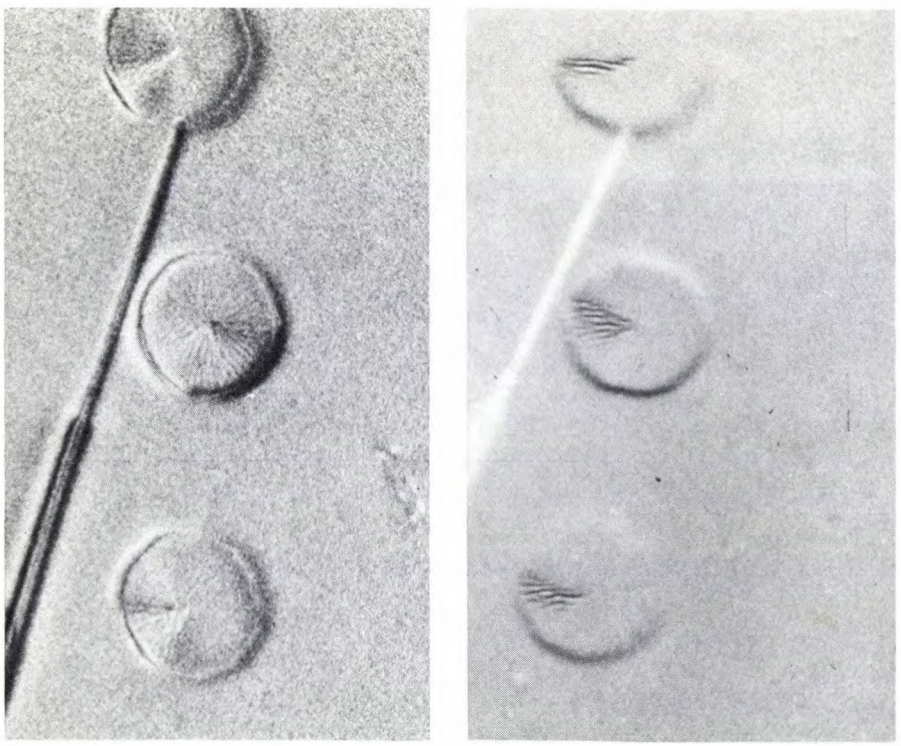
Domain-wall pinning due to the helices has been reported [9].

Facet regions

When the interface of a crystal with the melt is too curved during the growth of the crystal, *facets* appear in the interface at places where it would be tangential to certain crystallographic planes. Garnet forms (112) and (110) facets. They develop into 'facet regions' in the boule; growth striations in these regions are thus perfectly straight [10] (see *fig. 8a*). Facet regions in a platelet of GGG can be detected by X-ray topography or microscopy with polarized light.

A facet region in a substrate platelet interferes with the epitaxial film because it has a slightly larger lattice constant than the remainder of the crystal [11]. This causes the epitaxial layer to deform slightly to match this change; the result is a region with different magnetic properties [12]. Fig. 8*b* and *c* show some typical facet regions, and demonstrate the local deviation in the lattice constant.

Like dislocations, facet regions do not form a serious problem today. They can be completely eliminated



a *b*

Fig. 7. *a*) Transmission topograph, *b*) reflection topograph of helices viewed end-on; (*a*) comprises several turns of the helix, (*b*) only part of one rotation. This explains why the spoke structure, assumed to lie in the helicoidal surface of the helix, is shown around the complete circle in (*a*), but only in a triangular section in (*b*).

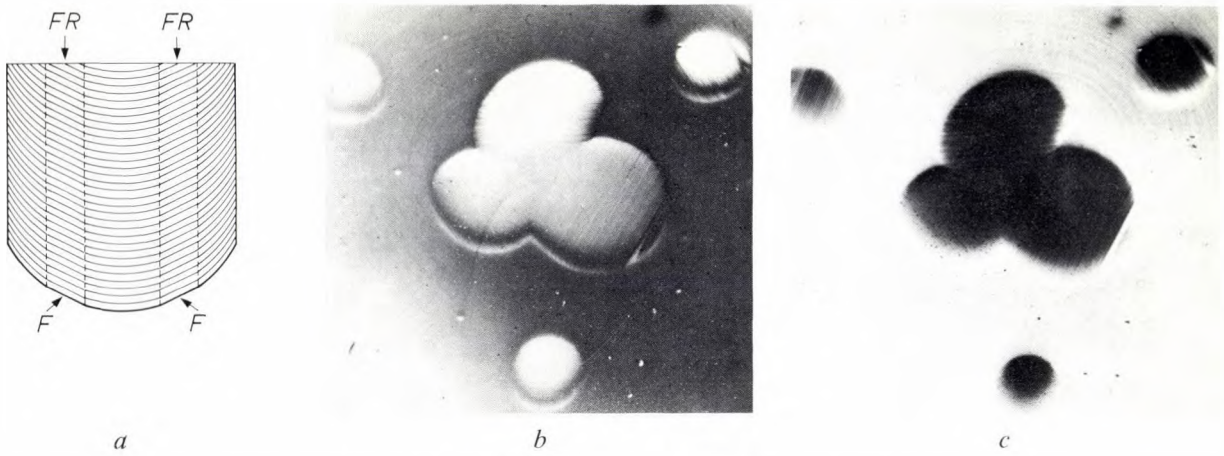


Fig. 8. Facet regions. *a*) Longitudinal section of a boule, schematic. *F* facets, *FR* facet regions characterized by perfectly straight growth striations. *b*) and *c*) Double crystal-reflection topographs of a platelet of GGG containing facet regions, showing that these regions differ in lattice constant from the remainder of the crystal. When regions appear dark, they are strongly diffracting, and the lattice constant of these regions can be deduced from the orientation of the platelet. The parts that appear dark in (*c*) are facet regions. The difference in orientation of the platelet between the two cases (52″) indicates a relative difference in lattice constants of 1.5×10^{-4}.

from the bulk of the boule by ensuring that the growth interface is nearly flat. This is done by carefully adjusting the pulling rate and the rate of rotation of the crystal about the vertical axis.

Growth striations

Growth striations form the one type of imperfection common to all GGG crystals currently available. The exact origin of growth striations in garnets is unknown. As in germanium and silicon, they are usually attributed to growth-rate fluctuations that cause localized compositional changes, which in turn lead to detectable bands of strain. They thus form a record of the history of the shape of the growth interface.

The striations would obviously have no influence on the epitaxial film if they were flat and parallel to the substrate surface. But perfectly flat striations can only be obtained on certain facet planes (the natural facet faces) and, in the case of garnets, never in the (111) or (100) planes preferred for bubble-domain substrates. *Nearly* flat striations that are *nearly* parallel to the

[8] W. T. Stacy *et al.* [6].
E. Nes, Scripta met. **7**, 705, 1973.
[9] W. T. Stacy and U. Enz, IEEE Trans. **MAG-8**, 268, 1972.
[10] J. A. M. Dikhoff, Philips tech. Rev. **25**, 195, 1963/64.
[11] H. L. Glass, Mat. Res. Bull. **7**, 1087, 1972.
B. Cockayne, J. M. Roslington and A. W. Vere, J. Mat. Sci. **8**, 382, 1973.
[12] H. L. Glass, P. J. Besser, T. N. Hamilton and R. L. Stermer, Mat. Res. Bull. **8**, 309, 1973.

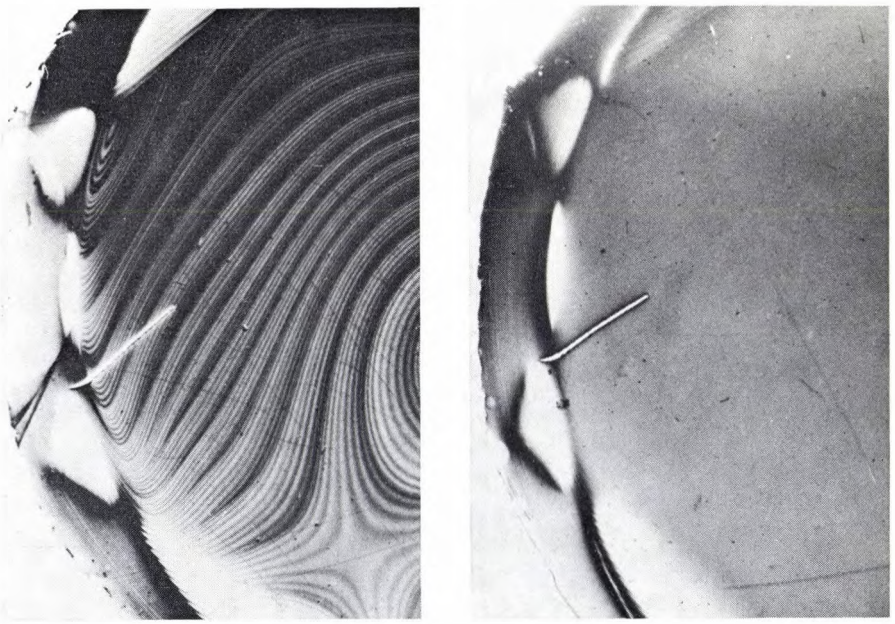


Fig. 9. X-ray reflection topographs of a GGG platelet covered with an epitaxial film, *left*: substrate reflection, *right*: film reflection. The shape of the pattern of striations in the left-hand picture indicates that the growth faces giving rise to the striations are nearly parallel to the platelet surface. Because of this, they do not deform the film, as is shown by the right-hand picture. Deformations in the film due to facet regions in the substrate are apparent on the left of the right-hand picture.

substrate surface, however, hardly affect the epitaxial film either, as is shown in *fig. 9*. This is because the film is sensitive to in-plane strains only, whereas the maximum displacement associated with a striation is perpendicular to its plane. Nearly flat striations mean a nearly flat growth interface, which is required in any case to prevent facet regions from occurring, as we saw above.

The growth technique; the solvent

The apparatus we use for growing garnet films by liquid-phase epitaxy is shown schematically in *fig. 10*. The melt is contained in a platinum crucible in a very precisely controlled constant-temperature zone in a furnace at about 900 °C. When equilibrium has been established in the melt, the substrate is slowly lowered into the furnace. It is held for a few minutes just above the surface of the melt to make sure that it will have the same temperature, then quickly immersed in the melt. After the required period of time for film growth the specimen is removed from the melt.

The key to the success of the whole technique is the ability to cool the solution of the garnet constituents to more than 20 °C below its saturation temperature without the spontaneous nucleation of garnet crystals or other phases in the melt. One of the most successful solvents making this possible has been PbO/B_2O_3 in a 50 : 1 weight ratio [13]. It has the disadvantages of being corrosive, poisonous and volatile. The apparatus must therefore be provided with a fume extractor.

The solution must also have a low viscosity at the growth temperature, so that no droplets adhere to the sample when it is removed. These would result in mesas and etch pits on the film, spoiling its surface. This requirement of low viscosity is also met when PbO/B_2O_3 is used as a solvent.

The substrate is usually in a horizontal position when dipped into the melt. It is then rotated at a speed of, say, 30 rev/min during growth. After removal from the melt the speed is increased, to spin off any remaining flux [14]. Experiments have also been made both at Philips Research Laboratories and at Philips Forschungslaboratorium Hamburg (PFH) with the sample in the vertical position, without rotation during film growth [15]. This method has the advantage of being much better suited to dipping a few (and in the future perhaps many) substrates together.

With vertical dipping, however, more attention has to be paid to the problem of convection currents in the melt. If these are not controlled, unacceptable variations in thickness of the film all over the sample are bound to arise. It has been shown at Philips Research Laboratories that variations of more than 5% (0.25 μm in a 5-μm film) interfere with the movement of the bubbles [16]. Using furnaces with three heating zones and baffle systems (*B* in fig. 10), we have been able to control the convection currents adequately. The effect can be seen by comparing *fig. 11a* and fig. 11*b* showing interference fringes due to variations in thickness. In growing the film of fig. 11*a*, the convection currents were controlled, in that of fig. 11*b* they were not.

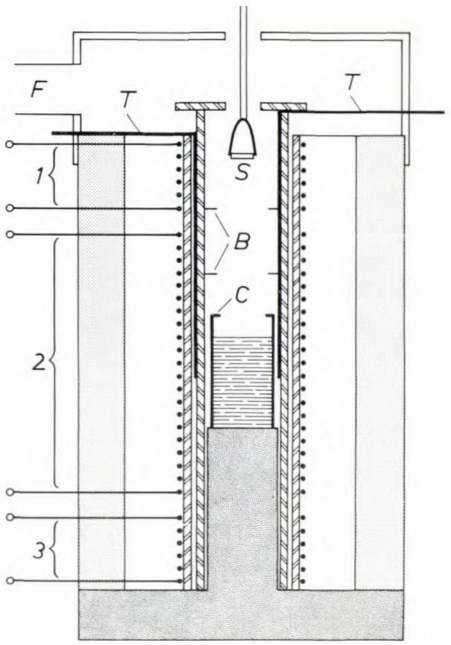


Fig. 10. Diagram of furnace for liquid-phase epitaxy. The substrate platelet *S*, to be dipped in the melt in the platinum crucible *C*, is drawn in the horizontal position; a vertical position has also been used. Temperature and temperature gradients are controlled by three separate heating zones *1,2,3*. *T* thermocouples. Convection currents in the vapour are reduced by platinum baffles *B*. With the help of the triple-zone heating and the baffles the convection currents in the melt can effectively be controlled. A fume extractor (*F*) is a necessity, since the melt is a corrosive volatile poison.

The choice of the growth temperature T_g is closely linked up with the composition of the melt. These two quantities determine the 'degree of supersaturation' ΔT, i.e. the difference between T_g and the saturation temperature T_s of the solution. But the crystal grower's choice of oxides to go into the melt can of course only be made after he has decided the composition of the film. This problem will be discussed first.

The required chemical composition of the film

The composition of the film is governed by the requirements of the device engineer. For a magnetic bubble memory with bubbles of 2-5 μm in diameter, for instance, a material is required with a saturation magnetic polarization J_s of about 15 mT (in the Gaussian system: saturation magnetization $4\pi M_s$ of about 150 gauss). Moreover, this parameter must be fairly

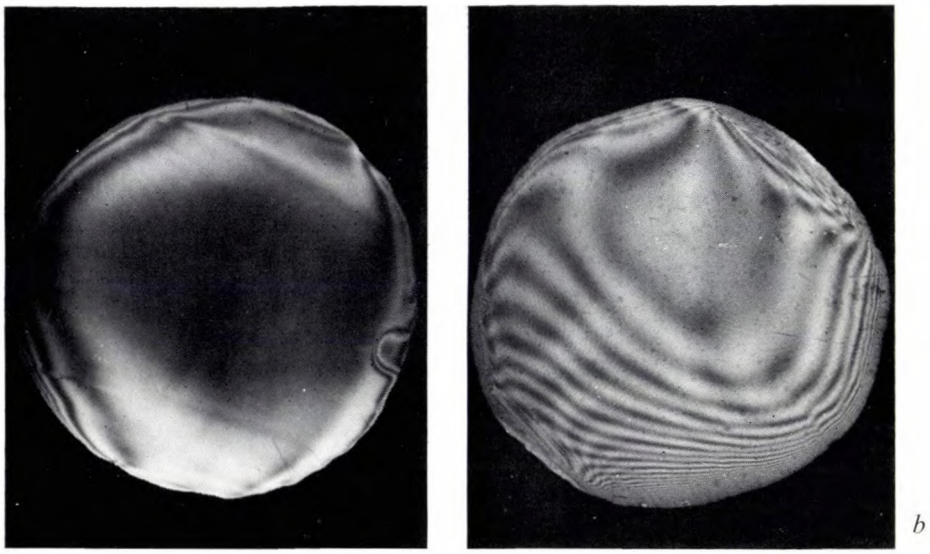


Fig. 11. Effect of the control of convection currents in the melt. The two samples, one (*a*) prepared with and the other (*b*) without convection-current control, are illuminated with sodium light. The interference fringes in (*b*) indicate a much larger variation in thickness than in (*a*).

temperature-independent in the operating range of 0-80 °C.

The required temperature-independence is obtained by a method developed several years ago for microwave garnets. If for example in YIG the yttrium ion (occupying the dodecahedral sites in the crystal) is replaced by some other rare-earth ion, e.g. gadolinium, then below the Néel temperature T_N a compensation temperature T_{cp} can be observed at which the polarization again goes through zero (*fig. 12*). The value of T_{cp} depends on the type of rare-earth ion and its concentration. At the correct composition T_{cp} and T_N will be evenly spaced about the room temperature. The J_s,T curve will then be more or less flat in the region of room temperature. Usually, however, the room temperature value of J_s of such mixed iron garnets is too high. This can then be remedied by diluting the garnet magnetically, i.e. replacing some of the iron ions (occupying octahedral and tetrahedral sites in the crystal) by nonmagnetic ions such as gallium or aluminium. Replacement of iron ions also affects T_{cp}, and is in fact also used to obtain a temperature independent J_s.

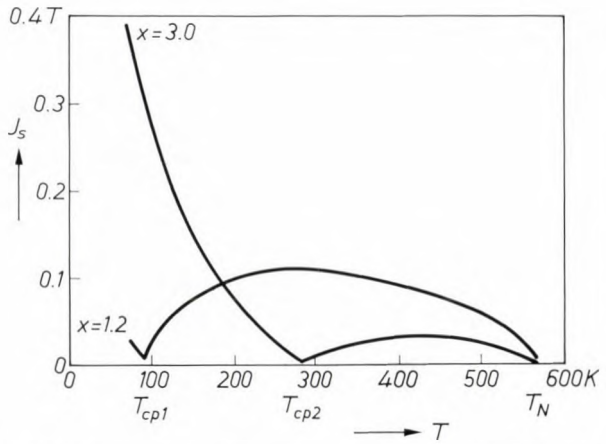
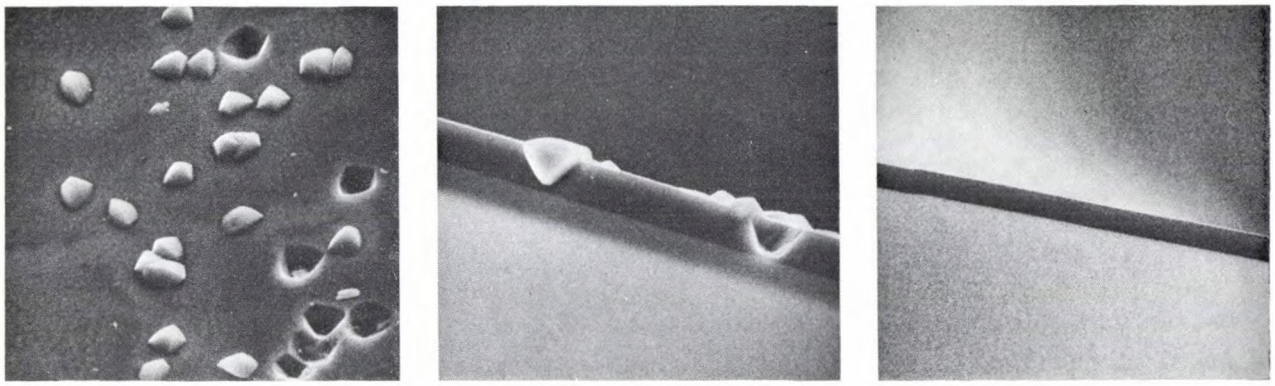


Fig. 12. The saturation magnetic polarization J_s of a ferrite can be zero at two separate temperatures. The higher one is the Néel (or Curie) temperature T_N. The lower is the compensation temperature T_{cp}, where the magnetizations of the magnetic sublattices compensate one another exactly. In iron garnets T_{cp} can be varied strongly by substituting different ions at the dodecahedral sites. As an example the figure shows the J_s,T curves of $Gd_xY_{3-x}Fe_5O_{12}$ for $x = 1.2$ and $x = 3.0$. By choosing T_{cp} as far below room temperature as T_N is above it, the polarization can be made nearly temperature-independent in the room-temperature range.

The film must also have a uniaxial anisotropy, giving the magnetization the required preference for the direction perpendicular to the film. This can be achieved by using 'growth-induced anisotropy'. An anisotropy of this kind and of the right order of magnitude can be obtained when rare-earth ions of different radii are combined in the film [17]. Moreover, if S-state ions are chosen such as La, Y or Gd, the bubbles in the film have the highly desirable property of being extremely mobile. The origin of the growth-induced anisotropy is not yet fully understood [18].

A film may also have a 'stress-induced' anisotropy. This occurs when the lattice constants a_f and a_s of film and substrate are not quite equal [19]. The difference $\Delta a = a_s - a_f$ must however remain smaller than 0.002 nm if crack-free films are to be obtained [13].

[13] S. L. Blank and J. W. Nielsen, J. Crystal Growth **17**, 302, 1972.
[14] E. A. Giess, J. D. Kuptsis and E. A. D. White, J. Crystal Growth **16**, 36, 1972.
[15] W. Tolksdorf, G. Bartels, G. P. Espinosa, P. Holst, D. Mateika and F. Welz, J. Crystal Growth **17**, 322, 1972.
[16] W. F. Druyvesteyn, J. appl. Phys. **46**, 1342, 1975.
[17] J. W. Nielsen, S. L. Blank, R. L. Burns and W. A. Biolzi, paper presented at the AIME Conference on Electronic Materials, Las Vegas 1973.
[18] E. M. Gyorgy, M. D. Sturge, L. G. Van Uitert, E. J. Heilner and W. H. Grodkiewicz, J. appl. Phys. **44**, 438, 1973.
A. Akselrad and H. Callen, Appl. Phys. Letters **19**, 464, 1971.
W. T. Stacy and C. J. M. Rooymans, Solid State Comm. **9**, 2005, 1971.
[19] P. J. Besser, J. E. Mee, P. E. Elkins and D. M. Heinz, Mat. Res. Bull. **6**, 1111, 1971.



a *b* *c*

Fig. 13. Surface (*a*) and cross-section (*b*) of a film that has been grown at too high a degree of supersaturation. As a result individual crystals have formed on the film. *c*) Cross-section of a film made at a lower degree of supersaturation. These photomicrographs were made with a scanning electron microscope; magnification 2000×. (Photograph J. L. C. Daams.)

Therefore only a limited use can be made of stress-induced anisotropy. In fact, it is often decided not to use it at all, and to make Δa as small as possible. This is our last main requirement for the film composition: its lattice constant should fit that of the chosen substrate.

We should note here that with gallium-arsenide epitaxy, for example, a slight lattice misfit between film and substrate very easily leads to the formation of 'misfit-dislocations' in the film. This is not the case with garnet epitaxy, since, as noted earlier, dislocations do not easily occur in garnet owing to its large lattice constant (see p. 3).

A film composition that we have found to satisfy the requirements discussed above, for bubbles of 2-5 μm in diameter and to be combined with GGG as a substrate, is $Y_{2.9}La_{0.1}Fe_{3.8}Ga_{1.2}O_{12}$.

Growth temperature and composition of the melt

First of all, we should mention that there is an upper limit to the growth temperature. It must not be too high, because of the possibility of *interdiffusion* between iron from the magnetic layer and gallium from the substrate, which would spoil the sharpness of the interface. It has been found that below 1000 °C this diffusion is sufficiently slow for a perfectly sharp interface to be obtained if the substrate is GGG. With other substrates such as $Sm_3Ga_5O_{12}$ or $Nd_3Ga_5O_{12}$ the growth temperature should be kept even lower, to avoid this risk. As a general rule, the larger the ionic radius of the rare-earth ion in the substrate garnet, the easier for the diffusion to take place and hence the lower the value that should be chosen for the growth temperature.

Secondly, the degree of supersaturation ΔT should not be more than about 20 °C. If it is higher, spontaneous nucleation of garnet crystals may occur. Such individual crystals in the film itself will of course spoil it. *Fig. 13a* shows the surface, and fig. 13*b* the cross-section, of a film containing such crystals; the perfect film of fig. 13*c* was made at a lower value of ΔT. However, even if the spontaneous nucleation only took place in the melt away from the film, or on the walls of the crucible, it would still be highly undesirable because it would change the composition of the melt. Thus ΔT, and therefore the growth rate, would be affected. This cannot be tolerated if films of uniform thickness are to be made reproducibly.

From these limitations on T_g and ΔT, it follows that the saturation temperature T_s is also limited, to around 1000 °C. This would imply that the amount of oxides in the melt must not be more than about 6 mol%. In practice, we use concentrations of 1 to 2 mol%, resulting in values of T_s and T_g in the region of 700-900 °C. At these lower growth temperatures the film thickness is more easily reproduced [20].

Having decided upon the chemical composition of the film and the growth temperature, the crystal grower would be able to calculate directly the required amount of oxides in the melt, if he knew the *segregation coefficient* of each garnet constituent, i.e. the ratio of its concentration in the melt to that in the film, at the growth temperature. In the earlier work on garnet liquid-phase epitaxy, such data were only available for bulk single-crystal growth, which is normally performed at about 1100 °C. For the much lower temperatures required in liquid-phase epitaxy, the data could only be guessed by extrapolation. From such extrapolation it can be seen that the range of compositions that will lead to the crystallization of the garnet phase only, without any simultaneous nucleation of magnetoplumbite or orthoferrite, is very narrow (*fig. 14*). In the meantime several

segregation coefficients have also been obtained for the temperature range in which liquid-phase epitaxy is performed [22].

Lead incorporation in garnet films

A disadvantage of any method of growing crystals from a solution is the incorporation of solvent ions in the crystal. Garnet films, as prepared in the way discussed in this article, always contain some lead [23]. The Pb^{2+} ions are grown into the dodecahedral sites of the lattice, i.e. the sites normally occupied by rare-earth ions [24]. This fact is of importance in the first place because the Pb^{2+} ion, being a very large ion, usually increases the lattice constant of the film. It

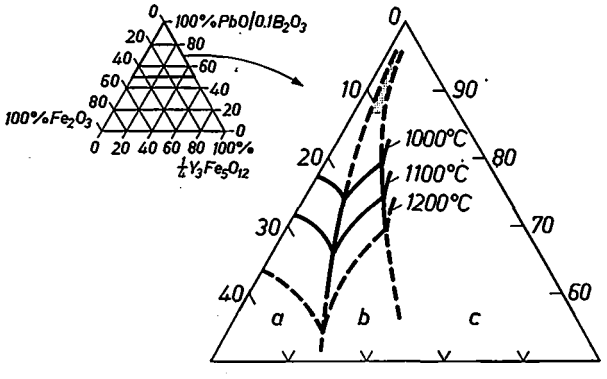


Fig. 14. Upper part of the phase diagram of the system $PbO/0.1B_2O_3$-Fe_2O_3-$\frac{1}{3}Y_3Fe_5O_{12}$, after H. D. Jonker [21]. The solid lines represent the molar compositions of the melt at which the saturation temperature has the indicated value, and also the boundaries between the regions where magnetoplumbite (*a*), garnet (*b*) and orthoferrite (*c*) will crystallize from the melt when the temperature is lowered below the saturation temperature. These data represent what is known from bulk crystal growth, which usually takes place at temperatures above 1000 °C. Extrapolation of the boundaries between *a*, *b* and *c* shows that there can only be a very narrow region (shown grey) from which garnet can be obtained at temperatures well below 1000 °C, as required for liquid-phase epitaxy.

therefore influences the choice of the film composition which, as we have seen before, must be such that the film and substrate lattices match. Secondly, there is evidence that the growth-induced anisotropy is related to the incorporation of lead.

At Philips Research Laboratories the incorporation of lead in garnet films and its effect upon the magnetic properties of the film have been studied in some detail, using the method of chemical analysis by radioactive tracers [25]. This technique is useful because it can distinguish ions in the film, e.g. Ga or Gd, from the same kind in the substrate. This makes accurate chemical analysis of very thin films (even down to 0.1 μm) possible where other methods (electron microprobe, spectrochemical analysis) would be confused by the presence of the substrate.

For this work special long-lived isotopes were used so that the measurements could be performed over a period of weeks. These isotopes were made at the Netherlands reactor centre (RCN) in Petten.) The isotopes were added to the melt, and after the specimen was prepared in the normal way, its radioactivity was measured. Up to five elements in one film could be analysed simultaneously with the spectrometer and analyser that we used for this work.

In this way we have studied in detail how the amount of lead in the film depends upon the growth temperature (T_g) and the growth rate (R_g). Our results suggest that less lead will be incorporated if T_g is higher (as has been reported before [26]) and if R_g is lower [27]. Thus, to prevent the incorporation of lead in the films, a *high* T_g and a *low* R_g should be used. Now a relatively *low* T_g should be used if film thickness is to be reproduced accurately, as we have seen above. Our results imply, however, that the low T_g thus required could be compensated in its effect upon the lead content by the use of a very low growth rate. We have indeed succeeded in making good-quality film, using growth rates that range from 0.1 μm/min to 0.1 μm/hour [28].

We have also studied the concentration of lead as a function of film thickness. The results are related to the problem that the thickness of a film as measured by optical interference always differs slightly but systematically from the result obtained by mechanical measurements. It has been pointed out by R. Ghez and E. A. Giess that this can be explained by a change of refractive index in the film because of the presence of lead [29]. This is confirmed by our measurements. We have measured the lead content in films of YIG grown on substrates of GGG, of widely varying thicknesses, by using widely varying deposition times.

[20] R. Ghez and E. A. Giess, J. Crystal Growth **27**, 221, 1974.
[21] H. D. Jonker, J. Crystal Growth **28**, 231, 1975.
[22] G. A. M. Janssen, J. M. Robertson and M. L. Verheijke, Mat. Res. Bull. **8**, 59, 1973. See also the articles of notes [13] and [15].
[23] E. A. Giess, B. E. Argyle, D. C. Cronemeyer, E. Klokholm, T. R. McGuire, D. F. O'Kane, T. S. Plaskett and V. Sadagopan, AIP Conf. Proc. **5**, Part 1, 110, 1972.
[24] J. M. Robertson, S. Wittekoek, Th. J. A. Popma and P. F. Bongers, Appl. Phys. 2, 219, 1973.
[25] G. A. M. Janssen *et al.* [22].
[26] S. L. Blank, B. S. Hewitt, L. K. Shick and J. W. Nielsen, AIP Conf. Proc. **10**, Part 1, 256, 1973.
[27] J. M. Robertson, M. J. G. van Hout, J. C. Verplanke and J. C. Brice, Mat. Res. Bull. **9**, 555, 1974.
[28] S. Wittekoek, T. J. A. Popma and J. M. Robertson, AIP Conf. Proc. **18**, Part 2, 944, 1974.
[29] R. Ghez and E. A. Giess, Mat. Res. Bull. **8**, 31, 1973.

The result of this work, for two experiments in which different values of T_g were employed, is shown in *fig. 15*. It can be seen that the lead concentration depends upon film thickness and is greatest for the thinnest films. It therefore follows that in all of the films the lead concentration is likely to be greatest near the interface.

This fact may be explained by supposing that just after immersion of the substrate into the melt a transient period is required for the film growth to attain stability. In this period the deposition of lead could easily be greater than normal. The extent to which this 'imperfection' of the film affects the behaviour of the bubbles, if at all, has yet to be investigated. We have found that the transient period can be shortened by using higher growth temperatures, and also by rotation of the sample during growth.

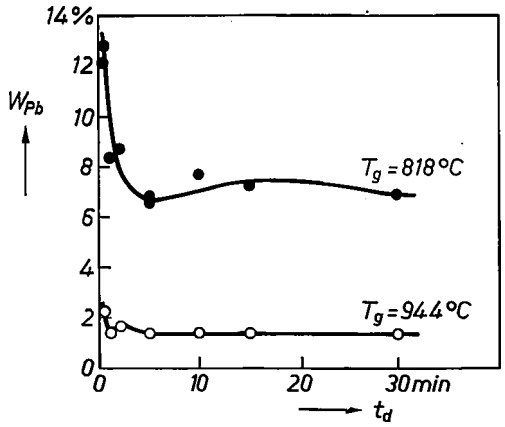


Fig. 15. Percentage weight of lead (W_{Pb}) in YIG films grown on GGG substrates, as a function of the deposition time t_d, at two growth temperatures. The films were analysed with the aid of radioactive tracers.

In conclusion, then, we can sum up by saying that we have made considerable progress in our knowledge of the various factors that influence composition, uniformity of thickness and perfection of films for magnetic bubbles. Taking full account of all these factors we are now able to produce films with properties as required for magnetic-bubble devices. However, several questions remain unanswered, and further investigations, in close cooperation with PFH, are therefore being made.

Summary. Magnetic-bubble memories require high-quality films with special magnetic properties. The most promising method of making such films is to grow an iron garnet single-crystal layer epitaxially on a single-crystal substrate platelet of a nonmagnetic garnet such as gadolinium gallium garnet. Substrate platelets are cut from boules grown by the Czochralski technique. The common defects in substrate crystals are dislocations, facet regions and growth striations. The first two present no major difficulties today, and growth striations are harmless when parallel to the substrate face. Dislocations may have an intricate helical structure. In liquid-phase epitaxy the substrate is dipped in a melt, usually of PbO/B_2O_3 (50 : 1), supersaturated with the required garnet oxides. The convection currents in the melt must be tightly controlled to obtain films of uniform thickness. The composition of the film is governed by the requirements of a temperature-independent saturation magnetization $4\pi M_s$ of, say, 150 gauss, a certain uniaxial anisotropy, and a fit of the lattice constant of the film to that of the substrate. The quantity of oxides in the melt is limited to about 6 mol%, because the growth temperature must remain below 1000 °C (to prevent interdiffusion between iron in the film and gallium in the substrate), and the degree of supersaturation within some 20 °C (to prevent the spontaneous nucleation of crystals). In practice concentrations of 1 to 2% are used. Some lead from the solvent is always incorporated in the film. It affects the lattice constant and the anisotropy. The quantity and distribution of the lead has been studied by chemical analysis with radioactive tracers. The results suggest that little lead is incorporated when the growth temperature is high and the growth rate is low. They also show that the lead concentration is greatest near the interface with the substrate.

The 'push-pull' spiral-groove bearing — a thrust bearing with self-adjusting internal preloading

The family of spiral- and helical-groove bearings [1–3] developed by Philips has recently been extended by a new type: a bearing that can take up thrust in either direction along the axis without the need for external preloading. We call it the 'push-pull' bearing.

As the earlier articles have shown, the grooves in a spiral-groove bearing pump the lubricant — often just the ambient air — between the moving surfaces, producing an overpressure and hence an upward thrust or lifting force (*figs. 1a,b*). The bearing surfaces consequently take up a position with a spacing that gives equilibrium between the lift and the load. The load capacity of the bearing increases as the speed of rotation is increased.

In the new bearing the moving surfaces are provided with two sets of grooves [4], rather than a single set. The two sets of grooves are of opposite sense and are quite different in size; the depths of the grooves, for example, differ by a factor of about five. Because the spirals are of opposite sense, the lubricant provides an overpressure in one part of the bearing and an underpressure in another part (fig. 1*c*). The underpressure creates an internal 'load' that balances the lift at a particular spacing between the surfaces. By paying careful attention to the differences between the two patterns of grooves, it is possible to make the stiffness of the bearing sufficiently large even though no external load is applied. The dimensions for the grooves can be calculated from the theory of spiral-groove bearings[5].

Fig. 2 shows an interesting possible application for such a bearing. This is a rotor with gas bearings that would be suitable for a gyroscope; since the two sets of spirals produce a push-pull action, only one thrust bearing is required. This has the great advantage that no serious problem is introduced by the thermal expansion of the rotor, which can often be a nuisance when thrust bearings are mounted at both ends.

Another kind of application, which we shall discuss here to give a better understanding of the design of the new bearing and its stiffness, is in a device for making

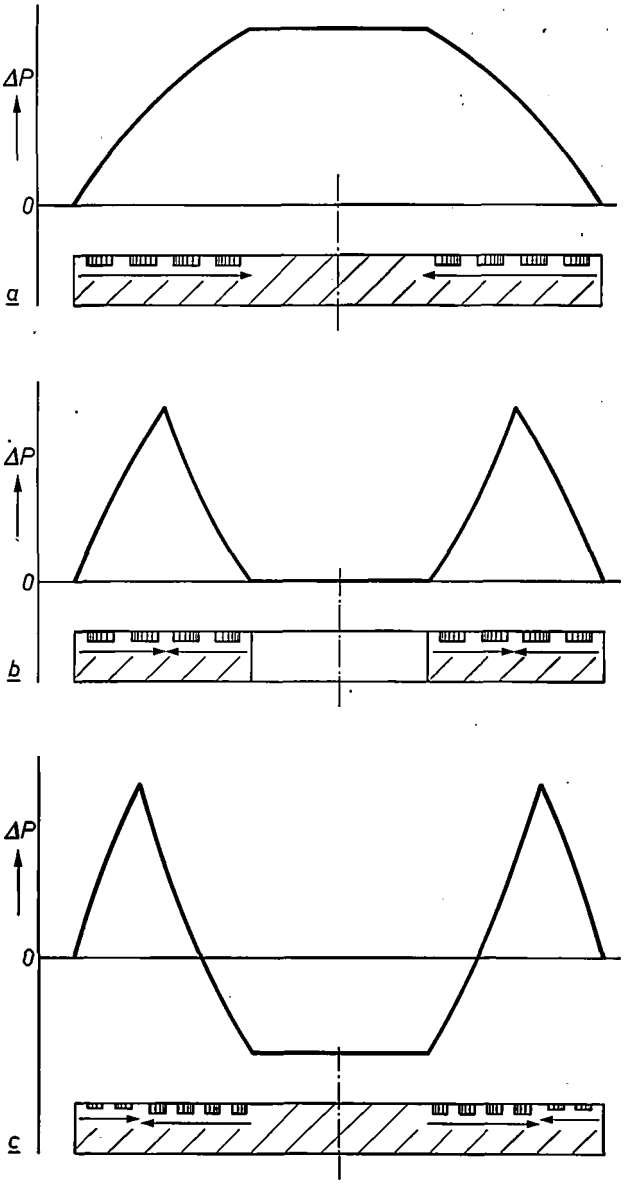


Fig. 1. *a*) The pressure distribution (ΔP is the overpressure) in the radial direction over the surface of a conventional spiral-groove bearing. The arrows indicate the direction in which the lubricant is pumped. In the central region where there are no grooves the pressure is uniform. *b*) The pressure distribution in a herringbone spiral-groove bearing. This type of bearing has a central opening; to obtain sufficient lift the lubricant is pumped between the bearing surfaces from both the outer edge and the inner edge. *c*) The pressure distribution in the new bearing — the 'push-pull' spiral-groove bearing. Besides the overpressure there is now an underpressure as well, generated by a second set of spiral grooves. In the absence of an external load, the bearing surfaces automatically adopt a spacing such that the forces set up by these pressures balance each other. When the spiral grooves are correctly designed, such a bearing has a high stiffness in the axial direction both for compression (push) and for tension (pull). This stiffness exists in spite of the absence of an external load. The three pressure distributions shown include parabolic curves because the radial pressure gradients occurring are proportional to the tangential velocity, i.e. proportional to the radial distance from the axis.

[1] E. A. Muijderman, New forms of bearing: the gas and the spiral groove bearing, Philips tech. Rev. 25, 253-274, 1963/64.
[2] G. Remmers, Grease-lubricated spiral groove bearing for a straight-through shaft, Philips tech. Rev. 27, 107-108, 1966.
[3] G. Remmers, Grease-lubricated helical-groove bearings of plastic, Philips tech. Rev. 34, 103-105, 1974.
[4] The first version of such a bearing was designed by J. P. Reinhoudt in 1972 at Philips Research Laboratories, Eindhoven.
[5] E. A. Muijderman, Spiral groove bearings, Philips Technical Library, Eindhoven 1966.

laser light incoherent (*fig. 3*). It consists of two discs of frosted glass, spaced by about 3 μm, which rotate in air at a relative speed to one another of about 1000 rev/min. As required for this application, the spacing remains very nearly constant when the speed varies. The device is used at Philips Forschungslaboratorium Hamburg in an equipment that writes an image on a projection screen by digital deflection of a laser beam [*]. If the laser light is not made incoherent, 'speckles' of light appear in the image.

The discs run at such an extremely small spacing because of the presence of patterns of spiral grooves on the two facing surfaces (*fig. 4*). These surfaces function as the moving surfaces of a push-pull bearing.

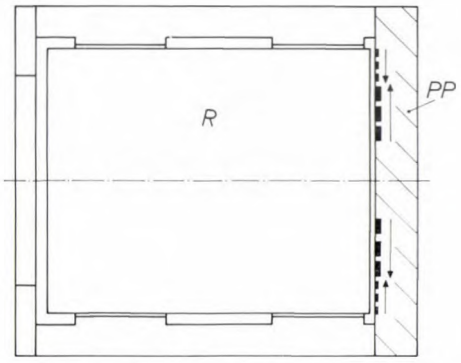


Fig. 2. Diagram of a thrust bearing for a cylinder (*R*), such as the rotor of a gyroscope, with a push-pull spiral-groove bearing (*PP*). One spiral-groove bearing is sufficient for the axial location of *R*.

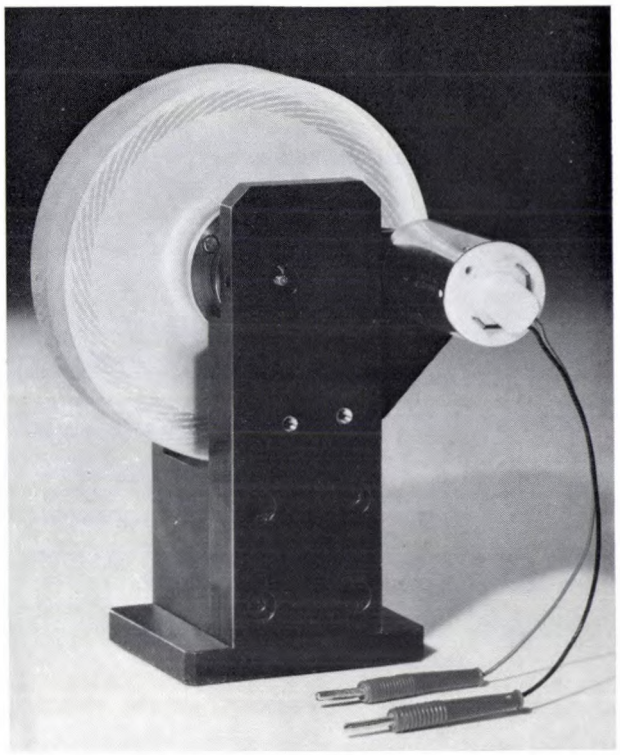


Fig. 3. Device that makes laser light incoherent. This is used at Philips Forschungslaboratorium Hamburg in an arrangement for projecting images by a digitally deflected laser beam [*]. The device consists essentially of two frosted-glass discs (diameter 100 mm) forming a push-pull bearing, one of which is rotated at about 1000 rev/min at a spacing of only 3 μm from the other. This spacing is almost independent of the speed of rotation. The drive motor is on the right. The spacing can adjust itself because the drive is transmitted from the motor to the glass disc by a thin metal diaphragm (not visible).

One of the discs remains stationary; the other is connected to the drive motor by a thin metal diaphragm, to give sufficient freedom in the axial direction. The diaphragm also ensures the radial location of the disc. The annular region between the inner set of spiral grooves and the diaphragm acts as a window for the laser light.

The underpressure around the centre provides the preload; in the surrounding region there is an overpressure that provides the lift. *Fig. 5* shows the varia-

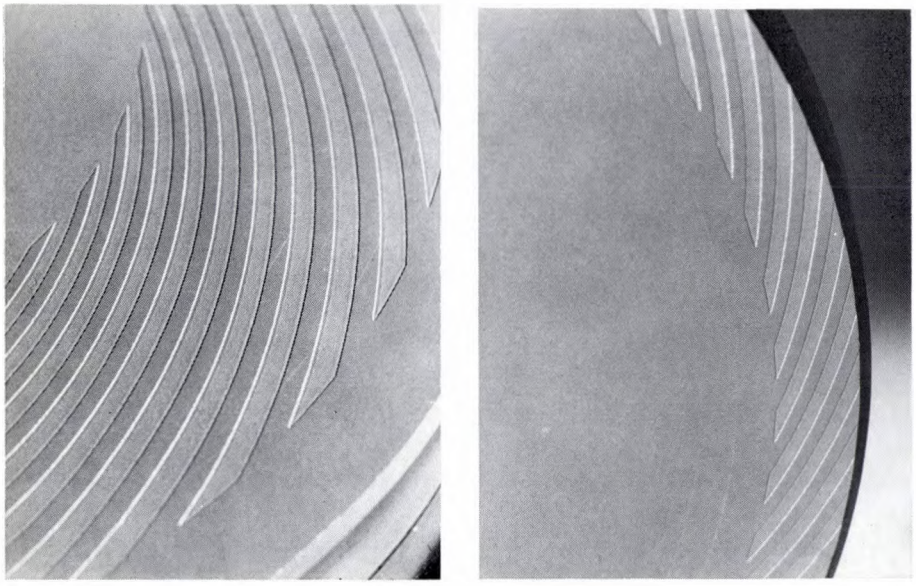


Fig. 4. Enlarged photographs of parts of the two sets of grooves etched in the glass discs of the device shown in fig. 3. The grooves shown on the left rotate concentrically inside those shown on the right. The outer grooves pump the air inwards from the outer edge to give an overpressure; the other set of grooves causes an underpressure at the centre. The grooves have a width of 0.8 mm. The outer grooves have a depth of about 2.5 μm; the inner grooves have a depth of about 13 μm.

[*] An article on this device will appear in the next volume of Philips Technical Review. *(Ed.)*

tion of the lift (curve *I*) and the internal preload (curve *II*) as a function of the spacing *h* between the two surfaces. The spacing that the discs adopt automatically is given by the point of intersection of the resultant curve (*III*) with the horizontal axis; the lift and the internal preload are then in equilibrium. The two pressures are both proportional to the speed of rotation; if this should vary, the forces continue to balance one another at the same spacing between the discs, and the spacing therefore remains constant.

The dimensions of the patterns of grooves are chosen in such a way as to make curve *III* as steep as possible at the point where it cuts the *h*-axis. The bearing therefore has the maximum stiffness, and strongly opposes any variation in the spacing of the discs.

There are no external loads of any significance. Even at quite low speeds the weight of the disc, for example, is negligible compared with the force due to the underpressure. The orientation of the axis makes no difference to the performance. Even when the rotating disc is suspended from the stationary one, the spacing between them remains close to 3 μm (see *fig. 6*).

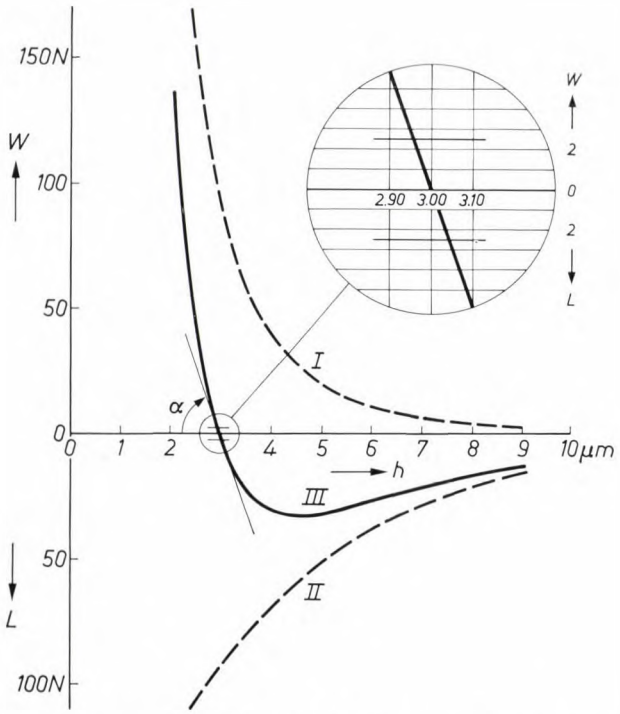


Fig. 5. Calculated characteristics of the push-pull spiral-groove bearing formed by the glass discs of fig. 3. Curve *I* is the lift *W* and curve *II* the internally generated preload *L*, both as a function of the spacing *h* between the moving surfaces. The curves were derived for a speed of 1000 rev/min. Curve *III* is the net force on the bearing, i.e. the sum of *I* and *II*. The operating point of the bearing is the point of intersection of *III* with the abscissa. The stiffness of the bearing (tan α) is about 58 N/μm. Although curves *I* and *II* and the stiffness change when the rate of rotation changes, the operating point does not move. The weight of the rotating disc (2.6 N) has hardly any effect on the position of the operating point (a displacement of only 0.05 μm, see inset).

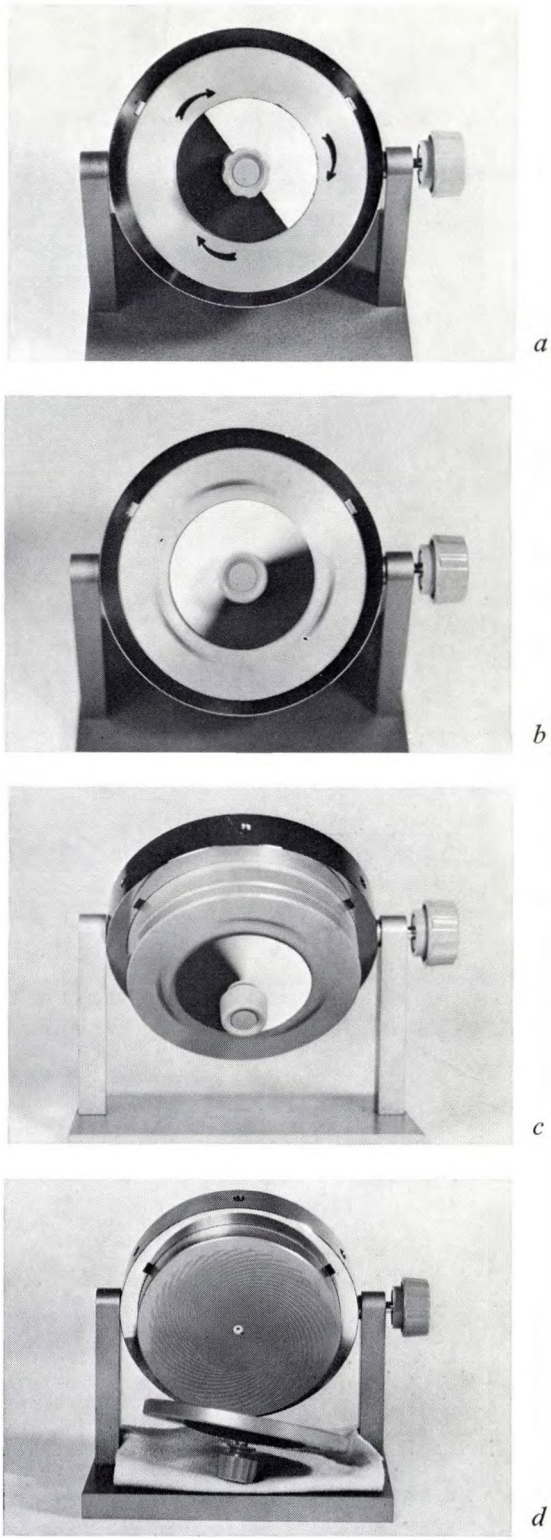


Fig. 6. Demonstration model of a push-pull spiral-groove bearing in operation. *a*) Stationary bearing; the arrows show the direction of rotation of the uppermost disc. The lower disc is clamped in a ring (dark in the photograph); the upper disc is centred by a pin that fits with a slight play in a hole in the lower disc. *b*) The upper disc now rotates (at about 500 rev/min). *c*) The bearing is set so that the rotating disc is suspended from the stationary disc. The rotating disc remains attached to the stationary disc until its speed has dropped to about 100 rev/min; the internal underpressure is then too small to support the weight of the disc (*d*).

The grooves displace only a relatively small amount of air in unit time. The volume of air contained between the discs is therefore kept as small as possible to give a rapid response to variations in speed. The bearing does not then tend to instability.

The diaphragm is preadjusted in such a way as to make the discs just touch each other when stationary. When motion commences the discs make light contact, but the desired spacing of 3 μm is soon reached. The discs do not wear, provided that the air is dust-free. The bearing will operate regardless of whether a particular groove pattern is on one disc or the other, or whether they are both on the same disc. The orientations of the patterns must be chosen to give the underpressure at the centre and the overpressure in the outer zone, since the bearing then presents the greatest stiffness to tilting [6].

H. J. W. M. Volman

[6] The problem of the tilting stiffness of a flat spiral-groove thrust bearing has been considered by J. P. Reinhoudt, On the stability of rotor-and-bearing systems and on the calculation of sliding bearings, thesis, Eindhoven University of Technology, 1972.

H. J. W. M. Volman is with Philips Research Laboratories, Eindhoven.

PINPOINT — a radio system for locating and monitoring vehicles

R. W. Gibson

The hyperbolic radio-navigation systems for ships and aircraft, developed during and after the second world war, have recently found counterparts in systems for the location and mapping of vehicles at a central point. PINPOINT is a system of this kind which is particularly simple and inexpensive because it is based on existing radio-telephony channels. Whole fleets of several hundred vehicles can be located with an accuracy of less than half a kilometre. This location system can be used as part of an extremely versatile command and control system. Thus, when the location and nature of an 'incident' are typed in, the computer will search a street directory to find the coordinates of the incident location and will request the display to indicate the vehicle best placed to attend.

Many organizations which use a large fleet of vehicles wish to control the vehicles from a central point. The advent of mobile radio-telephones has made this a possibility and, today, virtually every police car, ambulance and fire engine has a two-way radio-telephone. To an increasing extent buses, taxis and other commercial organizations are also using radio-telephones to control their fleets. If the controller is to use his vehicles most effectively he needs to know where each one is and what it is doing, i.e. he needs location and status information. At present this information is gathered by talking to the driver on the radio-telephone and is stored either in the control operator's mind or sometimes in manual plotting systems.

Recent advances in integrated circuits and small computers have now made it feasible to carry out this information-gathering and storage function automatically. A vehicle-location and monitoring system called PINPOINT has been developed at Mullard Research Laboratories in conjunction with Pye Telecommunications Ltd.

The technique used enables radio location to be fitted as an optional extra to a standard radio-telephone system. It locates the vehicles by time-difference measurements using a standard mobile-radio channel (3.5 kHz bandwidth, 12.5 kHz spacing). However, the system has been designed so that locations can also be determined with beacons or by other methods (such as the measurement of distance travelled by the vehicle). PINPOINT is also capable of fast data transmission, so that the system can be tailored exactly to meet individual requirements. Vehicles can be located in a wide area, while at the same time extra information from the vehicle can be obtained from automatic data signals. These can provide extra indications of position in important areas (for a taxi fleet this could be the station), status and other information.

The next section starts with a survey of the various methods for location and of the factors that determine the choice of method for PINPOINT. This is followed by a more detailed account of the method and a description of the equipment. The equipment, the range of applications for PINPOINT and its performance are described in the last two sections.

R. W. Gibson, B.Sc., is with Mullard Research Laboratories, Redhill (Surrey), England.

Location techniques

In vehicle location it makes a great deal of difference whether the vehicles are on fixed routes, like buses, or whether they can go anywhere, like police cars. For fixed routes roadside beacon systems can be used; the vehicles recognize signals from a particular beacon that they pass on the route and the information is transmitted back to base. The inverse system may also be used in which monitors at known points on the route recognize each vehicle as it passes and send the information back to base.

Various systems of this type have been proposed using a wide variety of recognition signals — optical, radio, mechanical, or electrical (inductive loop). However, the cost of the beacon system discourages their use except in cases where great accuracy is essential.

The alternative is to use some sort of radio-location technique. Radio location is well known for the navigation of aircraft, ships and space capsules; the same techniques can be used for vehicles but the operational requirements and economic considerations are different. One important aspect which discourages the use of many of the radar techniques is the large bandwidth necessary for transmission of the signals. It is difficult to justify any system which uses an unduly large share of the already overcrowded radio spectrum, and the idea of several users in the same area, each with hundreds of vehicles transmitting wideband signals, is decidedly unattractive.

In principle the range from a fixed site can be measured by measuring the 'there-and-back' time. This is not practical for narrowband mobile-radio channels, however, because the 'turn-around' time at the vehicle (delay between the mobile unit receiving the interrogation signal and re-transmitting the response) cannot be kept sufficiently constant to obtain the desired accuracy (about 0.5 km). (Because of other uncertainties this time should not vary by more than 1 μs.)

In theory it is possible to locate a vehicle from any radio signal it transmits; suitable receivers may be used to measure some characteristic of the signal, such as its bearing, time of arrival, or strength, and from several simultaneous measurements the position can be derived. The choice of which parameter to measure is mainly governed by the propagation characteristics of the radio signals. The signal strength can only be used for vehicle location if the statistical mean is known for signals from the vehicle in several positions and is not suitable for locating stationary vehicles. (Signal strength can be drastically affected by local obstacles, especially large metallic objects such as bridges or overhead cables.) Bearing and arrival time are both subject to errors due to multipath propagation (reflection, diffraction, etc.), but bearing errors are likely to be greater, unless the receiving site is entirely free from obstructions. The PINPOINT system therefore uses time *differences*.

Another criterion having an important bearing on the choice of system is that of price: since most vehicles of interest either have radio-telephones or will be fitted with them it is cheapest to use existing facilities for location.

The conventional navigation systems such as OMEGA, LORAN, or DECCA [1] are all in principle suitable for vehicle location, but because they need a special receiver in the vehicle they are inherently more expensive and the position data still has to be transmitted back to base. In terms of accuracy they are about the same as the techniques used in PINPOINT, their only real advantage being that they can cover a large area.

Location by time difference

The schematic diagram of *fig. 1* shows three fixed receiving sites and a vehicle equipped with a radio-telephone. The vehicle transmits a radio signal which is received at all three sites; the time of arrival of the signal is measured very accurately at each site. If the signal arrives at site *1* 10 μs earlier than at site *2* then we know that the transmitter is 3 km nearer to site *1* than it is to site *2* (because a radio signal travels 3 km in 10 μs). We can therefore plot a line on our diagram (or map) which shows all points 3 km closer to site *1* than site *2*, and know that our vehicle lies somewhere on this line. The line is shown on the diagram labelled $t_2 - t_1 = +10$ μs, and is a hyperbola. Systems of this class are therefore known as hyperbolic navigation systems.

Similarly if the time of arrival at site *2* is 20 μs later than at site *3* the vehicle is also on the hyperbola $t_3 - t_2 = -20$ μs, showing points 6 km nearer to site *3* than site *2*. Hence the location of the vehicle is at the intersection of the two hyperbolae.

Hyperbolic navigation systems were first introduced during the second world war for the guidance of ships and aircraft. For navigation purposes, the system used is the exact inverse of that used in PINPOINT. Three (or more) transmitters send out synchronized signals and the ship or aircraft measures the relative delays, from which it can plot its position on a grid of intersecting hyperbolae.

The first system of this kind, proposed in 1935 by R. J. Dippy [2], was developed in England during the early years of the war and was called the 'Gee' system. Transmitters based in England enabled bombers to be guided to target areas on the continent with an accuracy of a few miles. Later, in 1944, 'Gee' played an important part in the accurate navigation of the ships and the coordination of the landings for the invasion of Europe. The Gee system was based on pulse techniques and therefore occupied a wide frequency band. This makes it less suitable for civil applications where the allocation of spectrum space is not at a premium.

The Decca system [3] developed soon after the war, uses continuous-wave signals and therefore takes up a much smaller space in the spectrum than pulsed systems. A similar system but adapted to navigation over longer distances is 'Loran', developed in the U.S.A.

A hyperbolic location system therefore requires a pattern of receiver sites situated so that a least three of them can receive the vehicle signal. These sites relay the times of arrival of the signals (or the signals themselves) to a central point where they are processed. In PINPOINT this is done by a small computer.

For location systems to be capable of handling a whole fleet of vehicles the following additional features are also necessary:

[1] See for example G. J. Sonnenberg, Radar and electronic navigation, 4th edition, Newnes-Butterworths, London 1970, or R. A. Smith, J. IEE 93 IIIA, 331, 1946.
[2] R. J. Dippy, J. IEE 93 IIIA, 468, 1946.
[3] C. Powell, Proc. IEE 105B, suppl. No. 9, 225, 1958.

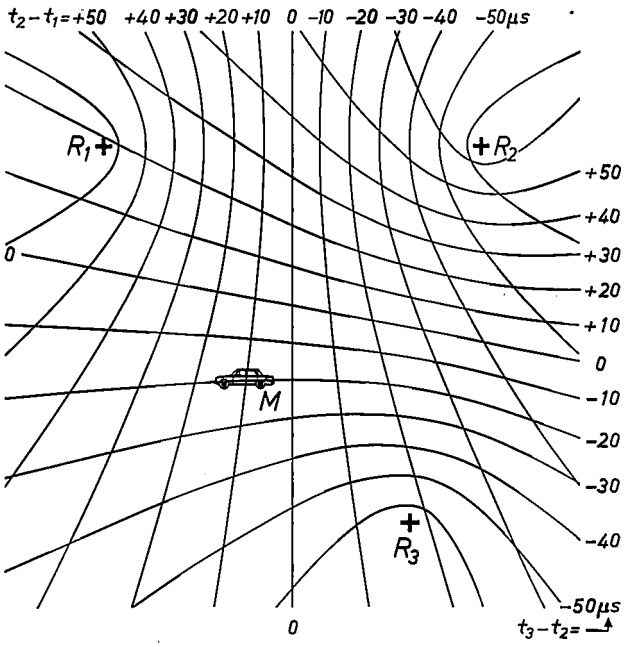


Fig. 1. Principle of the hyperbolic system of vehicle location. *M* represents a mobile vehicle with its transmitter. $R_{1,2,3}$ represent receiver sites. The lines 0-0 are the loci of points equidistant from receivers *1* and *2* and *2* and *3*. The two families of hyperbolae are loci of constant *difference* in 'transmission time' from the mobile transmitter *M* to the various receivers. The third line 0-0 and its family of hyperbolae are omitted for clarity. For the situation shown, the mobile transmitter *M* lies on the hyperbola $t_2 - t_1 = +10$ μs of the family *1-2* and on the hyperbola $t_3 - t_2 = -20$ μs of the family *2-3*. The signal from *M* is thus received at site *1* 10 μs earlier than at site *2*. In the PINPOINT system the appropriate hyperbolae are calculated from the measured time differences, and the point of intersection gives the position of the vehicle.

— Each vehicle of the fleet must be able to transmit individually without mutual interference.
— The position of each vehicle must be stored at the central point and updated at regular intervals.
— A display must be provided to show the controller where the vehicles are.

It is also desirable that other information, such as status messages, can be automatically transmitted from the vehicles and kept in a store, to be displayed automatically or on demand.

The methods used to achieve these various features in the system developed by MRL and Pye will now be explained.

Practical system

As mentioned above, the PINPOINT system is based on the use of a standard radio-telephone in each vehicle. A modulator-demodulator unit (data modem) is added to the equipment to interpret incoming data and to modulate the outgoing continuous-wave signal with the data to be transmitted (e.g. vehicle data, status, etc.).

The time measurements are made by measuring the phase of an audio tone that modulates the r.f. carrier. When, as is usually the case, data transmission is also required, it is simple to put the data in the form of binary code so that phase-shift key modulation (PSK) can be applied to the audio subcarrier. In this method of modulation the binary 0 is represented by a single period of the original subcarrier, while to represent a binary 1 the phase of this subcarrier is shifted by 180°. (An alternative form of PSK is obtained if each binary 1 causes a phase shift of 180°, while the phase remains unchanged for a binary 0.) The signal for the location of the vehicle is simply a burst of several cycles of unmodulated subcarrier.

Fig. 2 shows the waveform of a complete signal. The first 16 ms is occupied by the burst of unmodulated 2-kHz subcarrier, which serves not only as the location signal, but also to help establish the link with the data receiver (the receiver and data modem need a few ms of unmodulated carrier to prepare themselves for the reception of data). In the second 16-ms period there are 180° phase shifts in the subcarrier, so that the phase of each complete cycle is set to either 0 or 180° to denote either a 0 or 1 data bit as explained above.

The time measurement consists of a precision phase measurement made on the first 16 ms (about 32 cycles) of the incoming signal, against a quartz clock that also has a frequency of 2 kHz. This gives an ambiguity of half a cycle in the phase measurements. At a frequency of 2 kHz this corresponds to 250 μs, or a distance of 75 km. This ambiguity is so large that it is generally unimportant, because for most vehicle-location systems a range much less than this is demanded. If necessary, however, the ambiguity can be resolved by a time measurement on the data signal; such a measurement does not have to be particularly accurate.

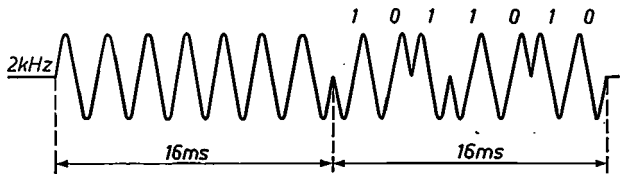


Fig. 2. A typical waveform (2 kHz) for the signal modulating the carrier in the mobile transmitter. Note that only 7 cycles of the 2-kHz subcarrier are shown in each period, for clarity. The location information is contained in the first 16-ms burst of unmodulated audio subcarrier signal, and is obtained as follows. The phases of the unmodulated subcarrier signal, as received at the three (or more) fixed sites, are measured against precision quartz clocks. The data contained in the second 16-ms period is in binary code, the carrier being modulated by 180° phase shifts. In the example shown a cycle of audio subcarrier of the same phase as the previous cycle represents a 0, while a cycle of opposite phase to the previous one represents a 1 (differential phase-shift key modulation).

Fig. 3 shows a block diagram of the complete system. The computer, which stores the information relating to position and status of the vehicles, decides which vehicle is to be interrogated and sends a data or interrogation signal to the transmitter *Tr*; the data signal includes the coded number of the required vehicle. When the vehicle receives and recognizes its own code number, it replies by transmitting a signal. This signal is received at the three (or more) receiver sites and at each of these the phase of the location tone is measured and the data contained in the modulated part decoded. The results of the phase measurements and the decoded data are then sent via modems over the links to the central control room.

The computer in the central control room keeps in its memory a complete list of all vehicles in the fleet, their latest positions, and any status information or other data received from the vehicles. The list is continuously and automatically updated as the new data flows in from each new set of measurements. The computer not only calculates the new positions of the vehicle in cartesian coordinates, it also applies a plausibility check to the new position. If this is satisfactory the stored position is updated; if not, the new position is rejected.

The system can scan the fleet at a typical rate of about 25 to 50 vehicles per second, depending on the amount of data to be transmitted.

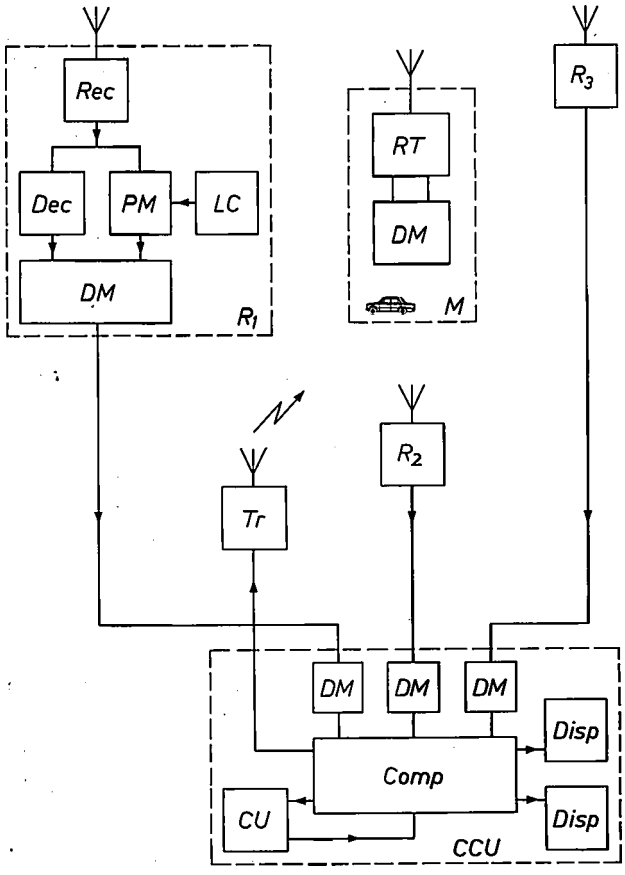


Fig. 3. Block diagram of the PINPOINT system. *M* is one of the vehicles to be located. All the vehicles have a standard radiotelephone *RT* and a modulator-demodulator (data modem) *DM*. There are at least three receiver sites $R_{1,2,3}$; one of these is shown in detail. The phase of the first part of the received signal (fig. 2) is measured (*PM*) against precision quartz clocks *LC* at each site. The data in the received signal is decoded (*Dec*) and the combined phase and data information is then sent in binary form to the command-control centre *CCU*, usually by telephone line. It is decoded there, and fed into the location and status computer (*Comp*). This calculates the position of the vehicle in cartesian coordinates. The operator can obtain this position on a display panel (*Disp*) showing a map of the area covered by the vehicles. Selected information can be shown on a second display. *CU* control unit. *Tr* central-command transmitter.

Mobile equipment

In order to monitor a whole fleet of vehicles on one radio channel it is necessary that each mobile unit only transmits for a short period; the 32-ms burst of fig. 2 is typical. It is therefore necessary to be able to switch the mobile transmitter on and off quickly; this is achieved very simply by using semiconductors for the transmit/receive switch instead of a relay.

It is important to note that the system calculates position only from the time *differences* measured at the fixed receiver sites, so that the mobile unit does not have to transmit at any precise instant.

Nevertheless each mobile unit does need to know to within a few milliseconds when it should transmit its signal. This is therefore done at the request of the central control room. The modem in each vehicle is fitted with a detachable code plug which identifies that particular unit and circuits are provided with which it can recognize its own code (*fig. 4*). The central system can then interrogate any unit by transmitting its digital code together with a message code (for example, 'canned' questions) to the vehicle and command it to reply with a location and data burst.

In cases where data transmission to the car is not required, a simpler 'time-slot' system can be employed. This is a system in which the whole fleet is scanned in a fixed sequence. Every few seconds the system broadcasts a synchronizing signal to every mobile unit. Instead of a data modem, each mobile unit is fitted with a timing unit which locks on to the synchronizing signal and transmits a reply in a pre-assigned time slot.

Fixed-site receivers

The phase measurement must be made to an accuracy of better than one microsecond (1 μs corresponds to a change in position of 200 m). Because of this, special radio receivers are necessary at the fixed sites. Although it is possible to use either amplitude-modulated (AM) or frequency-modulated (FM) radio, the FM system is preferred since it is easier to make suitable receivers.

In a conventional FM receiver the limiting action on strong signals causes the phase of the audio output to vary with signal strength. Regardless of signal strength, which can vary from a fraction of a microvolt to several millivolts, the phase delay must remain constant to within 1 μs, and to achieve this a special FM receiver has been designed. The i.f. amplifier in this receiver gives a constant output phase for signals varying from just detectable to very strong.

The transient response of the receiver must also be fast enough to permit a precise measurement to be made a few milliseconds after the start of the signal. The receiver is a narrowband device because in mobile-radio-telephone practice each radio channel is strictly limited to a working band only a few kilohertz wide. Since the response time is inversely proportional to bandwidth, it follows that the receiver will have an overall delay of several hundred microseconds, nearly all the delay occurring in the channel filter. Even with careful design it is not practical to keep this delay absolutely constant to better than one microsecond. However, this can be allowed for by including an automatic calibration routine in the system. Every few seconds a calibration transmission is measured and the result used to monitor the delays in the receivers and measuring channels. The calibration is effected with a radio unit that transmits from a fixed site on a high building, and functions as a 'vehicle' of accurately known position.

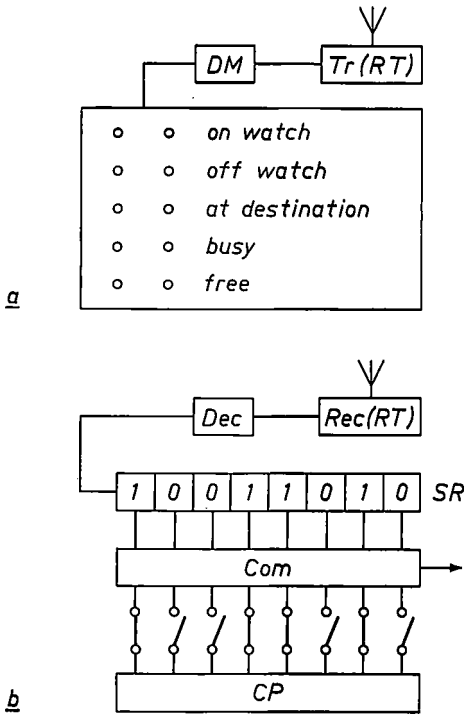


Fig. 4. *a*) Vehicle equipment for sending 'canned' information to the central command. The message is selected by pressing a button; the message is then encoded by the data modem *DM* and transmitted to the command centre by the vehicle's radio-telephone transmitter *Tr(RT)*. *b*) Data-reception equipment in vehicle. *Rec(RT)* vehicle radio-telephone receiver. *Dec* decoder. The signals transmitted by the central command are accompanied by an identification signal, which is simply the number of the vehicle called. This is sent as a binary number. After decoding this number appears in the shift register *SR*. If the comparator *Com* finds that the number is the same as the number on the code plug *CP*, the message is accepted and displayed in that vehicle. In other vehicles the message is not displayed.

Phase measurement

After the tone burst has been received and demodulated there is a choice of procedure. Either the tone signal from each receiver site is sent to the control centre where the phase differences are measured, or the phases are measured at each receiver site against a local clock and the resulting measurement sent. The second method has the advantage that the link need only transmit a digital measurement (via a data modem) and does not demand phase stability of the link: against this advantage must be offset the expense of providing a precision clock at each receiver site.

In either case the measurement is made on the zero crossings of the tone signal (*fig. 5*). The 2-kHz sine wave is first shaped into a square wave in a circuit carefully designed to avoid phase variation with signal strength. The positive edge of the square wave is then used to open a gate which allows clock pulses into a counter. The gate is shut by the corresponding edge of a 2-kHz reference signal. The length of time the gate is open, and hence the number of clock pulses counted, therefore depends on the phase difference between the incoming signal and the reference signal. By repeating this over several cycles a more accurate average phase is measured.

The process is complete in a few milliseconds and the content of the counter is the required measurement in the form of a binary number: this is then sent to the computer at the central control room. The computer thus receives three such time measurements, one from each of the receiver sites.

A data decoder is used to extract the phase shifts in the second half of the signal and the resulting binary data is also sent to the computer.

Display

Fig. 6 shows an early version of the control-room equipment as laid out for an exhibition. The obvious way to display the position of the vehicles is by superimposing symbols on a map. The display shown in fig. 6 uses a half-silvered mirror to combine symbol images on a television screen with a printed or back-

projected map. The result is shown in *fig. 7*. In the case shown, the position of the suffix symbol in each group indicates the position of the vehicle, the suffix symbol itself giving the status of the vehicle; the upper three digits give the vehicle number.

An alternative technique consists in projecting a map through a window in the back of a special TV picture tube. For large wall displays projection TV can be employed.

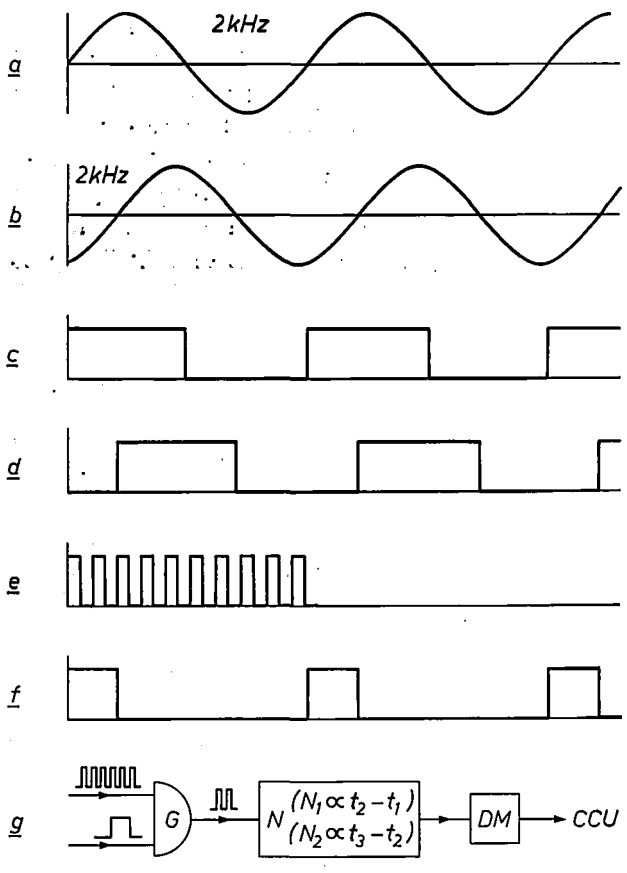


Fig. 5. Phase measurement of received signals against precision quartz clocks. *a*) Received signal (unmodulated 2-kHz subcarrier). *b*) Waveform of local precision clock (2 kHz). *c*) Square-wave signal derived from (*a*). *d*) Square-wave signal from precision clock (*b*). *e*) Clock pulses (e.g. 10 MHz). *f*) Gate waveform for admitting clock pulses to counter. *g*) The clock pulses are admitted to the counter only while the gate *G* is open. The number of clock-pulse counters is therefore proportional to the phase difference between the received signal and the precision clock. The count is sent as a binary number to the central computer. Comparison with the counts from the other receivers yields the time differences $(t_2 - t_1)$ and $(t_3 - t_2)$, see fig. 1.

A display which simply showed the position of every vehicle in the fleet simultaneously would be virtually useless unless the fleet is very small in number. The amount of information that can be usefully put on display simultaneously is very limited — too much data either distracts from or, worse still, obscures the vital information. To overcome this problem PINPOINT provides an interactive display so that the information presented is limited to that relevant to the job in hand.

A push-button keyboard gives the operator complete control over the display — he can add or delete any vehicle or group of vehicles instantly. Vehicles can be called by:

1. Name	e.g. Car No. 223
2. Type	e.g. All cars of Type 6
3. Status	e.g. All cars of Status 10/4
4. Absolute position	e.g. All cars within 1 mile of grid reference 213 457
5. Relative position	e.g. All cars within 5 miles of vehicle No. 321

If the system is operated by several operators, each one can have his own display, each display being completely independent, showing different data, different maps, different areas, etc.

The 'main list' within the computer memory, which holds all the position and data information on every vehicle, is continuously updated, regardless of what is being displayed. If the data from any vehicle is currently on the display its position and status will be automatically kept up to date as signals are received from that car: typically for a fleet of 300, each vehicle will be updated every 10 seconds. When the operator uses the keyboard to change the display or request additional information, the response is immediate because it only takes the computer a fraction of a second to search the list for the required data and transfer it to the display.

Other types of display can also be used if required. For instance a second visual display unit may be interfaced with the computer to give listings of particular vehicles and information about them.

The computer can also be used to do some special preliminary sorting or collating of data on the vehicles so as to give automatic display of some anomaly or unusual situation. Certain special conditions can be detected and brought to the controller's attention. Thus, an alarm signal could indicate some emergency, or an automatic warning could be given if there were no vehicles of status 'available' within a certain area.

Where the location facility is only part of a full command-and-control system the display is also available in an another automatic mode. This is a facility particularly useful to a city police department or to the operator of a taxi fleet. When the address and nature of an incident is typed in, the main computer will search a street directory to find the coordinates of the incident and then command the location system to display the best vehicle to deal with that incident, taking into account distance, status, type of vehicle, etc.

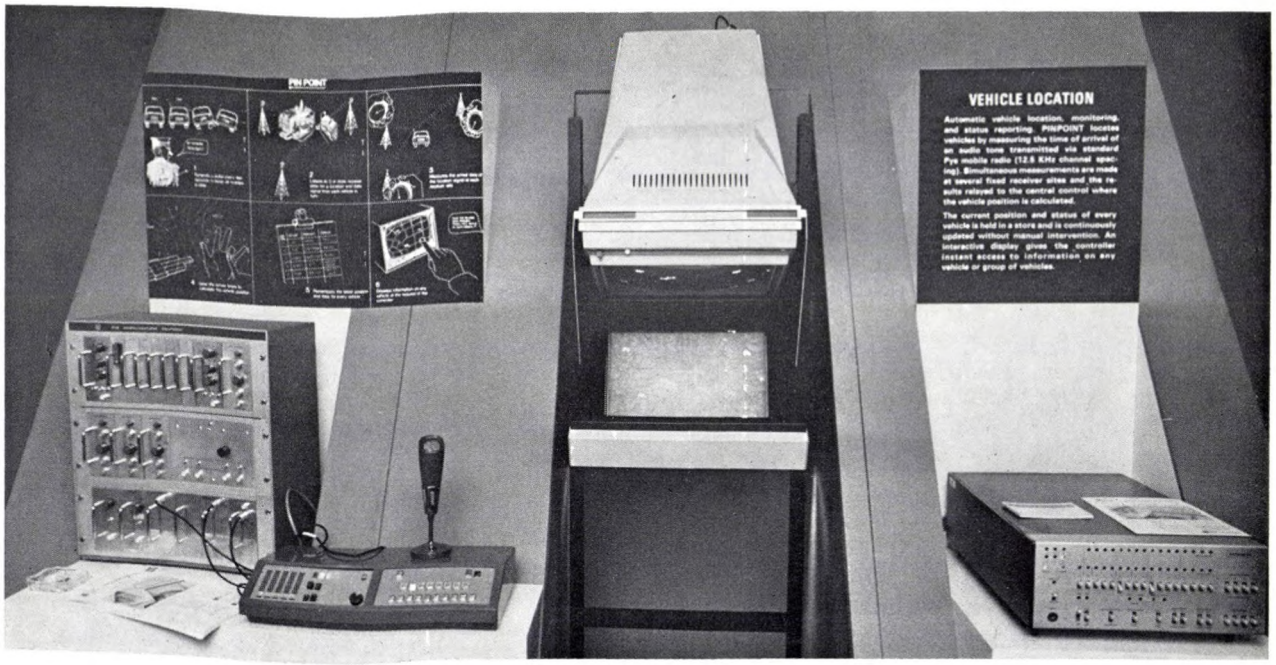


Fig. 6. Central command-and-control equipment, as set up for an exhibition. In the centre is the vehicle-location display (see also fig. 7). On the left is the radio-telephone with display control panel. On the right is the computer.

Fig. 7. Photograph of location display on a map of central London. The three-digit numbers are the vehicle identification numbers and the suffix is a status number (e.g. busy, free, returning to base, etc.). The actual location of each vehicle is marked by the suffix digit.

Accuracy and coverage

All radio-location systems are subject to errors because the propagation of radio waves is affected by buildings, trees, hills, power lines etc. The nature of the errors varies depending on the frequency and bandwidth of the channel employed but it is very doubtful if any practical radio-location technique can give true street-corner accuracy in a city. The PINPOINT system gives results comparable with other known methods but at less cost.

A single spot measurement can show an error of more than a kilometre due to combined effects of interference and multipath signals, but a considerable amount of averaging and signal processing is carried out in the computer so that these errors are much reduced. A simple three-station system set up in central London using a standard VHF channel (160 MHz) has shown mean errors on the display of about 0.5 km, with occasional errors of up to 1 km.

Although averaging several successive readings from a vehicle improves accuracy, there is a limit to this improvement. For a typical 300-vehicle system, where each vehicle sends a signal every 10 seconds, averaging four consecutive signals means that the displayed position of the vehicle lags the true position by 40 seconds; at a speed of 60 miles (96 km) per hour this corresponds to a kilometre. How serious this is depends on the vehicle speed but a compromise has to be struck so that the errors due to this lag are not greater than the gains due to the averaging. In practice it is reasonable to average between four and eight successive readings, corresponding to a lag of a minute or so. Where this is unacceptable, smaller fleet sizes on one channel or more frequent signals from particular vehicles can be used.

In order to cover a certain area it is necessary that a signal from anywhere within the area can be heard by at least three receiver sites. In small or medium-sized towns where the terrain is suitable it is sometimes sufficient to set up only three stations. For very large cities, country areas, and places where there are obstructing hills it is necessary to use extra receiver sites. (Extra sites also improve the accuracy slightly but the improvement is not commensurate with the extra costs.) Typical systems will use between 4 and 10 receiver sites, or even more for very large cities such as London.

Summary. PINPOINT is a system for locating vehicles which is based on the measurement of the differences in time of arrival of the signals received at three (or more) fixed receiver sites when each vehicle transmits a short burst of a relatively narrowband signal. PINPOINT provides the controller of the fleet with detailed knowledge of the whereabouts and status of every vehicle in the fleet. This display provides clear and easily read symbols superimposed on a map or diagram. A keyboard gives the operator complete control over the information displayed, and operation of the system is extremely simple. The system uses a standard mobile-radio channel with a bandwidth of about 3.5 kHz and a channel spacing of 12.5 kHz. It collects the data much faster than is possible with speech, thus saving valuable transmitting time.

The extra equipment required in the vehicle is simple. The complete system is less expensive than other systems, but is just as accurate (about half a kilometre). The location function is compatible with data transmission, so that hybrid systems using beacons, distance run, etc., can be assembled to suit any particular application.

A fast scanning microscope used in the development of avalanche photodiodes

P. M. Boers and L. J. M. Bollen

The avalanche photodiode

To achieve a high sensitivity in an avalanche photodiode it is necessary that the sensitivity should be very uniform over the whole surface of the diode. An avalanche photodiode is a light detector with internal amplification; it can be regarded as the solid-state analogue of a photomultiplier tube.

An important application of the avalanche photodiode is as a detector in a wideband communication system with an optical carrier. In such an application the diode must have a high sensitivity at the carrier wavelength. (In a communication system with a GaAs injection laser as the source the carrier wavelength is 0.8-0.9 μm.) Other requirements are a high signal-to-noise ratio and a response that is rapid enough to detect the highest modulation frequencies, which may extend to several GHz for optical communication systems.

The avalanche photodiode is essentially an ordinary *P-N* diode in which one of the electrodes is transparent to the radiation to be detected. The diode operates as an optical detector when a voltage is applied in the reverse direction. The reverse voltage causes a carrier-free zone, the 'depletion layer' (*fig. 1*) to appear on both sides of the junction. When radiation falls on the diode, electron-hole pairs are formed. The electrons and the holes formed in the depletion layer will be accelerated in opposite directions by the electric field there. Depending on the place where the electron-hole pair is formed, either the electron or the hole will move to the location of maximum field-strength. If this maximum field-strength is high enough, the charge carriers can obtain sufficient energy to enable them to initiate further ionization and thus produce new electron-hole pairs. These are then accelerated by the local field so that, for each photon absorbed, a whole avalanche of charge carriers is created, which appears as a photocurrent in the diode. The multiplication factor, i.e. the number of charge carriers per absorbed photon, is dependent on the field-strength in the avalanche region. For a large multiplication factor, the local field-strength must be high. It is therefore desirable that the total voltage across the depletion layer should be as high as possible, i.e. as close a possible to the 'dark' breakdown voltage of the diode. Owing to local variations in the semiconductor material of the diode, there will be local variations of this breakdown voltage. The lowest breakdown voltage anywhere in the layer will therefore determine the maximum reverse voltage that can be applied across the diode and hence the average multiplication factor that can be achieved by the complete diode. Good homogeneity of the material, and hence as high a field-strength as possible over the avalanche layer, is therefore very important for the optimum performance of an avalanche diode as a radiation detector. A scanning microscope has been found to be ideally suitable for assessing the homogeneity of avalanche diodes either after completion or in the various stages of manufacture.

P. M. Boers and Ir L. J. M. Bollen are with Philips Research Laboratories, Eindhoven.

The geometry of an avalanche photodiode

If an avalanche diode is to have a rapid response to the incident radiation, as many of the charge carriers as possible should be created in the depletion layer.

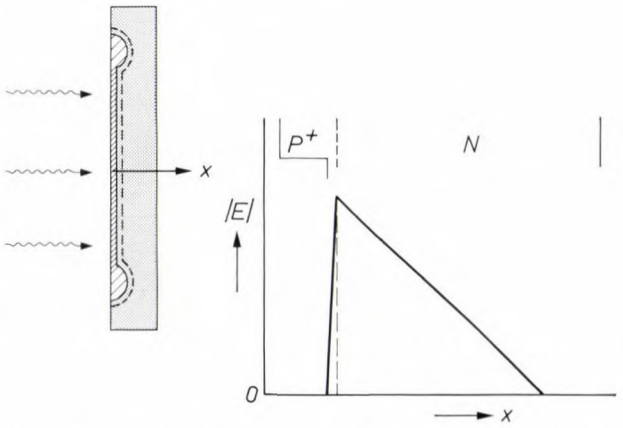


Fig. 1. Structure of a *P-N* photodiode (*left*) and the field-strength *E* in this diode (*right*) as as function of position *x*. The diode consists of a thin P^+ layer on *N* material. The P^+ layer is so thin that the light to be detected is transmitted through it to the *P-N* junction. The P^+ layer is surrounded by a *P* layer whose shape is such that excessively high fields along the edge of the P^+ layer are avoided. The field profile arising in this diode when a voltage is applied across it in the reverse direction is entirely determined by the doping profile; the gradient of the field at any point is proportional to the dope concentration. If the voltage across the diode is high enough, the field-strength in the neighbourhood of the *P-N* junction becomes so large that avalanches of charge carriers may occur.

The reason for this is that charge carriers created outside the depletion layer must first reach the depletion layer by diffusion before they can be accelerated and make their contribution to the photocurrent; this diffusion process is relatively slow and consequently introduces an appreciable delay in the response of the diode. This delay and the absorption of the light in the semiconductor together determine an optimum thickness for the depletion layer. If the layer is thicker than this value inconveniently high voltages are required, and it is difficult to avoid breakdown over the outside surfaces of the diode. If the depletion layer is thinner than this value many charge carriers originate outside the layer and this introduces delay in the response. However, a correction for this can be introduced by making the lifetime of free charge carriers outside the depletion layer very short. Electrons and holes generated outside the depletion layer will then recombine rapidly so that only a negligible number of them reach the depletion layer by diffusion.

The diode material is given a doping profile that will provide the field-strength variation shown in *fig. 2*. The field-strength reaches a value of about 40 kV/cm in the greater part of the depletion layer. This is sufficient to make the charge carriers move at the maximum mean velocity, the 'saturated' drift velocity, in the direction of the field. The field-strength in the avalanche layer reaches a value of about 300 kV/cm, accelerating some of the charge carriers to such a high velocity that the avalanche effect can take place. The drift velocity in this layer no longer increases, however (*fig. 3*).

A diode with such a doping profile performs well even at voltages below 200 V.

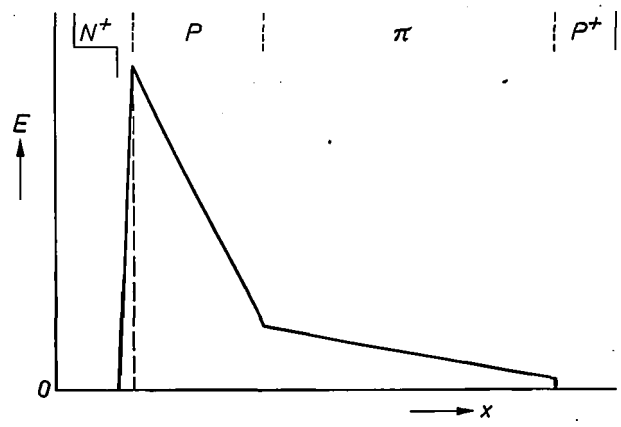


Fig. 2. The field-strength E in an $N^+P\,\pi\,P^+$ diode as a function of position x. The π layer is a very weakly doped P layer. The $N^+P\,\pi\,P^+$ structure gives a field across the P-N junction region that is sufficiently high for the avalanche effect to occur there although the total voltage across the diode (i.e. the area under the curve) is not excessive.

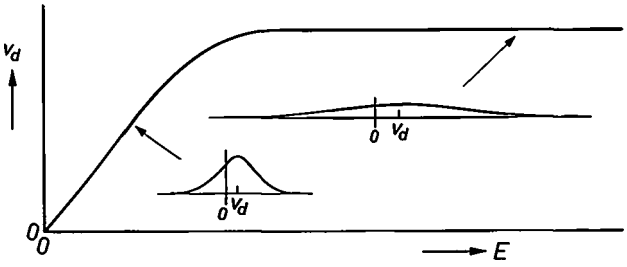


Fig. 3. Drift velocity v_d as a function of the field-strength E in the depletion layer. At low fields the relation between v_d and E is linear. As the field-strength is increased there comes a point where the drift velocity no longer increases with field-strength. The two inset diagrams show the distribution of actual velocities about the drift velocity (on a different scale) for two values of the drift velocity.

The scanning microscope

The use of flying-spot scanners for the investigation of photocells and photodiodes is certainly not new. For our purposes such a scanner has to have both a high resolution and a fast scan. The high resolution is necessary for observing small details on the diode surface and the scan must be fast so that we can follow variations in the voltage across the diode. None of the instruments of which descriptions have appeared [1] possess this combination of features. We have therefore developed an instrument which can scan an object with a very fine light spot at a rate of 50 scans per second. *Fig. 4* shows a block diagram of this instrument. *Fig. 5* is a photograph of the complete equipment. The diameter of the scanned area can be varied between 50 μm and some millimetres. For diameters greater than 200 μm, the resolution is about 1/200 of the diameter of the scanned field; for smaller diameters the resolu-

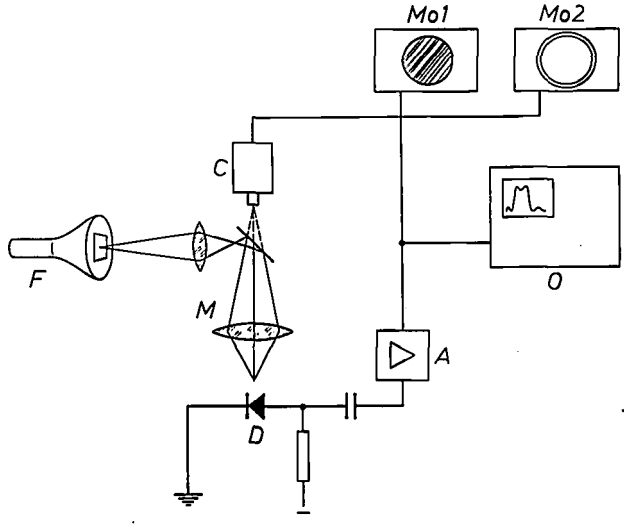


Fig. 4. Block diagram of the flying-spot microscope. *F* flying-spot scanning tube. *M* optical microscope. *D* diode under investigation. *A* amplifier for photocurrent. *C* television camera for observation of optical image. *Mo1* monitor for display of the distribution of the multiplication factor over the diode surface. *Mo2* monitor for display of the optical image. *O* oscilloscope for the display of the diode signal for one line of *Mo1*.

tion is limited to about 1 μm by the wavelength of the light used.

The flying spot is obtained from the screen of a cathode-ray tube specially developed for flying-spot scanners [2] designed for television purposes (films, transparencies, etc.). The electron beam in the cathode-ray tube writes a raster on the screen at a repetition rate of 50 Hz. The phosphors on the screen in this type of tube have a short afterglow, matched to the scanning frequency, so that the afterglow does not degrade the resolution of the instrument. We have used a specially manufactured flying-spot tube with a phosphor that emits light of wavelength between 500 and 600 nm; standard tubes of this type contain a second phosphor emitting light of a shorter wavelength. This short-wavelength light would be absorbed just beneath the surface of the diode and would be of no significance in our experiments because electron-hole pairs immediately vanish there by recombination.

A reduced image of the flying spot on the screen of the tube is projected by a microscope on to the sensitive surface of the diode to be investigated. The photocurrent generated by the moving light spot modulates the brightness on a monitor scanned in synchronism with the flying-spot tube, giving a representation of the variation of the sensitivity over the diode surface.

The amplifier, which brings the photocurrent up to a level suitable for the monitor input, has a noise level equivalent to a photocurrent of several nA. The bandwidth of the amplifier is several MHz to permit a rate of 50 pictures per second. This picture frequency has the great advantage that normal video equipment can be used for most parts of the electronics.

It is also possible to represent the photocurrent on an oscilloscope. This allows quantitative measurements of the diode current to be made easily; the local multiplication factor and hence the uniformity of the various diodes can therefore be compared. If the signals of adjacent scanning lines are given a small relative displacement a simulated three-dimensional picture of the photocurrent distribution is obtained (*fig. 6*).

A second monitor can display an optical image of the diode during the scan. This image is picked up from the microscope by a television camera, and is of use in deciding whether a particular irregularity observed with the flying-spot microscope has its origin on the surface or in the interior of the diode.

Fig. 7 shows oscillograms of the distribution of the multiplication factor over the surface of a given diode for three different voltages.

With this flying-spot microscope it was possible to make a rapid assessment of the suitability of the various batches of silicon for the manufacture of avalanche diodes (*fig. 8*). Different configurations and manufacturing methods could also be compared at an early stage in their development. In addition, specifications for the diodes could be rapidly translated into specifications for the semiconductor material and the doping process. It was found, for example, that the most homogeneous diodes were obtained when the *P-N* junctions were made by ion implantation in epitaxial silicon. Diodes made from ion-implanted material have given a multiplication of 1000×; this implies that deviations from uniform doping were of the order of

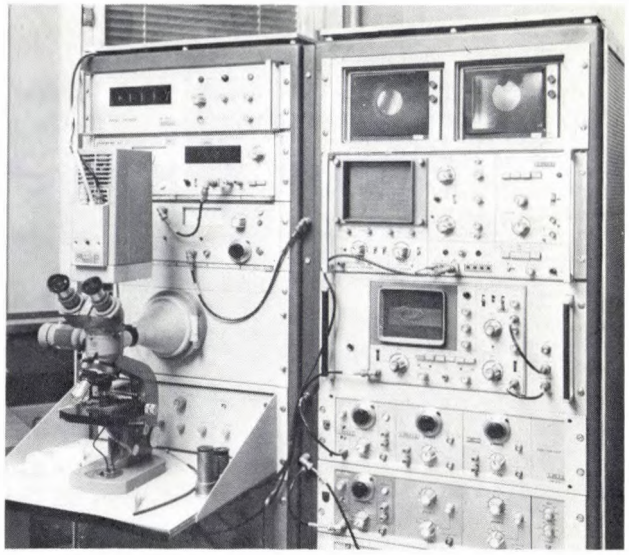


Fig. 5. The complete flying-spot microscope. The optical microscope is on the left. Behind it is the flying-spot tube and above it the television camera. The monitors for the display of the photocurrent distribution and the optical image are mounted in the right-hand rack in the uppermost panel. Underneath is the oscillograph on which the diode signal can be displayed in such a way as to give a simulated three-dimensional display of the distribution of the photocurrent over the diode surface.

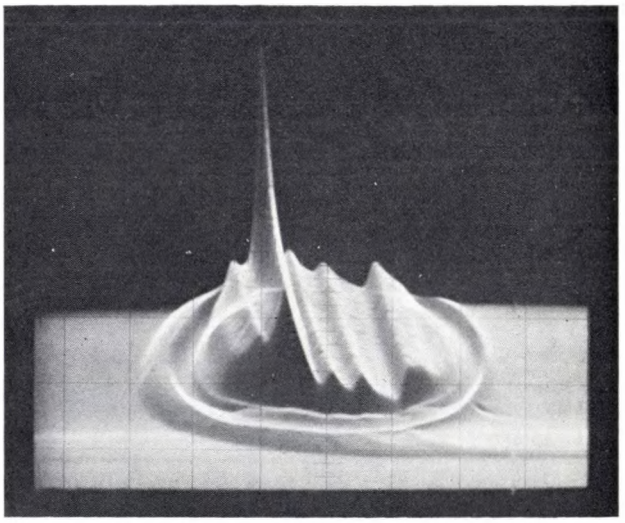


Fig. 6. Simulated three-dimensional display of the distribution of the photocurrent in a diode (the same diode was used for fig. 7).

[1] R. J. L. Lerou, Ingenieur **84**, ET 131, 1972 (in English).
[2] A. Bril, J. de Gier and H. A. Klasens, Philips tech. Rev. **15**, 233, 1953/54.

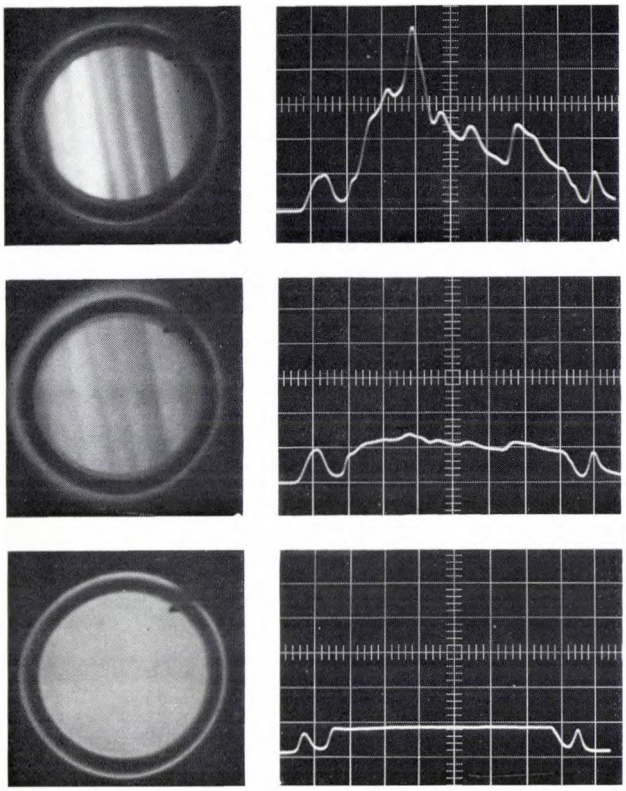


Fig. 7. The distribution of the photocurrent over the surface of a diode for three different voltages across the diode. *Left:* the distribution of the photocurrent over the whole diode surface; diode diameter 500 μm. The irregularities in the current distribution are caused by non-uniform doping. The ring-shaped structure round the circumference originates in the doping applied there to prevent breakdown at the outside edge of the diode. *Right:* the distribution of the current along a single line through the centre of the diode. The voltage across the diode when these displays were made was, from top to bottom: 118 V, 93 V and 1.5 V. The vertical deflection sensitivity of the oscilloscope was the same in all three cases.

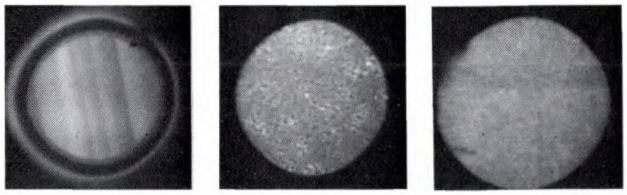


Fig. 8. Photocurrent distribution for diodes made from silicon of different quality. The right-hand picture shows the current distribution of a diode whose *P-N* junction was made by ion implantation in epitaxial silicon.

0.01 %. Such a high amplification is not however particularly suitable for an optical communication system, because of noise. A suitable diode for this purpose is a mesa diode with a diameter of 500 μm and a multiplication factor of about 100×. This multiplication requires a voltage of 175 V across the diode. The response time of the diode is 0.3 ns while its quantum efficiency is 35 %.

The flying-spot microscope described here is also useful for investigations on certain other semiconductor devices, provided that the *P-N* junction is no more than a few microns under the surface and that the surface is not coated with an opaque film.

Summary. In the development of avalanche photodiodes it is of great importance to know the distribution of the photocurrent density over the surface of the diode. To measure this distribution we have built a flying-spot microscope that scans the diode surface with a very small spot of light at a rate of 50 pictures per second. The diameter of the image field of this instrument can be varied between 50 μm and a few millimetres. For a diameter > 200 μm the resolution is 1/200 of this diameter; for smaller fields the resolution is 1 μm. The instrument has greatly assisted in the rapid design of a high-quality avalanche photodiode. Epitaxial silicon with a *P-N* junction made by ion implantation was found to be the most suitable material for such diodes.

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, 3 avenue Descartes, 94450 Limeil-Brévannes, 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*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.) *N*

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. M. van Alphen: Combined electrostatic focusing and deflection.
Adv. in Electronics & Electron Phys. **33A**, 511-525, 1972. *E*

P. Baudet & M. Binet: Transistor à effet de champ à barrière métallique sur arséniure de gallium.
Onde électr. **54**, 31-35, 1974 (No. 1). *L*

M. Berth: Technologie MOS à grille réfractaire.
Onde électr. **54**, 23-26, 1974 (No. 1). *L*

J. van den Boomgaard & F. M. Carpay: On the existence of a metastable eutectic between Co and Co_2Si.
Scripta metall. 8, 23-25, 1974 (No. 1). *E*

L. Boonstra, C. W. Lambrechtse & R. H. W. Salters: A 4096-b one-transistor per bit random-access memory with internal timing and low dissipation.
IEEE J. **SC-8**, 305-310, 1973 (No. 5). *E*

M. Bouckaert, A. Pirotte, M. Snelling, P. Wodon & D. Wybaux: Utilisation des attributs dans un système d'écriture de software.
Structure et programmation des calculateurs, Séminaire IRIA, Rocquencourt 1973, pp. 291-314. *B*

J. P. Boutot: Le HR 400: un photomultiplicateur à galette de microcanaux à gain élevé et faible bruit.
4ème Journées d'Optique Spatiale, Marseille 1973, pp. 291-300. *L*

M. J. G. Braken: Aangepaste werkstukconstructies, lastoortsvormen en gereedschap voor TIG-lassen.
Lastechniek **40**, 1-6, 1974 (No. 1). *E*

G. Brouwer & J. A. J. Jansen: Deconvolution method for identification of peaks in digitized spectra.
Anal. Chem. **45**, 2239-2247, 1973 (No. 13). *E*

M. C. W. van Buul & L. J. van de Polder: Standards conversion of a videophone signal with 313 lines into a TV signal with 625 lines.
Philips Res. Repts. **29**, 413-428, 1974 (No. 5). *E*

F. M. A. Carpay & W. A. Cense: In situ growth of aligned composites from the vitreous state. The Li_2O-B_2O_3 system.
Scripta metall. 8, 11-14, 1974 (No. 1). *E*

V. Chalmeton: Microchannel image intensifier for neutron radiography.
7th Int. Conf. on Nondestructive testing, Warszawa 1973, Vol. III, paper B-42, 6 pp. *L*

J. B. Coughlin, R. J. H. Gelsing, P. J. W. Jochems & H. J. M. van der Laak: A monolithic silicon wide-band amplifier from dc to 1 GHz.
IEEE J. **SC-8**, 414-419, 1973 (No. 6). *E*

M. Davio & G. Bioul: Taylor-expansion computation from cube arrays.
Philips Res. Repts. **29**, 401-412, 1974 (No. 5). *B*

H. Duifhuis (Institute for Perception Research, Eindhoven): Consequences of peripheral frequency selectivity for nonsimultaneous masking.
J. Acoust. Soc. Amer. **54**, 1471-1488, 1973 (No. 6).

G. Eschard, M. Fouassier & R. Polaert: Fonctions de transfert de modulation dans un intensificateur d'image à microcanaux à double focalisation de proximité. Application à la réalisation d'un tube à haute résolution pour expériences spatiales.
4ème, Journées d'Optique Spatiale, Marseille 1973, pp. 301-309. *L*

L. J. M. Esser, M. G. Collet & J. G. van Santen: The peristaltic charge coupled device.
Tech. Digest 1973 Int. Electron Devices Meeting, Washington, pp. 17-20. *E*

G. G. P. van Gorkom, J. H. Haanstra & H. v. d. Boom: Infrared and Raman spectra of the spinel $ZnGa_2O_4$.
J. Raman Spectr. **1**, 513-519, 1973 (No. 5). *E*

J. Hasker & J. J. M. J. de Klerk: Improved electron gun for the beam-indexing color-television display.
IEEE Trans. **ED-20**, 1049-1052, 1973 (No. 11). *E*

M. H. H. Höfelt: On the radial mode in bubble lattices.
IEEE Trans. **MAG-9**, 621-624, 1973 (No. 4). *E*

B. A. Joyce: The growth and structure of semiconducting thin films.
Repts. Prog. Phys. **37**, 363-420, 1974 (No. 3). *M*

W. L. Konijnendijk, M. van Duuren & H. Groenendijk: Preparation of homogeneous borosilicate glasses by wet-chemical techniques.
Verres Réfract. **27**, 11-14, 1973 (No. 1). *E*

K. Kosai & M. Gershenzon (University of Southern California, Los Angeles): New photoluminescence line-series spectra attributed to decay of multiexciton complexes bound to Li, B, and P centers in Si.
Phys. Rev. B **9**, 723-736, 1974 (No. 2). *N*

D. J. Kroon: Analytical aspects of environmental monitoring.
Meded. Fak. Landbouwwet. Gent **38**, 1826-1836, 1973 (No. 4). *E*

J. Krüger, G. Schulten & W. Jasmer: Theoretische und experimentelle Untersuchungen einiger Scanlaser-Resonatoren.
Optik **39**, 327-350, 1974 (No. 4). *H*

K. E. Kuijk: A fast integrated comparator.
IEEE J. **SC-8**, 458-462, 1973 (No. 6). *E*

P. N. Kuin & J. P. Reynders: The routine determination of boron, carbon, nitrogen and oxygen by charged-particle activation analysis.
J. radioanal. Chem. **16**, 403-411, 1973 (No. 2). *E*

D. E. Lacklison, G. B. Scott & J. L. Page: Absorption spectra of Bi^{3+} and Fe^{3+} in $Y_3Ga_5O_{12}$.
Solid State Comm. **14**, 861-863, 1974 (No. 9). *M*

F. H. de Leeuw: Influence of an in-plane magnetic field on the domain-wall velocity in Ga:YIG films.
IEEE Trans. **MAG-9**, 614-616, 1973 (No. 4). *E*

J. Lohstroh & M. Otala (University of Oulu, Finland): An audio power amplifier for ultimate quality requirements.
IEEE Trans. **AU-21**, 545-551, 1973 (No. 6). *E*

R. Metselaar & M. A. H. Huyberts: The influence of dopants on photomagnetic effects in yttrium-iron garnet.
Philips Res. Repts. **29**, 453-475, 1974 (No. 5). *E*

F. Meyer: Ellipsometry as a tool in surface studies.
Ned. T. Vacuümtechniek **11**, 77-83, 1973 (No. 6). *E*

C. H. de Minjer & P. F. J. v. d. Boom: The nucleation with $SnCl_2$-$PdCl_2$ solutions of glass before electroless plating.
J. Electrochem. Soc. **120**, 1644-1650, 1973 (No. 12). *E*

G. J. Naaijer: Bi-directional interval timer includes snap-action trip points.
Electronic Design **21**, No. 25, 142, 1973. *L*

U. Pick: A 'safe' high vacuum plant.
J. Physics E **7**, 249-250, 1974 (No. 4). *M*

G. A. Postema & S. van Heusden: Infrared emission of air pollutants induced by a CO_2 laser.
Int. J. environ. anal. Chem. **3**, 95-100, 1973 (No. 1). *E*

P. Schagen: Elektronische hulpmiddelen voor het nachtkijken.
T. Ned. Elektronica- en Radiogen. **39**, 41-46, 1974 (No. 2). *M*

E. Schröder: Digitale Laserstrahl-Richtungsmodulation und ihre Anwendungen.
Feinwerktechnik + Micronic **78**, 64-68, 1974 (No. 2). *H*

H. Schweppe & P. Quadflieg: Electromechanical properties of bismuth silicon oxide ($Bi_{12}SiO_{20}$).
IEEE Trans. **SU-21**, 56-57, 1974 (No. 1). *A*

J. M. Shannon: Reducing the effective height of a Schottky barrier using low-energy ion implantation.
Appl. Phys. Letters **24**, 369-371, 1974 (No. 8). *M*

L. A. Æ. Sluyterman & J. Wijdenes: An unusual type of enzyme inhibition.
Biochim. biophys. Acta **321**, 697-699, 1973 (No. 2). *E*

A. Thayse: Applications of discrete functions, Part III. The use of functions of functions in switching circuits.
Philips Res. Repts. **29**, 429-452, 1974 (No. 5). *B*

J. B. Theeten, J. L. Domange (ENSCP, Paris) **& J. P. Hurault:** Anomalous thermal dependence of extra spots from a Ni(100) + $Sc(2\times2)$ low-energy-electron-diffraction pattern: a diffraction from the adsorbed layer only?
Solid State Comm. **13**, 993-996, 1973 (No. 7). *L*

G. E. Thomas & E. E. de Kluizenaar: The influence of reactive gases on the ion bombardment-induced light emission from surfaces.
Le Vide **28**, 190-193, 1973 (No. 167). *E*

A. C. W. Verdonk (Philips Materials Laboratory, Eindhoven) **& J. J. van den Broek:** Reaction of iron with $SiCl_4$ gas.
Oxid. Met. **8**, 1-9, 1974 (No. 1). *E*

H. W. Werner: A study on a mass spectrometer ion source by means of computer-calculated ion trajectories.
J. Physics E **7**, 115-121, 1974 (No. 2). *E*

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 2/3

Electrical conduction in ferromagnetic metals

A. R. Miedema and J. W. F. Dorleijn

It has long been known that ferromagnetic metals and alloys differ from other metals not only magnetically but also in their electrical behaviour. One of the noteworthy effects exhibited by the ferromagnetic metals is the anisotropy of the electrical resistivity: the resistivity depends on the angle between the current and the magnetization. This effect was investigated some years ago at these Laboratories by J. Smit. In recent years interest has returned to this effect in connection with its application to the read-out of magnetic tape and bubble memories. A good theoretical explanation of resistance anisotropy has not yet been given. It is therefore not possible to predict theoretically the behaviour of a given iron or nickel alloy or to say which alloy might be best for a given application. We give here a simple empirical model with which such predictions can be made for nickel alloys.

Introduction

The most characteristic property of metals is their high electrical conductivity; the conductivity of metals can be more than 10^{25} times that of insulators. In the present article we shall be concerned with electrical conduction in ferromagnetic metals, which is fundamentally different from that in other metals. We shall concern ourselves here only with the ferromagnetic metals nickel, iron and cobalt, and their alloys. It will be shown that the electrons that are responsible for electrical conductivity in ferromagnetic metals possess an extra degree of freedom. This can lead to special effects, which, though long known, have often not yet been properly explained. Examples are the anomalous Hall effect and the resistance-anisotropy effect [1], which has recently been applied in magnetic recording [2].

In a ferromagnetic metal the electrical resistivity depends on the angle between the current and the magnetization (*fig. 1*). If the ferromagnetic metal is in the form of a thin film as obtained by evaporation, the direction of magnetization in the plane of the film can be very easily changed. 'Easily' here means that only a weak external field is necessary, like that near the surface of the magnetic tape of a tape recorder. For the two extreme orientations of the magnetization, parallel or perpendicular to the current through the film, the electrical resistance may differ by several per cent (at room temperature): this is the resistance-anisotropy effect. Since this is a *percentage* effect, and the total resistance of such a ferromagnetic film, if sufficiently thin, can be relatively high, this effect is of importance for the detection of weak magnetic fields or their corresponding magnetizations. In magnetic recording an important aspect of the effect is that low-frequency alternating magnetic fields can be detected.

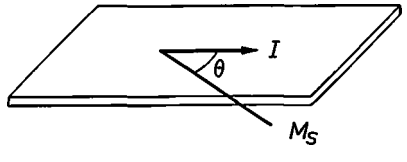


Fig. 1. The electrical resistivity of a specimen of a ferromagnetic metal depends, even at room temperature, on the angle θ between the current I and the magnetization M_S. For a thin specimen (a film) the direction of M_S can be changed even by a weak field.

Dr A. R. Miedema and Ir J. W. F. Dorleijn are with Philips Research Laboratories, Eindhoven.

[1] The history of resistance anisotropy in ferromagnetic metals goes back to 1857 (W. Thomson) and that of the anomalous Hall effect to 1893 (A. Kundt). One of the first more systematic investigations is that of J. Smit, Physica **17**, 612, 1951; see also H. C. van Elst, Physica **25**, 708, 1959.

[2] See for example R. P. Hunt, IEEE Trans. **MAG-7**, 150, 1971, and F. W. Gorter, J. A. L. Potgiesser and D. L. A. Tjaden, IEEE Trans. **MAG-10**, 899, 1974.

Magnetoresistivity (dependence of resistance on magnitude of magnetic field) and direction-dependent magnetoresistivity are also observed in non-magnetic metals (such as copper or aluminium) but only under certain (not entirely independent) conditions. A strong magnetic field is necessary, the temperature must be low and the metal must be very pure (e.g. a magnetic field of 10^4 Oe, the temperature of liquid helium and a metal with impurities $< 0.01\%$).

Our aim in this article is to explain why there is a difference between the electrical conductivity of ferromagnetic and ordinary metals. We shall start with a recapitulation of the modern solid-state-physics explanation of why metals are good conductors and why certain substances are metals and others not.

We shall discuss resistance anisotropy in some detail. Although the fundamental background of this effect is not yet adequately explained we have found a model that gives a very good description of the experimental results for nickel and its alloys. With the aid of this model it is now possible to predict very generally the magnitude of the resistance anisotropy for nickel alloys.

When is a substance a metal?

Using *fig. 2* and *fig. 3* we shall first show why some substances are conductors and others not. Our first example is lithium. Lithium is a monovalent element, that is to say, each atom has only one electron in its outer shell. For free, isolated atoms this electron is in a certain well defined quantum state (fig. 2*a*, *left*) which can be indicated on an energy scale at the value corresponding to the ionization energy of the free lithium atom. It is important to note that this quantum state may be occupied by two electrons because in each state the electron spin (angular momentum and corresponding magnetic moment) has two permissible directions. In free lithium atoms only one of the two spin directions is occupied.

If now, in imagination, we 'construct' a lithium crystal by bringing a number of isolated lithium atoms nearer to each other until they are at their normal lattice positions, then the collection of sharply defined quantum states (which are only half-occupied) is spread out into a band of states with different energies. For N Li atoms the band will be made up of $2N$ states. In this simple picture the centre of the band will lie at about the same level as the energy of the outer-shell electron in a free lithium atom. With N electrons, just half of the states in the band are of course occupied.

The existence of a partially filled band of states of the outer-shell electrons is essential for the occurrence of metallic properties.

Unconventionally, we shall illustrate this by first considering the effect of a magnetic field on the energy

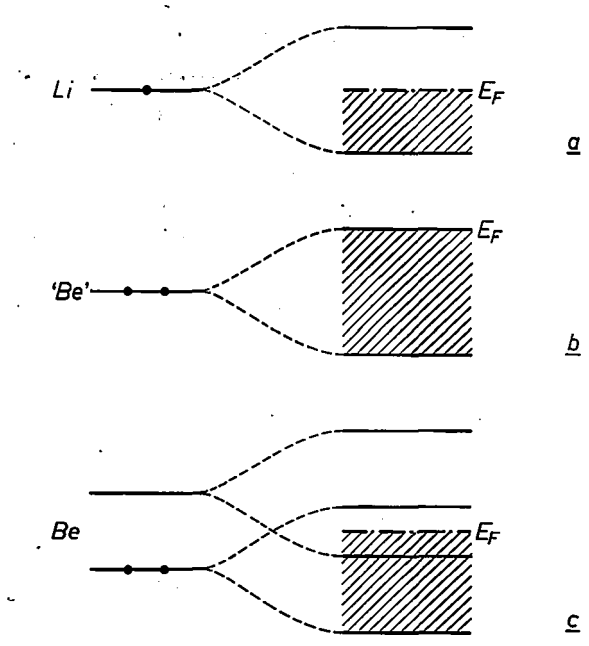


Fig. 2. The energy levels of electrons in a crystal (*right*) regarded as originating from the energy levels of a free atom (*left*). The energy scale is vertical. If N free atoms are brought together, the N originally identical energy levels are spread out into $2N$ different levels. The closer the atoms, the broader this band of energy levels. *a*) The situation for lithium. In the outermost shell of lithium there is only one electron; this shell is therefore only half-filled in the sense that for one energy there are two permissible orientations for the electron spin. The energy band of a lithium crystal is also half-filled. A crystal with a partly-filled band is a conductor. The shading shows the occupied energy levels. E_F is the Fermi energy. *b*) The analogue of (*a*) for an element with two electrons in the outer shell, such as beryllium. In the crystal the band is now full: there is no electrical conduction. *c*) The true situation in beryllium. A higher unoccupied level in the free Be atom is so spread out that the band overlaps the highest occupied energy level and therefore takes over a number of electrons. These bands are therefore not filled, as in (*b*), and electrical conduction is therefore possible.

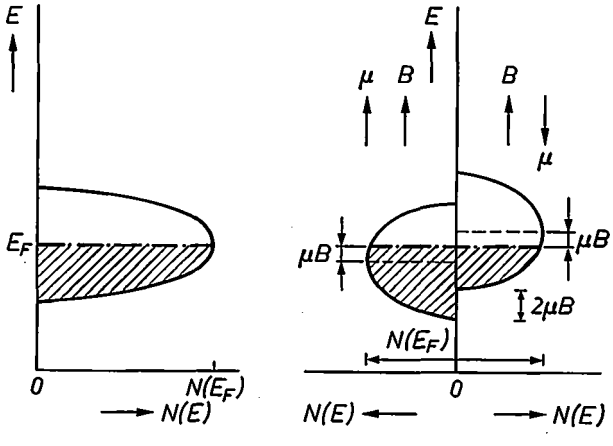


Fig. 3. *a*) The density of quantum states $N(E)$, i.e. the number of states per unit energy interval, of a monovalent metal. In the neighbourhood of $E = E_F$ the value of $N(E)$ is relatively large. *b*) If the metal is placed in a magnetic field of flux density B, the conduction band may be regarded as split into two sub-bands, one in which the electrons have their spin μ (magnetic moment) parallel to B (to the left of the origin), and the other in which μ and B are antiparallel (to the right). These bands are displaced relatively by an amount $2\mu B$, so that some electrons are transferred from the one sub-band to the other.

levels of the quantum states. We do this because the effect of an electric field can be usefully described in an analogous way, and also because ferromagnetic metals can be regarded as metals in which a strong internal magnetic field is always present. In fig. 3*a* we first show the conduction band of lithium as in fig. 2*a*, except that we now plot the density of states per unit energy $N(E)$ as a function of the energy. The total number of states for N atoms of lithium is $2N$. The greatest density of states occurs near the centre of the band, i.e. near the original energy level of the electron of the isolated atom. The energy to which the band is filled is called the Fermi energy E_F; the corresponding density of states is $N(E_F)$.

In fact we must again divide the electrons in the band into two groups. If a magnetic field is applied, an electron can occupy each state in two ways: with its magnetic moment μ parallel or antiparallel to the external flux density B. In an external field the two bands are relatively displaced over an energy $2\mu B$ (fig. 3*b*). The occupation of the electron states will now be such as to make the total energy a minimum. If the two sub-bands (distinguished by their spin direction) were originally equally filled, on applying the external field a number of electrons in the band with the 'wrong' spin direction will be transferred to states in the band where the magnetic energy is lower as shown in fig. 3*b*. The external flux density B induces a magnetization M which is proportional to the sum $N(E_F)\mu B$ of the two shaded regions and proportional to the magnetic moment μ per electron:

$$M = \mu^2 N(E_F) B.$$

In this way it can be seen that the magnetic susceptibility [3] of the conduction electrons in a metal, the Pauli susceptibility, is a measure of the density of states in the conduction band at the level of the Fermi energy E_F.

We shall now use the same picture to describe the effect of an electric field F applied in the longitudinal direction of a metallic specimen (of length L). When the electric field is applied, the quantum states at one end of the specimen are reduced in energy with respect to those at the other end. There arises in effect an excess of electrons at one end of the specimen and a deficiency at the other end. An electric current then flows which is proportional to this difference, to the electronic charge and also to the rate at which electrons can diffuse from the one end of the specimen to the other. The excess of electrons (*fig. 4*) is proportional to $N(E_F)$, just as in the magnetic case, and is also proportional to the potential difference FL. In addition the electrical conductivity is also proportional to the mean free path of the electrons in the diffusion process and inversely proportional to the length L. In a pure metal at low temperature the mean free path is large and the electrical resistance therefore low. In general the mean free path depends on the concentration of the impurities — i.e. the number of foreign atoms — and on the probability that an electron coming close to these foreign atoms is indeed scattered (the scattering cross-section).

In the literature the process of electrical conduction in metals is usually described in terms of the acceleration of the electrons by the electric field F between one collision and the next. This reasoning leads to the same result, as can be seen by analysing the situation in which an electric field is applied but no current is allowed to flow. In such a case a (small) concentration gradient of electrons is set up in the metal, but this has the result that the applied field is largely (but not entirely) compensated (leaving a residual field F'). The total current can, if desired, be regarded as being made up of two equal but opposite currents. One is the current due to the acceleration of the electrons in the field F'; the other is the diffusion current that would flow if diffusion in the electron-concentration gradient were considered separately.

In our example we considered the case of a monovalent element (Li). This was chosen deliberately. If an atom has two valence electrons, like beryllium, then there is no *a priori* certainty that the substance will be metallic. The energy-level diagram of a divalent element is shown in fig. 2*b* ('Be'). For free atoms there are again two possible states per atom, both of the same energy, but these are now both occupied. If the free atoms are brought together as a crystal, forming an energy band, all the states in the band will be filled. It is no longer possible for an excess of electrons to arise in a magnetic or electric field. There can be no such excess because $N(E_F)$ is

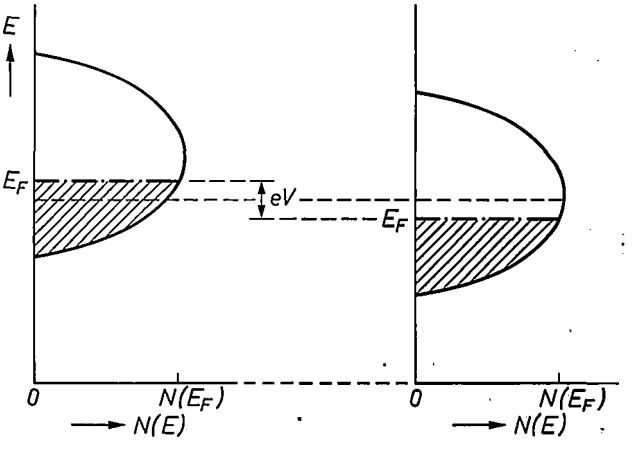


Fig. 4. The conduction band at the two extremities of a specimen (length L) across which a voltage is applied. The conduction of electricity can be considered as a diffusion of electrons from left to right under the influence of the electric field V/L.

[3] Besides the paramagnetic susceptibility calculated here there is also a negative contribution from the diamagnetism of the free electrons. This reduces the result to 2/3 of the value given.

zero; there is no Pauli susceptibility and no metallic conduction.

In reality the element beryllium *is* a metal. As explained in fig. 2*c* this is because a free Be atom has available a number of further quantum states (e.g. those of p electrons) which are not occupied, at levels close to those of the occupied s states. If we attempt to picture the transition from free beryllium atoms to solid-state beryllium, then both the occupied and the empty levels would be spread out to a band. In beryllium these bands overlap, giving a combined band with more than enough states for the $2N$ electrons; beryllium is therefore a metal. $N(E_F)$ is thus not necessarily zero for an element with two valency electrons per atom.

Thus, whether the crystal produced is a metal or an insulator depends on the energy levels of the states of the free atoms and on how much these states are spread out (and mixed) when the atoms are brought together. An insulator is possible only if there are an even number of valency electrons per unit cell. If a band is partly filled, as will be the case when there are overlapping bands, then we have a metal. If a band is full and there is a large difference in energy between it and the next empty band, then the substance is an insulator. If the energy difference between a filled band (at absolute zero) and a higher energy band is so small that at higher temperatures the charge carriers can be excited thermally from the full band into the empty band (*fig. 5*), then the substance is a semiconductor. If two bands just overlap at the level of the Fermi energy, so that $N(E_F)$ is very small, the substance is called a semimetal. Bismuth, antimony and grey tin are examples of semimetals. The value of $N(E_F)$ for bismuth is about 10^6 times smaller than that for an ordinary metal such as copper.

From the foregoing it is evident that although $N(E_F)$ is not zero for all metals, this quantity can assume widely different values in different metals. On the basis

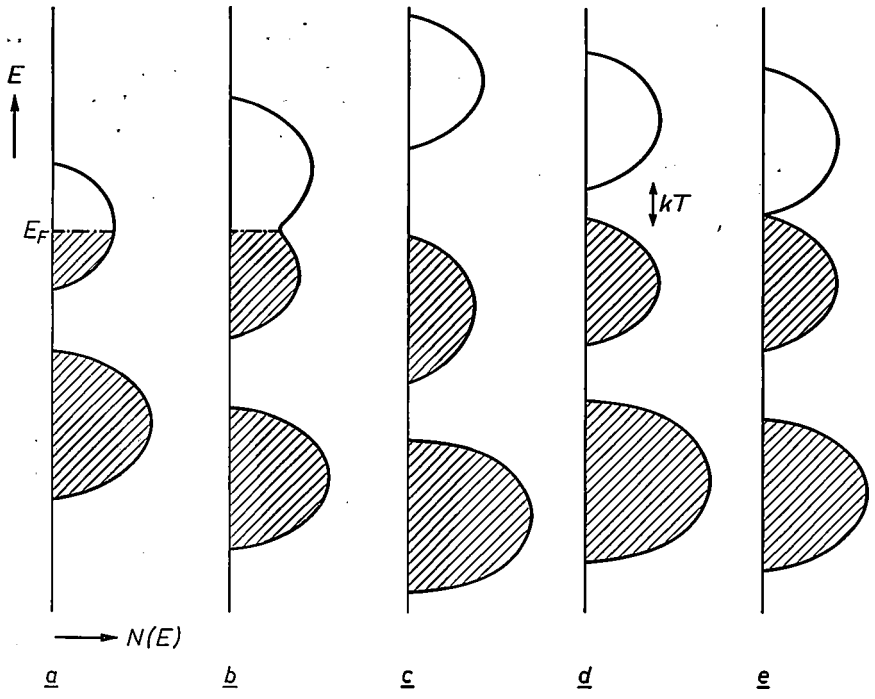


Fig. 5. Energy diagrams as in fig. 3 ($N(E)$ *vs* E). *a*) Metal in which only one energy band is of importance for the conduction. *b*) Metal in which two bands are involved. *c*) Insulator. *d*) Semiconductor. *e*) Semimetal. The semiconductor differs from the insulator only because the empty band is within the reach of a small number of electrons by thermal excitation.

of the proportionality between conductivity, electron excess and $N(E_F)$ as suggested by fig. 4, it might be expected that the conductivity of different metals at the same temperature would also have widely differing values. In fact, however, the differences between the conductivities of the metals are fairly modest, as can be seen from *Table I.*

For a good comparison between the conductivity of the various metals the values of the resistivity should be compared not at some fixed arbitrary temperature (room temperature in Table I) but at the temperature at which the same number of lattice vibrations are excited, i.e. at the Debye temperature Θ_D (the resistivity of each metal at its Debye temperature is given between brackets in Table I). It can be seen that for bismuth,

Table I. The electrical resistivity (in μΩ cm) of a number of pure metals at room temperature. The metals are arranged as in the periodic table of the elements. The value between brackets is the resistivity at the Debye temperature for each element [4].

Li 9.3 (11)	Be 3.2 (15)					
Na 4.7 (2.5)	Mg 4.3 (6)			Al 2.75 (4.0)		
K 7.2 (2.2)	Ca 3.6 (2.8)	Cu 1.70 (2.0)	Zn 5.9 (6.5)	Ga 14.8 (16)		
Rb 12.5 (2.4)		Ag 1.61 (1.2)	Cd 7.3 (5)	In 8.7 (3.2)	Sn 11.0 (7.5)	Sb 41 (30)
Cs 20 (2.6)		Au 2.20 (1.2)		Tl 16 (4)	Pb 21 (7.5)	Bi 116 (46)

for which $N(E_F)$ was a factor of 10^6 times smaller than for copper, the resistivity is only about 10 times larger.

This apparent contradiction arises because $N(E_F)$ is also of significance for the conductivity in another way: via the scattering process. The mean free path of the conduction electrons, which is one of the factors that determine the conductivity, depends not only on the temperature and the purity of the metallic element but also on the probability that an electron passing near an impurity atom is in fact scattered, as we noted earlier. We are concerned here with the concept of transition probability. Scattering of the electron changes its direction of motion; in principle the electron goes to another state of different energy within the conduction band. This is possible only if there are unoccupied states in the neighbourhood of the Fermi level. The number of such states and hence the transition probability is proportional to $N(E_F)$.

We now see that $N(E_F)$ enters into our description of electrical conductivity twice. The larger $N(E_F)$, the larger the number of electrons contributing to the conduction process but also the higher the probability that a given electron is scattered by an impurity atom or by a lattice vibration, i.e. the shorter the mean free path. These two effects of $N(E_F)$ largely compensate each other so that the electrical resistivity of semimetals such as bismuth and antimony is not so very different from that of ordinary metals.

Electrical conduction in the transition metals

In this section we examine more closely the group of metals that include the ferromagnetic metals — the subject of this article. This group is that of the transition metals, the central group in the periodic system (the three rows $Z = 21$ to 28, 39 to 46 and 57 to 78).

In a transition metal the conduction is not entirely due to the 'conventional' conduction electrons, which can be considered to arise from the atomic s and p electron states, as in non-transition metals: the d electrons also have a part to play. In fact d electrons are also present in copper, silver and gold. The energy bands of these three metals are shown in *fig. 6*. The energy scale is stretched compared with the previous figures and the $N(E)$ scale is compressed. The conduction band, which will contain the s electrons, is only partly filled: copper, silver and gold are in fact good metallic conductors. In addition, however, there is an energy band containing states that can be thought of as originating from the atomic d electron states; copper has ten 3d electrons and for silver and gold there are ten 4d or 5d electrons. These 10 states form a (narrow) band, which is completely filled. The presence of the d band therefore has no effect on the conduction properties. The d band does however contribute to the cohesion of the atoms in the crystal (relatively high heat of evaporation) and its presence accounts for the characteristic colours of copper and gold [5].

In *fig. 7* we consider the elements lying to the left of Cu, Ag, Au in the periodic system: Pd, Pt and paramagnetic nickel. The energy-level diagrams can again be constructed from a conduction band derived from s- and p-like atomic states and a band corresponding to the d states. Now, however, the d band is not full; Pd, Pt and Ni have ten valence electrons which are distributed over the conduction band and the d band in such a way that the energy is at a minimum. In effect these three metals have about 9.5 electrons in the d band and 0.5 electrons in the s,p band.

What is the effect now of the unfilled d band on the electrical resistance? At first sight it would appear that the resistance ought to be lower since, besides the conduction-band electrons, the d electrons should now also make a contribution to the conductivity. This contribution is virtually absent; d electrons move slowly (they have a large effective mass) and, because $N(E_F)$ is large, they have a short mean free path. The unfilled d band actually increases the resistance. The s,p conduction electrons can also be scattered into

[4] Resistivities taken from G. T. Meaden, Electric resistance of metals, Plenum Press, New York 1965. The values at the Debye temperatures are from C. Kittel, Introduction to solid state physics, 4th edition, Wiley, New York 1971.

[5] Light of a wavelength such that it can excite electrons from the d band into a non-occupied state in the conduction band at an energy greater than E_F will be relatively strongly absorbed. The energy difference ΔE between the top of the d band and E_F (see fig. 6) determines the absorption edge. For copper and gold ΔE is about 2 eV, which corresponds to a wavelength of 500 nm. Light of wavelength shorter than this is strongly absorbed by copper and gold; the reflected light has the complementary colour, i.e. yellow-red. In silver the d band is deeper; ΔE is about 4 eV ($\lambda \doteq 250$ nm), so that no selective absorption or reflection occurs in the visible region.

states in the d band as well as into states with energies near E_F in the conduction band itself. This increases the transition probability for scattering and hence the resistance. For this reason the electrical resistivity of the transition metals is generally higher than that of the non-transition elements. If the d band were absent all the transition metals would have a resistivity comparable with that of Cu, Ag or Au.

As can be seen in *Table II*, the resistivity of the transition metals is not only considerably higher than that of Cu, Ag and Au, but there are also wide variations. The reason for this lies in the structure of the d band. If $N(E)$ has a minimum near to the Fermi energy then the resistivity may be fairly low, as is the case for Mo and W.

Finally, we shall consider the difference between the ferromagnetic transition metals Ni, Co and Fe and a paramagnetic transition metal. As already explained (fig. 3) the energy bands of fig. 6 and fig. 7 ought really to be divided into two sub-bands depending on the orientation of the magnetic moment of the electron. This division can be observed in a magnetic field: the two bands acquire a difference in energy and because of the high value of $N(E_F)$ the paramagnetic Pauli susceptibility becomes very much larger in the transition metals than in the non-transition metals. In a ferromagnetic metal, even in the absence of an external magnetic field, the occupation of the two sub-bands of opposite spins is unequal (*fig. 8*). In nickel, for example one of the reasons for this is that there is an electrostatic repulsion interaction between the d electrons, which has a magnitude different for electrons with parallel spins than for electrons with antiparallel spins. It turns out that the energy is at a minimum when there are different numbers of d electrons in the two spin directions.

We see here that the magnetic moment (per atom) of a ferromagnetic metal can assume any value, i.e. it need not be an integral number of Bohr magnetons. The ferromagnetic moment is given by the difference in occupation between the two sub-bands of oppositely directed spins. In nickel one sub-band is exactly full; the other is occupied to the extent of 4.4 electrons per atom so that the ferromagnetic moment is approximately equal to 0.6 Bohr magnetons per atom.

It will now be clear that electrical conduction in ferromagnetic metals is a complicated matter. Firstly, the contribution from the electrons in the conduction band depends strongly on the scattering to the d band. Secondly, there are two d sub-bands in which the density of states may differ very appreciately. Finally, there is a further complication, or perhaps it might be called a simplification. N. F. Mott [6] suggested that for the electrical resistance of ferromagnetic metals a

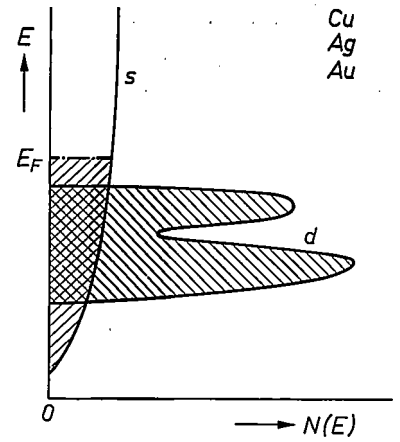


Fig. 6. Energy diagrams for the metals copper, silver and gold. In addition to the conduction band, which contains s-like electrons, there is a relatively narrow band of d states. This band is completely filled and makes no contribution to the conduction.

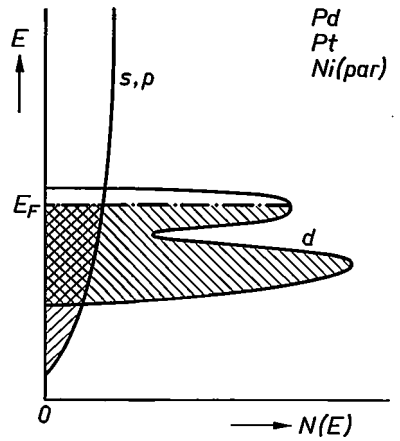


Fig. 7. Energy diagram of a transition metal (for Pd, Pt or Ni above the Curie temperature, i.e. paramagnetic). The situation is very similar to that of fig. 6, except that the d band is now not filled.

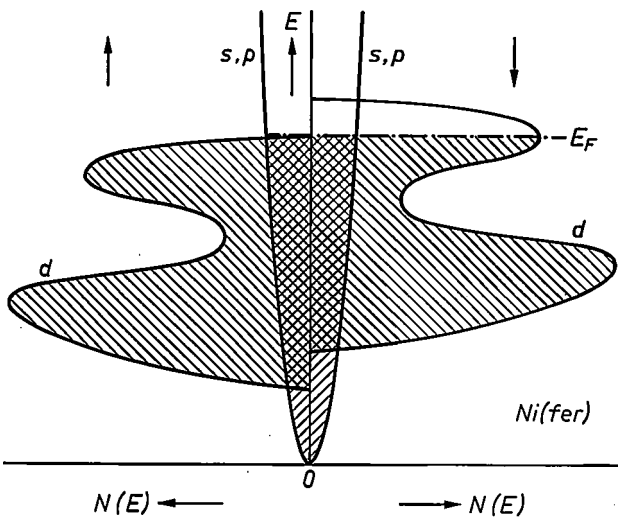


Fig. 8. Energy diagram of a ferromagnetic transition metal (Ni). The two possible orientations of the electron spins make it necessary to regard the band as split into two sub-bands. Even in the absence of a magnetic field these two sub-bands are not equally filled (see fig. 3*b*). The two s,p sub-bands remain to a first approximation equally filled.

sharp distinction should be made between scattering processes in which the spin reverses and processes in which the direction of spin is maintained. Scattering with spin reversal is less probable. According to Mott's suggestion there are two parallel electron streams when a current flows in a ferromagnetic metal: those with the spin parallel to the ferromagnetic magnetization and those antiparallel (in fig. 8 those on the left are parallel, those on the right antiparallel). These two currents are subject to quite different resistances; for the electrons on the left of fig. 8, which we shall now call ↑-electrons, there is very little possibility of being scattered to a d state (and therefore the resistance of these electrons is small), whereas the possibility that ↓-electrons are scattered into a d state is large. It is interesting now to consider how the existence of two independent contributions to the electric current in nickel, for example, can be demonstrated. In the next section we indicate that this can be done through a study of the electrical resistance of ternary nickel alloys.

The residual resistivity of ternary alloys

The residual resistivity of a metal is the value of the resistivity at such a low temperature that the contribution from the lattice vibrations is negligible. The residual resistivity of a given metal depends on the nature and concentration of the lattice imperfections, e.g. impurities. The residual resistivity caused by a given impurity can be characterized by the value per atomic per cent of that impurity, the 'residual resistivity' ϱ_A. The total resistivity caused by metal A, present in the concentration c_A, is then $c_A\varrho_A$. The residual resistivity increases the more the impurity metal differs from the host metal. Differences in valency and in atomic volume, for instance, are important. If we take as an example the residual resistivity due to various impurities in copper (*Table III*), then we see that the residual resistivities of neighbours to Cu in the periodic table (Ag, Zn and Cd) are indeed relatively small. We note that the resistivity caused by one per cent of a dissolved metal is comparable with the resistivity produced in pure copper by lattice vibrations at room temperature and in some cases is even much larger.

If two elements are dissolved in copper their separate residual resistivities may be added up, provided that the total concentration of the two elements is not too large, i.e. that the host metal is still effectively copper. This is Matthiessen's rule:

$$\varrho = c_A\varrho_A + c_B\varrho_B. \qquad (1)$$

If the host metal is ferromagnetic, the residual resistivity due to impurity metals cannot be characterized by only one quantity: the residual resistivity for the ↑-band can differ very considerably from that for the ↓-band. The residual resistivity for element A in nickel (for example) is in fact the equivalent resistance of the two parallel resistances $\varrho_A{}^\uparrow$ and $\varrho_A{}^\downarrow$ (*fig. 9*). We have:

$$\varrho_A = \varrho_A{}^\uparrow\varrho_A{}^\downarrow/(\varrho_A{}^\uparrow + \varrho_A{}^\downarrow). \qquad (2)$$

If we now dissolve two impurities, metals A and B, in a ferromagnetic metal, then Matthiessen's rule is valid for each of the two currents:

$$\varrho^\uparrow = c_A\varrho_A{}^\uparrow + c_B\varrho_B{}^\uparrow,$$
$$\varrho^\downarrow = c_A\varrho_A{}^\downarrow + c_B\varrho_B{}^\downarrow. \qquad (3)$$

However, we cannot calculate the equivalent resistance

Table II. The resistivity (in μΩ cm) of pure transition metals at room temperature compared with that of the metals Cu, Ag and Au (right-hand column). The values between brackets are again the values at the Debye temperature [4].

Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu
47	43	20	13	139	9.8	5.8	7.0	1.70
(58)	(62)	(26)	(28)	(195)	(16)	(9)	(11)	(2.0)
Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag
58	42	14	5	14	1.4	4.8	10.5	1.61
(55)	(41)	(13)	(8)		(15)	(8)	(10)	(1.2)
La	Hf	Ta	W	Re	Os	Ir	Pt	Au
79	31	13	5	19	9.1	5.1	10.4	2.20
(38)	(27)	(10)	(7)	(28)	(16)	(7)	(9)	(1.2)

Table III. The residual resistivity of impurity metals (in μΩ cm per atomic %) dissolved in copper [7] and in nickel [8–10]. The upper value refers to copper alloys, the next to the resistivity of the ↑-band and the lowest to the ↓-band of the conduction electrons in nickel alloys.

A							Mg	Al	Si
ϱ_A(Cu)							0.8	0.95	3.1
$\varrho_A{}^\uparrow$(Ni)								3.6	
$\varrho_A{}^\downarrow$(Ni)								5.2	
V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge
–	4.0	2.9	9.3	6.9	1.1	–	0.3	1.4	3.7
11	39	0.86	0.53	0.24	–	1.28			
7.2	5.6	4.2	4.0	2.9	–	4.1			
Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn
–	–	–	–	4.4	0.95	0.2	0.3	1.1	3.1
16[9]	29[9]		54	7.7	0.30[10]				4.8
7[9]	8[9]		5.5	2.3	0.30[10]				6.2
Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb
–	–	–	–	6.1	2.0	0.55	1.0	–	3.3
16[9]	17[9]	26	59[9]	35[9]	3.9	0.47			
8[9]	7[9]	6.4	6[9]	5[9]	1.0	2.0			

[6] N. F. Mott, Proc. Phys. Soc. 47, 571, 1935.
[7] F. J. Blatt, Physics of electronic conduction in solids, McGraw-Hill, New York 1968, p. 199.
[8] J. W. F. Dorleijn and A. R. Miedema, J. Physics F 5, 487, 1975, and 1975, in press.
[9] J. Durand and F. Gautier, J. Phys. Chem. Solids 31, 2773, 1970.
[10] T. Farrell and D. Greig, J. Physics C 1, 1359, 1968. The experimental values of the sub-band resistivities in [9] and [10] are derived from the temperature dependence of the resistivity of binary alloys.

in the same way as (2). This can be seen from *fig. 10*. To make Matthiessen's rule valid, the switch S would have to be closed. Only in that case are the sub-band resistivities due to A and B in series as well as the equivalent resistivities. When the switch is open the total resistance is generally higher than when the switch is closed; we can therefore conclude that in ternary nickel alloys only positive deviations from Matthiessen's rule (1) ought to occur.

The following calculation illustrates the point nicely. We assume that 1% of the metal A in nickel gives a resistance 10 in the ↑-band and a resistance 1 in the ↓-band, whilst for 1% of metal B the opposite applies. The equivalent resistance and hence the residual resistance for each element separately in Ni is then 0.9. If 0.5% of A and 0.5% of B are now dissolved in nickel then the residual resistance in both the ↑-band and the ↓-band is 5.5; the equivalent resistance is then 2.75. We see that while 1% of each metal separately

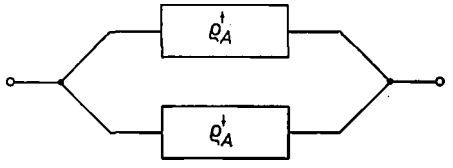


Fig. 9. The residual resistivity of a binary ferromagnetic alloy (metal A in nickel) in the two-current model. When electrons are scattered by the A atoms (the origin of the residual resistivity) the spin orientation remains unchanged, so that two kinds of moving electrons may be distinguished, each of which experiences resistance. The corresponding resistivities are denoted by $\varrho_A{\uparrow}$ and $\varrho_A{\downarrow}$. These resistivities are sometimes very different.

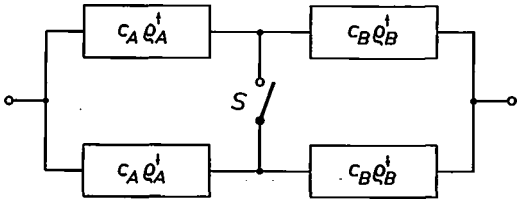


Fig. 10. Equivalent circuit (see fig. 9) of a ternary nickel alloy. c_A is the concentration of element A, c_B that of B. Matthiessen's rule is not valid here; it would only be valid if the switch S were closed, which it is not.

gives a resistance of 0.9, 1% of an equal mixture of the two gives a resistance that is at least three times larger.

Measurements corresponding to the foregoing case are plotted in *fig. 11*. The residual resistivity of ternary nickel alloys, with Co and Rh, is plotted against concentration. The sum of the Co and Rh concentrations is held constant at 3 atomic %. The dashed line is the curve to be expected from Matthiessen's rule. In fact the resistance of the ternary alloys is always higher, as would be expected. From the measurements the four

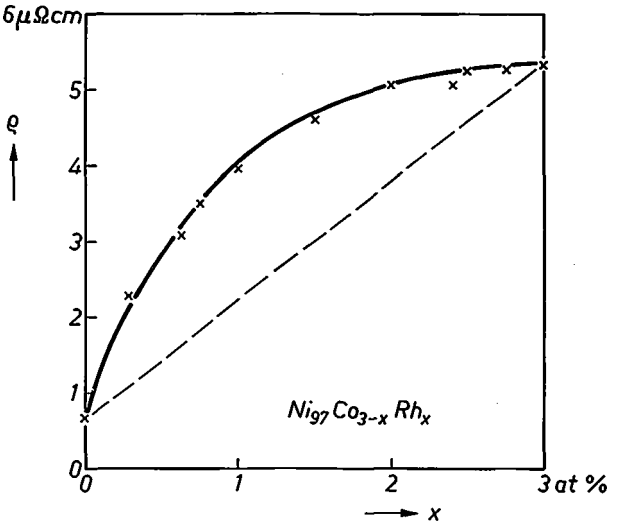


Fig. 11. Residual resistivity of ternary alloys of nickel with Co and Rh. The total concentration of Co and Rh taken together is maintained at 3%. The dashed line corresponds to Matthiessen's rule.

relevant parameters $\varrho_{Co}{\uparrow}$, $\varrho_{Co}{\downarrow}$, $\varrho_{Rh}{\uparrow}$ and $\varrho_{Rh}{\downarrow}$ can be derived that determine the electrical behaviour of Co and Rh in nickel (*Table IV*). The solid line in fig. 11 is calculated by using values of these parameters that give the best fit.

Experimental results such as those of fig. 11 can be regarded as a proof of the existence of two-current conduction in nickel. The deviations from the linear behaviour are very large and other explanations for the deviations are not at all obvious. Further convincing evidence for the correctness of the explanation of the deviations from Matthiessen's rule can be obtained by repeating the experiment for alloys of nickel and cobalt and another third element. If the model is correct and the measurements reliable, the results for $\varrho_{Co}{\uparrow}$ and $\varrho_{Co}{\downarrow}$ should not depend on the nature of this third element. This is in fact true within reasonable limits (see Table IV). A fairly complete set of resistivity values derived from this type of experiment for different solute metals in nickel is given in Table III.

It is interesting to consider why one impurity tends to scatter the conduction electrons in the ↑-sub-band while another gives the greatest resistance in the ↓-band. (The energy diagram for nickel (fig. 8) shows which of the two sub-bands gives the larger resistance). The answer to this question can be found in local deviations in the values of $N^{\uparrow}(E_F)$ and $N^{\downarrow}(E_F)$ occurring at the positions of the impurity atoms responsible for scattering. Lattice vibrations as well as impurities such as the neighbours of nickel in the periodic table (Co and Cu) give the largest resistance in the band that would be expected from fig. 8 (the right-hand ↓-band). If on the other hand a metal far from nickel in the periodic table is dissolved in nickel, there need

Table IV. The residual resistivities $\varrho\uparrow$ and $\varrho\downarrow$ ($\mu\Omega$ cm per atomic %) of cobalt and rhodium dissolved in nickel, derived from measurements on groups of ternary nickel alloys [11] (5 or 6 values of x per group). The values depend only slightly on the nature of the third element present in the alloy.

Alloy	$\varrho_{Co}\uparrow$	$\varrho_{Co}\downarrow$	Alloy	$\varrho_{Rh}\uparrow$	$\varrho_{Rh}\downarrow$
$Ni_{97}Co_xRh_{3-x}$	0.264	2.52	$Ni_{97}Co_xRh_{3-x}$	8.13	2.30
$Ni_{97}Co_xRu_{3-x}$	0.261	3.06	$Ni_{97}Au_xRh_{3-x}$	4.61	2.98
$Ni_{97}Co_xAu_{3-x}$	0.259	1.59	$Ni_{97}Cu_xRh_{3-x}$	7.31	2.40
$Ni_{97}Co_xCr_{3-x}$	0.21	4.3	$Ni_{97}Ru_xRh_{3-x}$	7.62	2.23

Table V. The resistance anisotropy for the two currents in a number of binary nickel alloys. The elements are arranged in order of decreasing value of the ratio $\alpha = \varrho\downarrow(//)/\varrho\uparrow(//)$.

Impurity element	α (see eq. 5)	$\Delta\varrho\uparrow/\varrho\uparrow$ (%)	$\Delta\varrho\downarrow/\varrho\downarrow$ (%)
Co	12.0	+ 14.5	— 1.7
Fe	7.5	+ 14.3	— 0.7
Mn	4.9	+ 9.1	— 1.2
Au	4.2	+ 8.5	— 0.5
Cu	3.2	+ 9.7	— 2.5
Al	1.4	+ 7.5	— 1.2
Sn	1.3	+ 6.1	— 1.3
Ti	1.1	+ 4.4	— 3.4
V	0.66	+ 7.3	— 4.5
Rh	0.30	+ 7.5	— 2.3
Pt	0.25	+ 4.2	— 0.2
Re	0.25	+ 6.0	— 1.7
Cr	0.15	+ 5.3	— 1.1
Ru	0.10	+ 5.7	— 1.1

be no relation between the local and the total densities of states. For a metal such as chromium in nickel there is the difficulty that as a free atom chromium has available only 5 of the 10 possible d electrons. Chromium atoms in nickel rather resemble free chromium atoms, as in the picture of fig. 2. Because atomic chromium is quite different from atomic nickel, a local density of states can arise for d electrons near the chromium atoms that is very different indeed from that of the pure nickel. For the scattering of conduction electrons by chromium in nickel these deviations in the local density of states are of considerable significance.

Resistance anisotropy

In the introduction we mentioned that the resistance of a ferromagnetic metal depends on the angle θ between the current and the magnetization. In an investigation of the resistivity of nickel alloys, as described in the previous section, it is therefore important that this angle is always the same during the measurements. Only results measured with $\theta = 0$ have so far been used, i.e. the current parallel to the magnetization (and to the external magnetic field).

It is interesting to repeat the analysis for the case in which the current and the magnetization are at right angles. The characteristic residual resistivities found in the previous section should have been denoted by $\varrho_A\uparrow(//)$ and $\varrho_A\downarrow(//)$. We now find the residual resistivity for the sub-bands for the orthogonal configuration $\varrho_A\uparrow(\perp)$ and $\varrho_A\downarrow(\perp)$. Now it is interesting to consider whether the anisotropy in the resistance occurs equally in the ↑-band and the ↓-band, or whether it depends substantially on one band only.

Table V summarizes the results of our analysis. We can express the results in terms of the quantities $\Delta\varrho_A\uparrow/\varrho_A\uparrow$ and $\Delta\varrho_A\downarrow/\varrho_A\downarrow$, which are defined as:

$$\Delta\varrho_A\uparrow/\varrho_A\uparrow = \{\varrho_A\uparrow(//) - \varrho_A\uparrow(\perp)\}/\varrho_A\uparrow(//), \quad (4a)$$

$$\Delta\varrho_A\downarrow/\varrho_A\downarrow = \{\varrho_A\downarrow(//) - \varrho_A\downarrow(\perp)\}/\varrho_A\downarrow(//). \quad (4b)$$

The tabulated results exhibit a clear pattern. For all elements in nickel the anisotropy is positive for the ↑-sub-band and negative for the ↓-sub-band. In approximate terms $\Delta\varrho\uparrow/\varrho\uparrow$ lies in the region of 10% while the mean value of $\Delta\varrho\downarrow/\varrho\downarrow$ is about —2%. The approximately constant values of $\Delta\varrho/\varrho$ for each sub-band for different elements are particularly surprising since there are large differences in the actual resistivity values (see Table III).

If $\Delta\varrho\uparrow/\varrho\uparrow$ and $\Delta\varrho\downarrow/\varrho\downarrow$ were true constants, there would be a simple relation between the experimentally measured resistance anisotropy and the ratio α of the total resistivities of each of the sub-bands:

$$\alpha = \varrho\downarrow(//)/\varrho\uparrow(//), \quad (5)$$

$$(\Delta\varrho/\varrho)_{exp} = \frac{\alpha}{1+\alpha}\left(\frac{\Delta\varrho\uparrow}{\varrho\uparrow}\right) + \frac{1}{1+\alpha}\left(\frac{\Delta\varrho\downarrow}{\varrho\downarrow}\right). \quad (6)$$

For metals that preferentially scatter the ↓-conduction electrons, α is large; the observed effect is then substantially the same as that of the ↑-band only. If α is small, the converse is true. For constant $\Delta\varrho\uparrow/\varrho\uparrow$ and $\Delta\varrho\downarrow/\varrho\downarrow$, the experimental resistance anisotropy, plotted as a function of α, will give a hyperbola.

We know from our analysis the values of $\varrho\uparrow$ and $\varrho\downarrow$ for most elements in nickel, and we can therefore calculate α for our ternary alloys from (3) and (5). In *figs. 12* and *13* the measured anisotropy at 4.2 K is plotted against the calculated α. Fig. 12 refers to the results for small α. The circles denote results for ternary alloys; the black dots relate to measurements on binary alloys. The correlation between the anisotropy and the value of α as given by (6) can be recognized although the curve is not an exact hyperbola. The correlation appears to be so good, even though $\Delta\varrho\uparrow/\varrho\uparrow$ and $\Delta\varrho\downarrow/\varrho\downarrow$ are only very approximately constants, largely because the deviations of $\Delta\varrho\uparrow/\varrho\uparrow$ from

[11] The values for $Ni_{97}Co_xCr_{3-x}$ (left-hand column, below) are taken from P. Leonard, M. C. Cadeville, J. Durand and F. Gautier, J. Phys. Chem. Solids **30**, 2169, 1969. Those for the seven other alloys are from J. W. F. Dorleijn and A. R. Miedema [8].

the mean value vary systematically with α; see Table V. We have no explanation for this, but such a simple correlation between α and the anisotropy is of course most welcome for prediction of the resistance anisotropy in alloys for which there are no experimental data.

Fig. 13 shows that the resistance anisotropy of nickel alloys at low temperatures is largest (about +13%) for the alloying elements Fe and Co. This should not be taken to mean that Ni-Co and Ni-Fe alloys will be the most suitable materials for practical application of the anisotropy effect since applications of the effect are at room temperature.

The situation at room temperature differs from that at low temperatures mainly because the conduction electrons are now scattered not only by impurities and their defects but also by phonons (lattice vibrations). For our way of looking at the situation this is not a serious complication because we can also treat the phonons in much the same way as the alloying elements in the nickel [12]. A binary alloy of Ni and Co at high temperatures is thus comparable with a ternary alloy of Ni and Co and (say) Cu at low temperature. For nickel at room temperature (300 K) the resistivities due to phonons are $\varrho_{ph}{}^{\downarrow} = 27\ \mu\Omega$ cm and $\varrho_{ph}{}^{\uparrow} = 6.7\ \mu\Omega$ cm. We have then $\alpha = 4$. The contribution from the phonons is therefore not too great a disadvantage in the application of the magnetoresistance-anisotropy effect.

A second difference from the situation at low temperature is more serious. So far we have assumed two completely independent currents in nickel. At high temperatures this is no longer justifiable: it is also necessary to take account of scattering processes in which the magnetic moment of the conduction electrons is reversed, e.g. scattering of electrons by magnetic thermal excitations, i.e. spin waves. The probability of these spin-reversal processes increases with the temperature, approximately as T^2. This scattering probability can be expressed as a resistivity for spin-reversal processes $\varrho^{\uparrow\downarrow}$. At room temperature the value of $\varrho^{\uparrow\downarrow}$ for nickel is about 11 $\mu\Omega$ cm. From this it may be concluded that for nickel alloys with a total resistivity of about 10 $\mu\Omega$ cm the resistance-anisotropy effect at room temperature will be considerably weaker than at low temperatures. If the resistivity of the dissolved metal is much larger than 10 $\mu\Omega$ cm, then at room temperature the effect will be almost as large as at low temperature, as can be seen from fig. 13. The effect of spin reversal can be included in eq. (6) by introducing an effective value of α, which can be found from:

$$\alpha = \{\varrho^{\downarrow}(//) + 2\varrho^{\uparrow\downarrow}\}/\{\varrho^{\uparrow}(//) + 2\varrho^{\uparrow\downarrow}\}. \qquad (7)$$

With $\varrho^{\uparrow\downarrow} = 11\ \mu\Omega$ cm at room temperature and the phonon resistivities given above and the residual resistivities given in Table III, the resistance anisotropy of any dilute nickel alloy can be calculated [13].

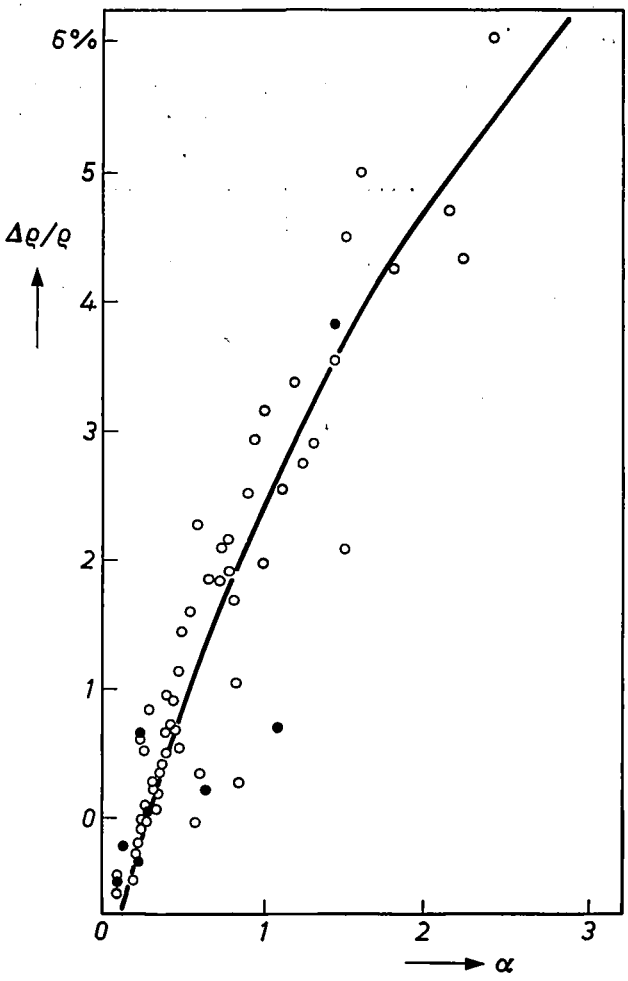


Fig. 12. The resistance anisotropy (see eq. 4) of nickel alloys at 4.2 K plotted as a fuction of α, i.e. the ratio of the resistivities in the two sub-bands (see eq. 5). The circles refer to ternary alloys, the black dots to binary alloys. The measured points agree only approximately with the predictions from eq. (6).

From Table III it can be deduced that the residual resistivity of Fe or Co in nickel, each of which have such a high anisotropy (see fig. 13), is relatively small, so that the requirement for a residual resistivity much larger than 10 $\mu\Omega$ cm cannot be fulfilled with practicable percentages of Co and Fe. It is not therefore possible to state immediately which alloy will have the largest anisotropy effect at room temperature.

Equation (7) and fig. 13 are even more important if, in a particular application of the resistance-anisotropy effect, it is necessary to have not only a high value for the effect, but also other requirements for the material. These could include a high resistivity, low magnetization (i.e. easy rotations of the direction of magnetization of a layer) or a minimum value of the magnetostriction. In all such cases the parameters and relations used in this section can be successfully employed to predict the anisotropy effect.

Other alloys; other conduction-related effects in ferromagnetics

Although we have mainly considered nickel alloys, the above approach should in principle be valid for all ferromagnetic metals. Comprehensive experimental results on the resistance anisotropy are however only available for nickel alloys; for iron and cobalt little more is known than that the effect does exist in the pure metals. However, it has been convincingly shown from an analysis of the temperature dependence of the resistivity that two distinct parallel currents occur in alloys based on Co or Fe.

For nickel the slope of the curve at the higher field-strength corresponds to the conventional Hall effect; extrapolation of this part of the curve to $B = 0$ gives the anomalous Hall effect. A noteworthy property of the metals showing the anomalous Hall effect (investigated in some detail at Philips Research Laboratories in the fifties [15] [16]) is that alloys with a high resistivity also have a large Hall effect. Over a considerable range of resistivity values, the anomalous Hall effect varies approximately as ϱ^2. In pure nickel it vanishes completely at low temperatures [15]. In dilute nickel alloys at low temperatures the effect is strongly dependent on

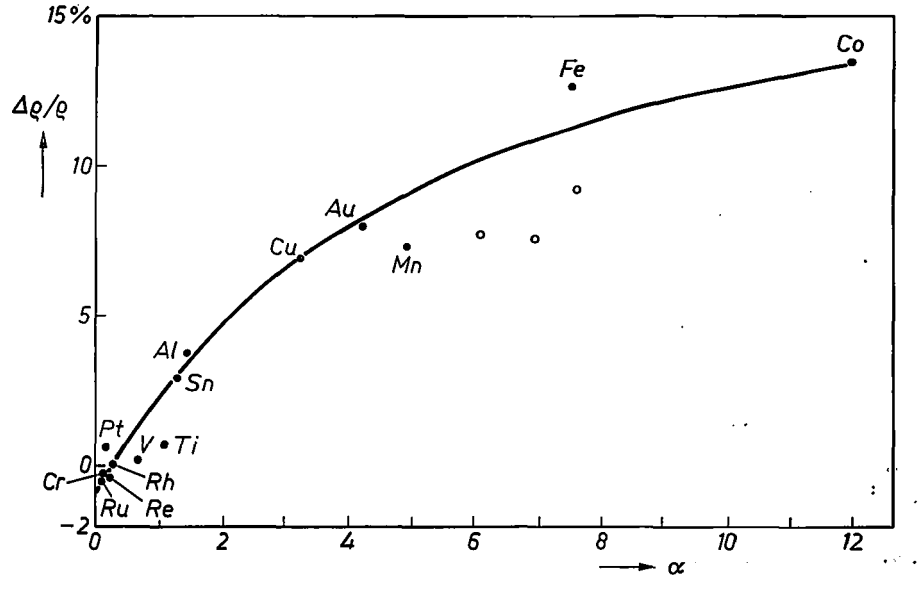


Fig. 13. As fig. 12 but for larger values of α.

Finally, we shall briefly discuss the other remarkable conduction effects that can be observed in ferromagnetic metals. The anomalous Hall effect has been mentioned in the introduction. In a non-ferromagnetic metal the Hall effect is caused by the Lorentz force acting on the conduction electrons in a magnetic field perpendicular to their direction of motion: when the current in a conductor is perpendicular to a magnetic flux density B, the Lorentz force accelerates the charge carriers in the direction perpendicular to B and the current I. A Hall voltage E_H is thus set up that is proportional to B, with a corresponding Hall resistance E_H/I. In ferromagnetic metals, besides this conventional Hall effect, there is a large additional component present that is not proportional to B, but depends only on the presence of a magnetization. This anomalous Hall effect is extremely large for some metals. *Fig. 14* shows the conventional Hall effect in the simple metals silver and gold and the Hall effect in nickel at room temperature as measured by A. Kundt [14], in 1893.

the nature of the impurities responsible for the electrical resistance [17].

In addition to the anomalous Hall effect, for which the theoretical explanation is still incomplete, many other unusual effects related to the conductivity can be expected in ferromagnetic metals. The existence of two independent, magnetically polarized, conduction bands has far-reaching consequences.

We note that in the two-current model the presence of an electric current implies the presence of a moving magnetization or, rather, of two equal moving magneti-

[12] A detailed investigation of the two-current model in nickel and iron was made by A. Fert. See A. Fert, dissertation, Paris (Orsay) 1970, and A. Fert and I. A. Campbell, J. Physique 32, Colloque No. 1, p. 46, 1971.
[13] O. Jaoul, dissertation, Paris (Orsay) 1974.
[14] A. Kundt, Das Hall'sche Phänomen in Eisen, Kobalt und Nickel, Ann. Physik u. Chemie 49, 257-271, 1893.
[15] J. Smit and J. Volger, Phys. Rev. 92, 1576, 1953.
[16] J. Smit, Physica 21, 877, 1955 and 24, 39, 1958.
[17] A useful review is given in: C. M. Hurd, The Hall effect in metals and alloys, Plenum Press, New York 1972.

zations of opposite sense which move with different velocities. Just as the Hall effect occurs not only when a charge carrier moves in a (homogeneous) magnetic field but also when the conductor itself moves in a magnetic field, we would expect that the magnetization that moves with the electric current in a ferromagnetic metal will give rise to new and unusual galvanomagnetic effects. This is a field as yet unexplored.

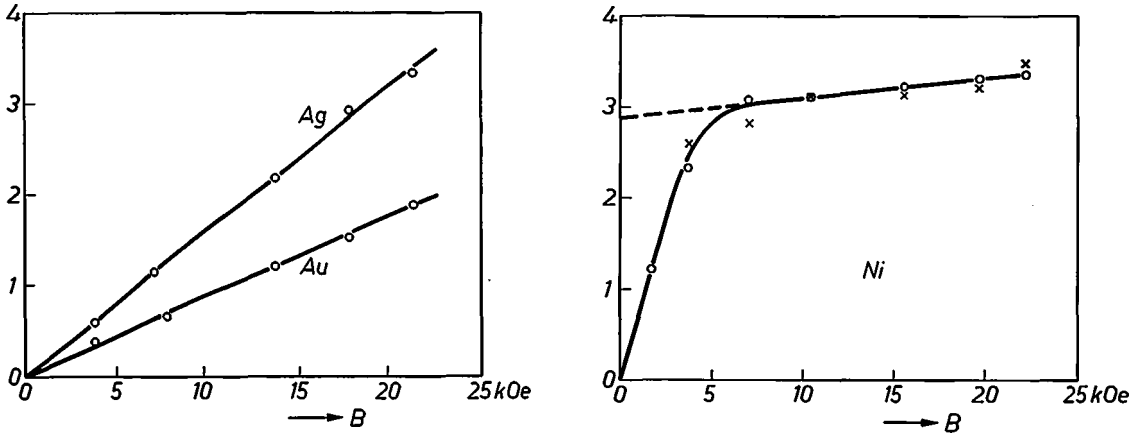


Fig. 14. *Left:* Hall effect in silver and gold, measured in 1893 by A. Kundt [14]. The Hall voltage is plotted (in arbitrary units) against the strength of the applied magnetic field. In these metals only the conventional Hall effect is observed. *Right:* Hall effect in nickel. In addition to the conventional Hall effect we now also observe the anomalous Hall effect. The slope of the upper part of the curve is a measure of the conventional Hall effect; the intercept on the y-axis obtained by extrapolation of this line to $B = 0$ is a measure of the anomalous Hall effect (the corresponding Hall resistivity is about $10^{-7}\ \Omega$ cm).

Summary. Ferromagnetic metals differ in several ways from ordinary metals in their conduction properties. The orientation of the magnetization with respect to the direction of the electric current has an important effect on quantities such as the resistivity. Many experimental measurements have been made on conduction in ferromagnetic metals but the theoretical description still leaves much to be desired. This article first gives a general introduction to metallic conduction, followed by a review of the experimental results on the resistivity and the resistance anisotropy of nickel alloys. In the interpretation of the measurements it is assumed that the conduction electrons can be divided into two independent groups — those with their magnetic moment (i.e. spin) parallel to the ferromagnetic magnetization and those with spin antiparallel to the magnetization. The usefulness of this model is apparent from the analysis of the resistance anisotropy.

Electron resists for the manufacture of integrated circuits

E. D. Roberts

In semiconductor technology the patterns on the wafers are usually produced by an optical method. However, in recent years an alternative method, in which the wafer is irradiated by electrons, has been investigated for cases in which high resolution is desired. This method requires the use of electron resists that behave like the well known photoresists when they are irradiated by electrons. Both positive and negative working electron resists are available to serve in various ways — etch resists, diffusion barriers, dielectrics, etc. It seems likely that in the early stages methacrylate-type resists will be favoured, since they are very versatile, necessitate no changes to other processes used in semiconductor technology, and lend themselves readily to the formation of metal patterns in the elegant 'lift-off' technique. At Mullard Research Laboratories a new methacrylate resist has been developed that is exceptionally suitable for the lift-off technique. Resists such as polymethylcyclosiloxane (PMCS) have also been developed at MRL for the direct application of silicon-dioxide patterns to silicon without etching. The use of this method in the manufacture of transistors and integrated circuits will allow the full resolution of electron-beam techniques to be obtained.

Introduction

One of the procedures used in making monolithic integrated circuits is photolithography. A coating of a photosensitive lacquer (photoresist) is deposited on a wafer of silicon oxidized at its surface, and is exposed to the light of a mercury UV lamp through a photomask. Depending on the type of photoresist (positive or negative) either the exposed or the non-exposed areas are dissolved in a liquid developer. At the places no longer covered by photoresist the layer of SiO_2 can be dissolved in a selective etching bath and the remaining photoresist is then removed, leaving a pattern of SiO_2.

Electron-beam techniques for defining the pattern required in semiconductor technology have been investigated for a number of years. Used as an alternative for the mask and the mercury lamp in photolithographic procedures, they offer the following advantages.

The beams used have high electron energies (5-30 keV) so that diffraction effects are negligible. This allows better definition of details in the range 0.5-2 μm than is possible with photolithographic procedures, and it is expected that the yield of devices with such detail should be substantially better than is obtainable by photolithography. It is also feasible to focus electron beams to very small diameters (3 nm upwards) and the use of finely focused beams can allow patterns to be made with detail much smaller than is possible by photolithography. It has been shown that aluminium conductors only 0.1 μm wide can be made by electron-beam techniques [1]. The ability to make finer detail can lead to an increase in the packing density of the components of integrated circuits with consequent reduction in their size and cost. Smaller components will also reduce power requirements and increase the speed of operation. The ease and precision with which electron beams can be deflected by electric or magnetic fields makes electron-beam techniques relatively amenable to automation. The energy of the electrons is much greater than that of the UV photons, which will not exceed a few electron volts. This broadens the range of materials from which a choice of a resist can be made.

After a brief description of electron-beam equipment now in use and the modes of operation, this article discusses a number of resist systems which are particularly applicable to electron-beam techniques, and explains the mechanisms by which they operate. Several applications of electron-beam techniques will be mentioned. Special attention will be given to the possibility

E. D. Roberts, B.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] M. Hatzakis and A. N. Broers, Record 11th Symp. on Electron, ion, and laser beam technology, Boulder 1971, p. 337.

of producing silica patterns *directly* on the silicon slice without resorting to etching processes, which allows the full resolution of the electron-beam technique to be obtained.

Electron-beam equipment

Two types of electron-beam equipment are being developed, one using scanning and the other projection techniques.

Scanning equipment such as that shown schematically in *fig. 1* has a finely-focused electron beam which is scanned across the resist-coated substrate. The required exposure pattern may be generated by a flying-spot scanner that reads the information from a conventional mask. The electron beam moves in a TV-type raster synchronized with the flying-spot scanner, and is modulated by the output of the latter, being switched on and off at appropriate times during the scanning process [2].

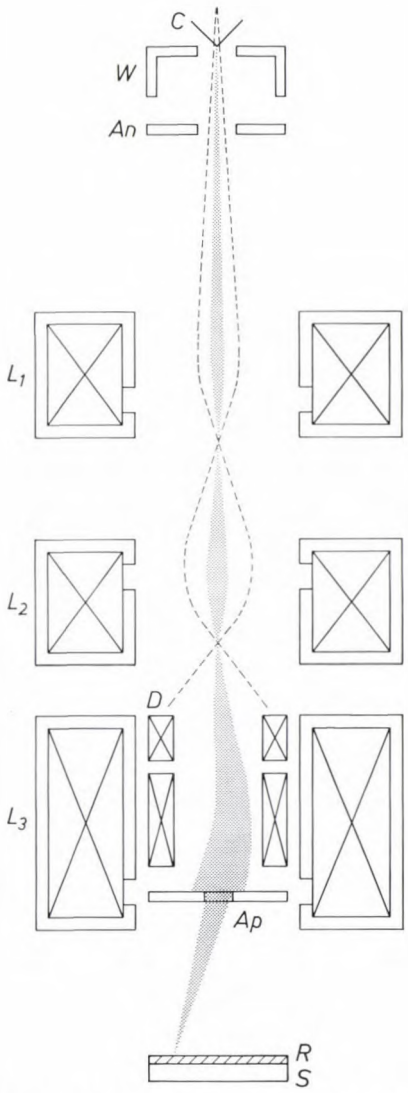


Fig. 1. Schematic diagram of a scanning electron-beam apparatus. *C* cathode. *W* Wehnelt cylinder that modulates the intensity of the electron beam. *An* anode. L_1, L_2 and L_3 magnetic lenses. *D* deflection coils. *Ap* aperture. *S* substrate, coated with resist *R*.

The other method of pattern generation in common use employs a computer which controls the deflection of the electron beam to produce the required pattern[3]. In this case, a TV-type scanning raster is not necessarily used. Both these methods are employed at Mullard Research Laboratories.

In projection electron-beam equipment, the electron beam is always a relatively broad one, of uniform current density and covering the whole of the field it is desired to expose. The pattern is produced by placing a mask somewhere within the beam, the mask having appropriate areas opaque to the electrons. The mask may be in contact with the resist-coated substrate or it may be supported within the beam, the situations being analogous to those in photographic contact printing and projection printing, respectively. In the latter version of the projection method, the final image would usually be reduced from the mask with the aid of a suitable system of electron lenses.

An elegant variation of this technique places the mask in contact with the cathode. This arrangement employs a photocathode formed over a mask bearing the desired pattern in a material opaque to ultraviolet light [4]. It is shown schematically in *fig. 2*. When ultraviolet light shines through the mask electrons are emitted from the cathode in a beam bearing the pattern, and are accelerated towards the resist-coated substrate, serving as an anode. Uniform axial magnetic and electric fields between the cathode and substrate cause the photoelectrons to travel effectively in a straight line from the point of emission to the substrate, exactly reproducing the pattern on the photocathode upon the substrate. A projection system of this type enables the whole area of a silicon slice to be exposed simultaneously to a beam from a repetitively patterned photocathode, the time required for exposure being only a few seconds. This method avoids the sequential exposures with 'step-and-repeat' operations required in scanning methods and in the reducing projection-printing system described above, with the consequent long exposure times for the whole slice. Since the mask only has to be made once, projection with the aid of a photocathode gives a considerable saving in time in mass production.

The nature of electron resists

Resists must be capable of forming uniform pinhole-free films on a substrate by a simple process such as spinning, dip-coating or spraying, and it is usually only polymeric materials which fulfil these requirements.

When energetic electrons interact with organic chemicals, the molecules may either be broken down to smaller fragments, or may link together to form larger molecules. Both types of change occur simultaneously, but when any given material is irradiated one of these processes predominates and determines the net effect upon the material. For polymers it is known that if every carbon atom in the main chain is directly linked to at least one hydrogen atom, larger molecules are generally formed on irradiation, by a process known as cross-linking. If a polymer is cross-linked on irradiation it forms a three-dimensional network and it becomes insoluble and infusible. This type of polymer, which is the most common one, forms the basis of negative-working electron resists, it being possible to dissolve and remove unirradiated material while irradiated material cannot be dissolved

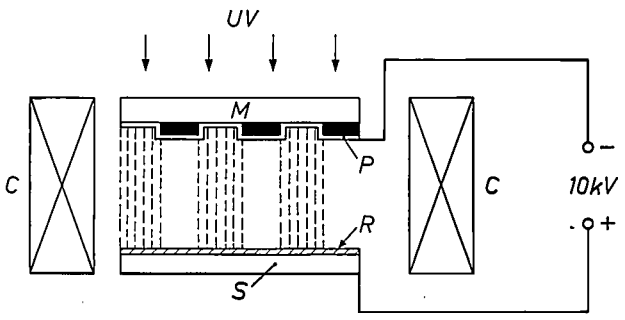


Fig. 2. Schematic diagram of a projection electron-beam apparatus using a photocathode. Ultraviolet light passes through a mask *M* with appropriate opaque areas on the photocathode *P*. Magnetic coils *C* produce an image of the pattern on the substrate *S*, which acts as an anode and is coated with electron resist *R*.

away. Thus a residual film pattern remains on irradiated areas after development, and the situation is analogous to the formation of a silver image in a photographic 'negative'.

A relatively small number of polymers is known in which irradiation in moderate doses predominantly causes a break in the main polymer chain, resulting in irradiated material having a lower average molecular weight than unirradiated material. By judicious choice of solvent or mixtures of solvents, it is possible to dissolve the irradiated polymer selectively, which allows positive-working electron resists to be formulated. At higher electron exposures, however, cross-linking predominates again and the irradiated material becomes completely insoluble. Thus, there is always a limited range of electron exposures in which positive action may be observed.

Another consideration of importance in the case of positive-working resists is the glass-transition temperature of the polymer, i.e. the temperature above which the polymer changes from a glass-like structure to a rubbery state. If the glass-transition temperature is low, the polymer can be easily deformed and the resolution of fine patterns developed in a film of such a polymer may be easily impaired. Good positive electron resists should therefore have a glass-transition temperature above the highest temperature to which the resist will be subjected after development of the irradiated pattern in it. It should be noted that the requirement of a high glass-transition temperature does not necessarily apply to negative resists. The resist material in this case is cross-linked during irradiation, which automatically increases its glass-transition temperature. Indeed, even liquids can be used as negative electron resists, though there is sometimes difficulty in applying films of uniform thickness of liquid initially.

Almost all polymers which give positive action upon irradiation consist of chains in which some of the carbon atoms are not directly linked to hydrogen (e.g. polymethyl methacrylate, poly-α-methylstyrene, polyisobutylene, polyvinylidene chloride, polytetrafluoroethylene, etc.). Many of these are unsuitable because their glass-transition temperatures are too low, or because they are not easily soluble (e.g. polytetrafluoroethylene), which makes application of the film and development of irradiation patterns difficult, and the only material which has yet found wide use is polymethyl methacrylate [5].

Positive electron resists

Polymethyl methacrylate

Polymethyl methacrylate (PMMA) is readily available commercially. It is a long-chain polymer, the main chains containing alternate quaternary carbon atoms

[2] J. M. S. Schofield, H. N. G. King and R. A. Ford, 3rd Int. Conf. on Electron and ion beam science and technology, Boston 1968, p. 561.

[3] J. P. Beasley, 4th Int. Conf. on Electron and ion beam science and technology, Los Angeles 1970, p. 515.

[4] T. W. O'Keeffe, J. Vine and R. M. Handy, Solid-State Electronics **12**, 841, 1969. Further developments at Mullard Research Laboratories are described by J. P. Scott, in 6th Int. Conf. on Electron and ion beam science and technology, San Francisco, May 1974, to be published.

[5] See for instance:
I. Haller, M. Hatzakis and R. Srinivasan, IBM J. Res. Devel. **12**, 251, 1968;
H. Y. Ku and L. C. Scala, J. Electrochem. Soc. **116**, 980, 1969;
E. D. Wolf, L. O. Bauer, R. W. Bower, H. L. Garvin and C. R. Buckey, IEEE Trans. ED-**17**, 446, 1970.
Several other positive-resist systems, including the polysulphones, have been investigated in various establishments. See M. J. Bowden and L. F. Thompson, 2nd Symp. on Scanning electron microscopy of polymers and coatings, Dallas, April 1973, p. 99.

(i.e. connected directly to four other carbon atoms), and upon irradiation these chains are broken.

It has been shown theoretically by H. Y. Ku and L. C. Scala [5] that, using a given developing solvent (or mixture of a solvent and a non-solvent), and provided the molecular weight of the resist material (e.g. PMMA) is sufficiently high, the exposure required to produce a positive image is independent of the molecular weight of the resist material. Also, the molecular weight of the irradiation-degraded material is almost independent of the original molecular weight. In the development process it is necessary to remove irradiation-degraded polymer without affecting the unirradiated material. Clearly, the greatest chance of accomplishing this successfully will occur if the original material has the highest possible molecular weight.

Most investigators working with PMMA have used developers containing mixtures of a good solvent for PMMA with a non-solvent. Given a liquid which is a good solvent for a particular polymer of any molecular weight, and a liquid which is not a solvent for the same polymer irrespective of its molecular weight, it is, in general, possible to find a composition of a mixture of the solvent and the non-solvent that will act as a solvent for the polymer below a critical molecular weight, but will act as a non-solvent for the same polymer above the critical molecular weight.

A typical combination of solvent and non-solvent for PMMA is methyl isobutyl ketone with isopropyl alcohol mixed in the volume ratio 1 : 3. We have found, surprisingly, that a satisfactory developer can also be made (and less expensively) by adding a small quantity of the non-solvent water to the other non-solvent isopropyl alcohol.

The minimum electron exposure required to form a developable image depends upon the developer composition. *Fig. 3* shows the minimum exposures required for developers of two compositions. We have found that for a number of commercial PMMA homopolymers and copolymers [6] with 10% isobutylmethacrylate, covering a range of molecular weights, the exposure curves are almost identical, confirming the theoretical predictions of Ku and Scala [5].

When a PMMA film is irradiated by an electron beam, some of the electrons are scattered within the resist, and some are also back-scattered from the substrate. All these scattered electrons can also break down the resists, and the net effect is that the exposed pattern may be slightly larger than the irradiated area. This does not usually matter unless it is necessary to produce details that are comparable in size to the beam cross-section. A more important consequence of the effect of scattered electrons is that the edge profiles of developed patterns in positive resists exhibit some degree of undercutting. This has been demonstrated in a number of ways by various investigators, and has been used to advantage in a very elegant technique for producing fine patterns of metal. The method is known as the 'lift-off' metallization technique and will be described below.

Uses of PMMA resists

PMMA resist forms a good etch mask, resistant to many etchants. We have used it, after defining patterns in it by electron exposure, to etch patterns in thermal oxide on silicon with buffered HF, to etch patterns in chromium on glass [3], as a mask in ion-etching gold films, and even as an ion-implantation barrier against phosphorus ions accelerated to 100 keV.

An example of lines etched in thermal oxide using a PMMA mask is shown in *fig. 4*. Thus, diffusion-barrier and passivating-layer patterns could be prepared, allowing semiconductor devices (e.g. transistors, integrated circuits) to be made by the same methods as are used in conventional processing. PMMA is not photosensitive, so safelight conditions are not required, and the other advantages offered by electron-beam techniques are available without changes to other established processes.

Perhaps the most interesting use of PMMA is in the lift-off metallization technique [7], in which the unwanted metal is lifted off together with the resist. The sequence of steps is shown in *fig. 5*. The success of this method depends upon achieving and maintaining the undercut profile of all edges in the developed pattern, which allows a clear separation between the metal deposited on the substrate and that over the remaining resist. The applied resist film must, of course, be rather thicker than the metal film it is desired to apply, and great care must be taken to ensure that development of

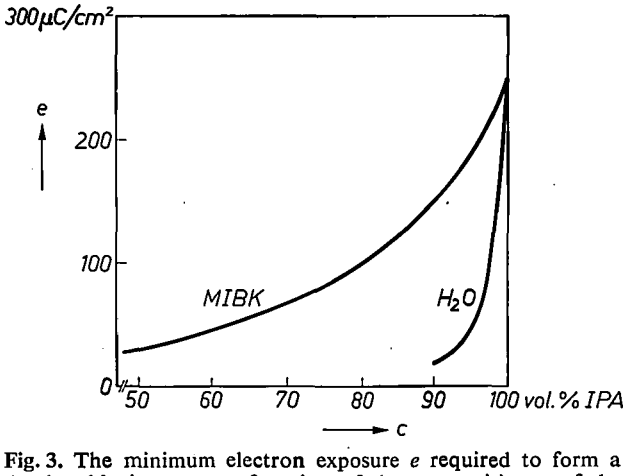


Fig. 3. The minimum electron exposure e required to form a developable image as a function of the composition c of the developer, which in one case consists of isopropyl alcohol and water, and in the other of isopropyl alcohol and methyl isobutyl ketone.

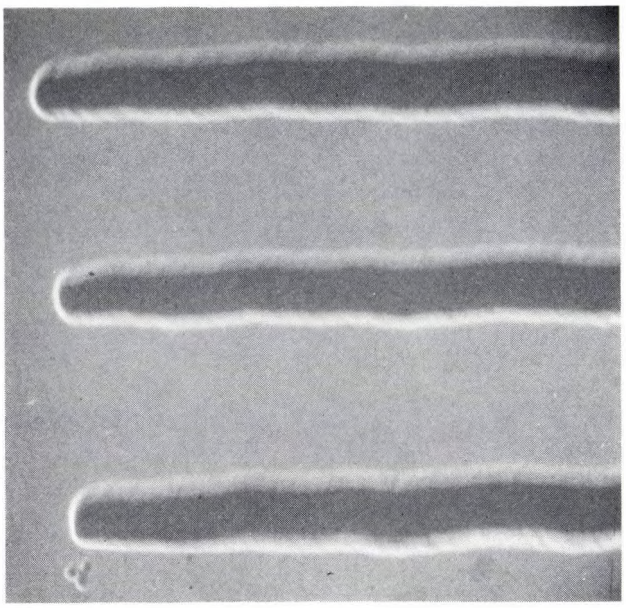


Fig. 4. Lines etched in a thermal-oxide film, the pattern being defined by a scanning electron-beam equipment in PMMA resist. Magnification 10 000 ×. (Because the object was positioned obliquely in all the photographs reproduced in this article, the magnifications given are valid only for one direction, in this case the vertical direction.)

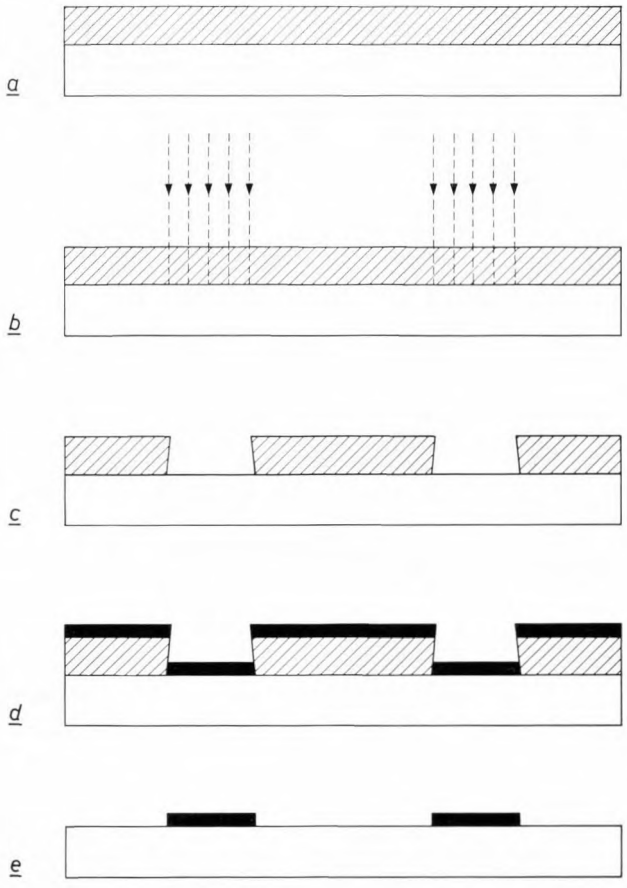


Fig. 5. Stages in the lift-off metallization technique. A film of PMMA or another positive electron resist is applied to a substrate (*a*), subsequently irradiated by an electron beam (*b*) and then developed (*c*). After metal is evaporated over the whole surface (*d*), the resist and metal deposited on it are removed (*e*).

the pattern is complete, and that the developed pattern is washed completely free of irradiated resist. Unless this is done a thin film of polymer may remain under the metal deposited upon the substrate so that this lifts off together with the unwanted metal at the final stage. With very fine patterns, washing assisted by ultrasonic agitation is often advantageous.

Modified methacrylate resists

PMMA is a thermoplastic material and patterns defined in it can be easily deformed if the film becomes heated after development. If it does get hot, as for example during evaporation of a metal, it will flow, spoiling the undercut edge profile. No break then occurs in the metal film at the edges of the patterns, and it is impossible to apply the lift-off technique.

The minimum electron exposure of PMMA is limited to about 50 $\mu C/cm^2$ by the need to produce sufficient decrease in average molecular weight during exposure to allow the developing solvent to discriminate between irradiated and unirradiated areas. At low electron exposures a more active developer must be used, but then the unirradiated polymer, which it is desired to retain upon the substrate, dissolves appreciably in the developer, and the undercut edge is spoiled. The lift-off technique cannot therefore be used unless very thin metal films are to be applied.

We have developed a modified methacrylate-type resist which can overcome these disadvantages [8]. It comprises methacrylate chains which are cross-linked in the unirradiated film by anhydride bridges. The mechanism by which this resist operates is different from that encountered in PMMA. Electron irradiation breaks the cross-links, restoring the straight-chain structure, and although some breaks in the main chain may occur this is not crucial to its operation. The resist film, being cross-linked, is insoluble in solvents, but becomes soluble when its cross-links are broken upon irradiation. Thus, a more active solvent may be used to develop patterns than is permissible with PMMA, and the development conditions (time, temperature, etc.) are less critical since it is only necessary to distinguish between soluble and insoluble regions, rather than between more-soluble and less-soluble regions, as when PMMA is used. The cross-links also confer more rigidity upon the film, making it relatively infusible, so that it has greater thermal stability (i.e. cross-links raise its glass-transition temperature) and undercut edge profiles are not deformed if it becomes heated.

[6] A homopolymer is built up from identical monomer units, a copolymer from different ones.

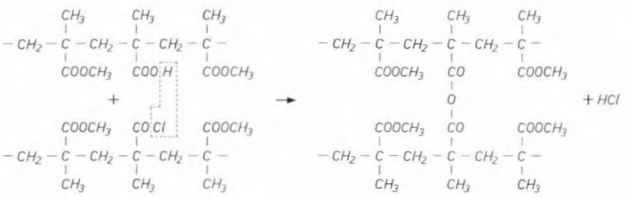
[7] M. Hatzakis, J. Electrochem. Soc. **116**, 1033, 1969.

[8] E. D. Roberts, 2nd Symp. on Scanning electron microscopy of polymers and coatings, Dallas, April 1973, p. 87.

Table I. Properties of cross-linked methacrylate films.

Fraction of cross-linking groups in copolymers (mol %)	Heat treatment to effect cross-linking (min, °C in dry N_2)	Minimum exposure at 1 min development in methyl isobutyl ketone ($\mu C/cm^2$)	% of unexposed film dissolved by 1 min immersion in	
			acetone	methyl isobutyl ketone
0 (standard PMMA)	(15, 200)	—	90	6
1	15, 200	15	20	0
3	30, 115	—	85	7
	15, 130	12	45	0
	15, 150	20	9	0
	15, 200	20	0	0
	15, 200 (heated in air)	20	10	1
10	30, 115	6-8	25	0
	15, 130	20	2.5	0
	15, 150	35	0	0
	15, 200	35	0	0
	15, 200 (heated in air)	35	0	0

As cross-linked films are insoluble, they cannot be applied directly, but the cross-links must be formed in the film after it has been deposited from solution upon the substrate. A convenient method of doing this is to prepare two copolymer solutions, one containing methyl methacrylate (CH_2:$C(CH_3)COOCH_3$) and methacrylic acid (CH_2:$C(CH_3)COOH$), the other containing methyl methacrylate and methacryloyl chloride (CH_2:$C(CH_3)COCl$) as co-monomers. The solutions are mixed in appropriate quantities (the mixture containing approximately equimolecular quantities of carboxyl and acid-chloride groups) and the films are spun from the mixed polymer solution. Upon heating, anhydride cross-links are formed by the following reaction:



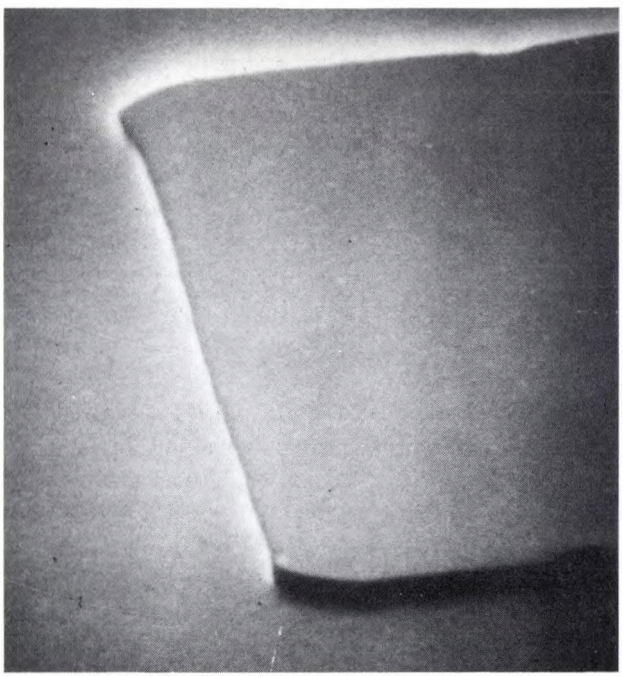


Fig. 6. Scanning-electron micrograph of a cross-linked methacrylate pattern (10% cross-linkable polymers). The resist is heated for 15 minutes at 115 °C, exposed at 10 $\mu C/cm^2$ and developed for 1 minute with methyl isobutyl ketone. Magnification 10 000×. The image shows that the new resist gives sharply defined edges.

The extent of cross-linking may be controlled by the relative number of reactive groups in the copolymers and by the heat treatment to which the film is subjected to effect the cross-linking. *Table I* shows the effect of these variables upon the solubility of the unexposed films (columns 4 and 5), and the minimum electron exposure required to produce a positive image developable by methyl isobutyl ketone is shown in column 3. The Table indicates that in all cases the film has become completely insoluble after heating at a temperature between 150 and 200 °C for 15 minutes. It is desirable if the copolymers contain only a small proportion of cross-linkable groups that the heating should be carried out in a dry atmosphere. The results for 3% cross-linkable polymers heated in an undried atmosphere show that complete insolubility is not achieved. This is probably because acid-chloride groups near the surface of the film are hydrolysed by the available moisture, thus losing their ability to form anhydride cross-links and producing a surface layer of acid copolymer which is still soluble in the developer.

Fig. 6 shows a pattern made by exposing a cross-linked resist film to a broad electron beam, the pattern being produced by a fine copper grid placed in contact with the resist to act as a shadow mask. *Fig. 7* is a pattern of fine lines produced by scanning a finely-focused beam across the resist film. The undercut edges in the developed patterns can be clearly seen. Fully cross-linked films are unaffected even by immersion in boiling acetone. *Fig. 8* shows the superior thermal stability of patterns in cross-linked methacrylate.

Uses of cross-linked methacrylate films

The films with cross-linked copolymer chains may be used for the same purposes for which unmodified PMMA is used, i.e. as an etch mask, or in the lift-off technique, most benefit being gained in the latter process. *Fig. 9* shows metal patterns produced by the lift-off technique applied to cross-linked methacrylate. The cross-linking makes the lift-off stage a little more difficult as the cross-links must be destroyed to remove the polymer. This may be accomplished, however, by immersion in fuming nitric acid or in a mixture of ammonia solution and acetone. While fuming nitric acid destroys the polymer completely, the

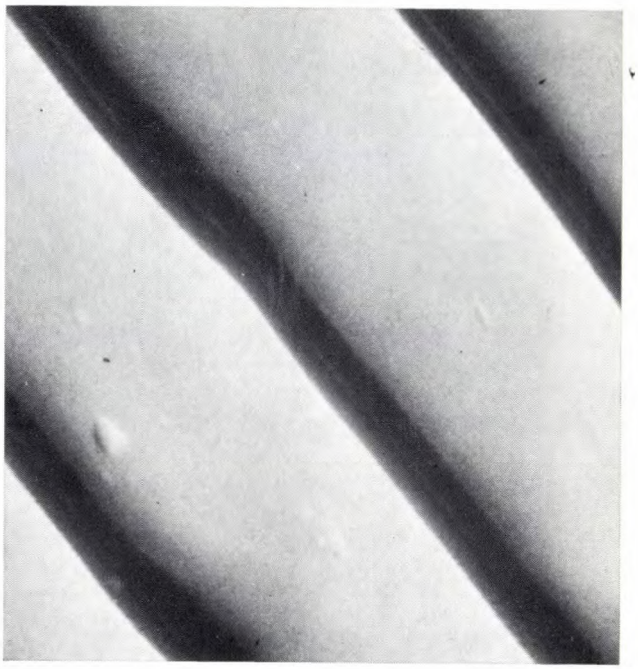
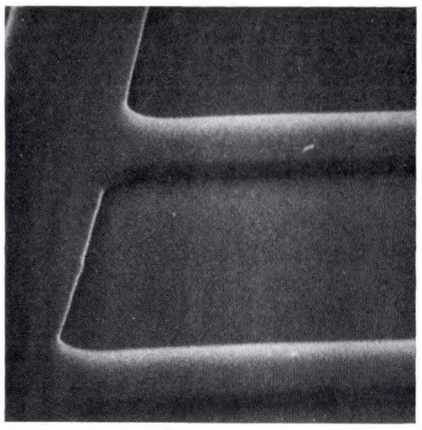
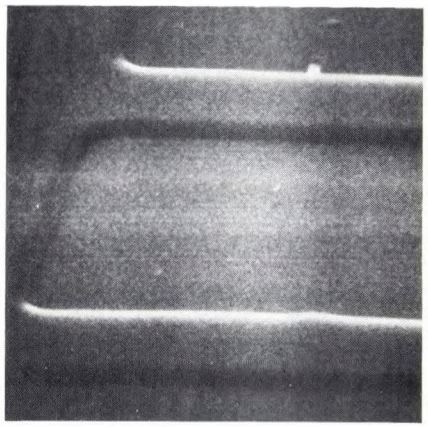


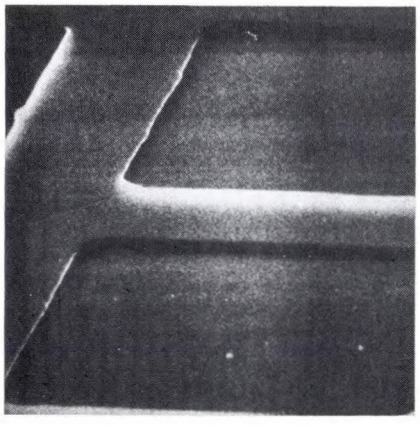
Fig. 7. Scanning-electron micrograph of narrow grooves (dark) in a film of cross-linked methacrylate as in fig. 6. The resist is heated for 15 minutes at 200 °C, exposed at 40 $\mu C/cm^2$, and developed for 2 minutes in methyl isobutyl ketone. It can be seen that the edges are undercut. Magnification 12 000×.



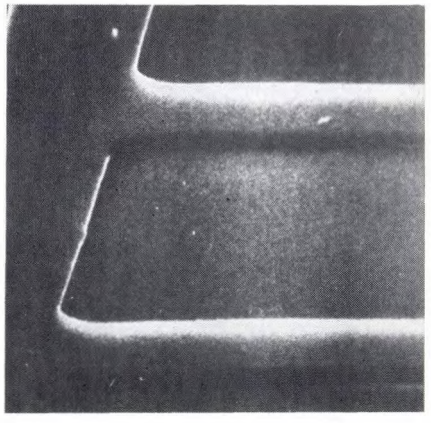
a



b



c



d

Fig. 8. PMMA grid patterns, obtained with standard (upper pictures) and cross-linked (10%) material, *before* (*a* and *c*) and *after* heating (*b* and *d*) for 15 minutes at a temperature of 130 °C. Photograph *d* demonstrates the very good thermal stability of patterns in the new resist. The preparation of the standard PMMA pattern consisted in heating for 15 minutes at 200 °C, exposure at 50 $\mu C/cm^2$ and development for 1 minute with isopropyl alcohol diluted with 5% of water. The cross-linked methacrylate pattern was made by heating for 15 minutes at 200 °C, exposing at 40 $\mu C/cm^2$ and developing for 1 minute with methyl isobutyl ketone. Magnification 3000×.

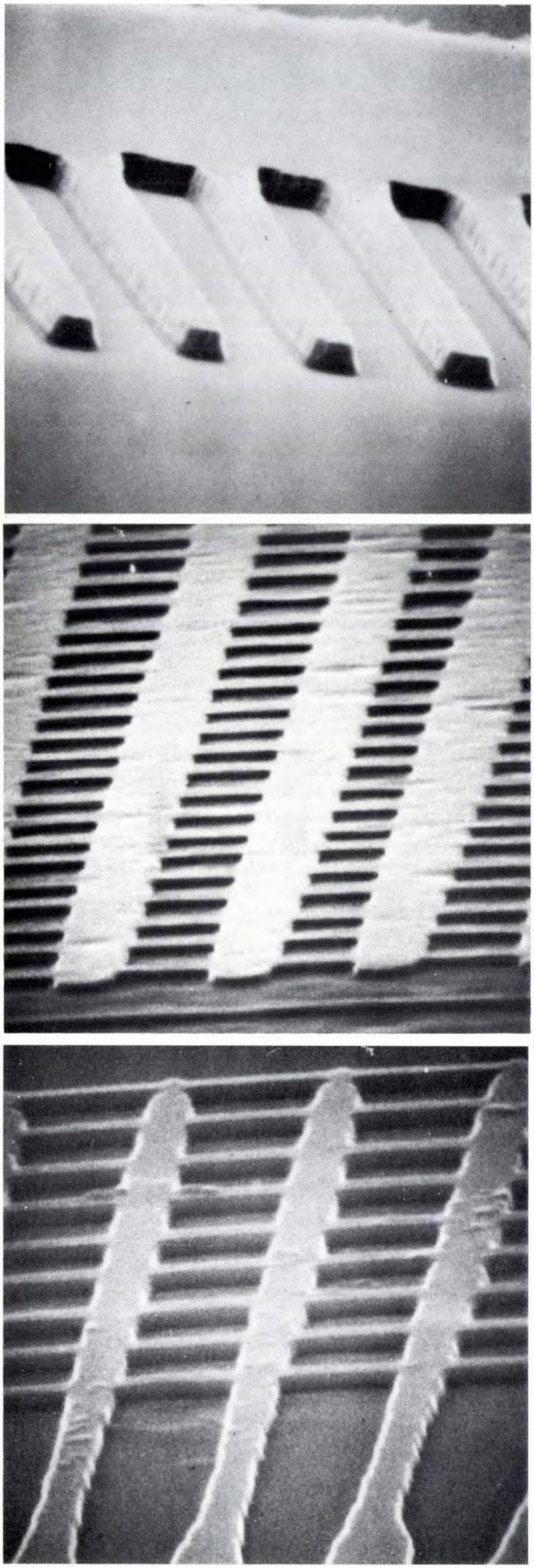
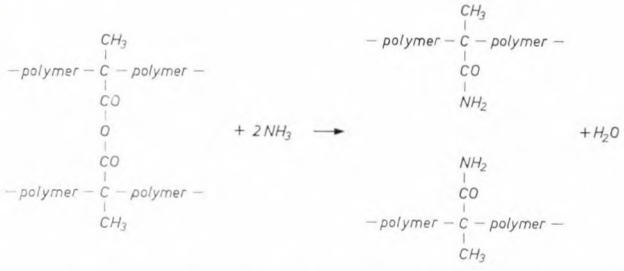


Fig. 9. Aluminium patterns on silicon produced by the lift-off technique with cross-linked methacrylate (3% cross-linkable polymers) resist (*a*), gold patterns on silicon (*b*) and gold patterns on oxidized silicon (*c*). Magnifications: 6750×, 8500× and 8500×, respectively.

acetone/ammonia mixture only breaks the cross-links, thus:



The separated chains then dissolve in the acetone just as PMMA does. We developed the acetone/ammonia mixture for metals that are attacked by fuming nitric acid.

Negative electron resists

In principle, many materials are available for use as negative resists, for most materials are cross-linked upon irradiation by electrons. The choice may be limited, however, by the sensitivity required, solubility of the resist, etc. [9].

In our work on negative resists we have concentrated upon siliceous (silicon-containing) materials. These enable siliceous film patterns to be produced directly by the beam, and it is possible by subsequent treatment to convert these films to a material having properties identical with thermal silica.

Polymethylcyclosiloxane (PMCS)

G. H. Wagner *et al.* have shown that polymers resembling silica in structure could be prepared by hydrolysis of certain organotrichlorosilanes [10]. In order to avoid excessive formation of three-dimensionally cross-linked polymers at this stage, the hydrolysis was carried out in a mixture of diethyl ether and ice. We have found that this same procedure can be applied to the hydrolysis of methyltrichlorosilane [11]. Since this molecule contains the smallest possible alkyl group (methyl), the polymer derived from it will have a structure as near as possible to that of silica. The presence of some alkyl groups is necessary to ensure that the polymer has the solubility required in its use as a resist. Hydrolysis of trichlorosilane or silicon tetrachloride, for example, would give a carbon-free polymer like silica directly, but it would not have the required solubility. In any case, the methyl group introduced by methyltrichlorosilane can be removed afterwards effectively.

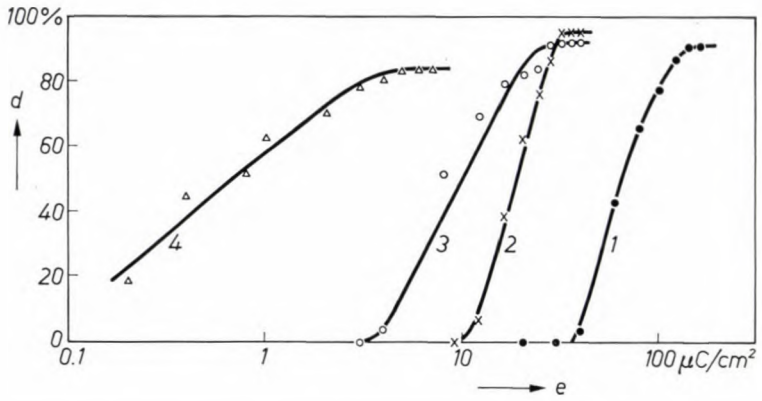
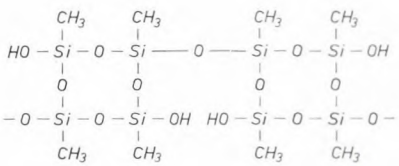


Fig. 10. The thickness d, as a percentage of the initially applied thickness, produced as a function of the exposure e, using standard PMCS resist (curve *1*), high-molecular-weight PMCS (*2*), poly(methyl/vinyl)cyclocylosiloxane copolymer (*3*) and polyvinylcyclosiloxane (*4*).

The only drawback in using methyltrichlorosilane instead of the higher homologues is that the loss during preparation by cross-linked-polymer formation is greater. Yields are only about 50-60 % of the theoretical yield. The polymer formed in this process is believed to consist of chains of tetrasiloxane rings joined through siloxane bridges, thus:



It has an average molecular weight of about 600, contains about 40 % w/w silicon and about 10-12 % w/w hydroxyl as silanol groups (OH groups linked to silicon). Pairs of the silanol groups become cross-linked on irradiation, producing the desired insolubility.

Upon irradiation, PMCS becomes progressively more cross-linked and insoluble. Curve *1* of *fig. 10* shows the thickness of cross-linked film remaining after various electron exposures. It can be seen that the maximum thickness of film remains at exposures above 140 $\mu C/cm^2$, presumably because the film is then cross-linked sufficiently to be completely insoluble in the developing solvent. The slight shrinkage in fully exposed films occurs during the irradiation stage.

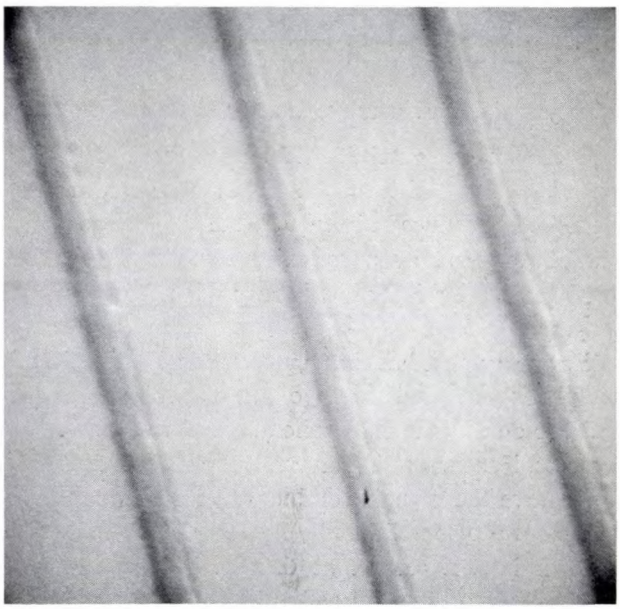


Fig. 11. Fine lines of siliceous film produced by electron exposure of PMCS resist. Magnification 9000×.

Greater sensitivity can be achieved either by increasing the molecular weight of the material, though there is a limit to this process, or by introducing vinyl groups into the molecules. Exposure curves for these modifications are also given in fig. 10 (curves *2-4*). *Fig. 11* gives an electron micrograph of some fine lines produced in PMCS resist.

Examination of infrared spectra of PMCS films at various stages during electron irradiation has shown that the main mechanism by which cross-linking occurs is the formation of siloxane bridges by the elimination of water between two silanol groups. There is also some reaction between methyl groups, probably to form methylene and ethylidene bridges, thus:

though this mechanism appears to occur more slowly

[9] Among negative resists, epoxidized polyolefines and polyepoxybutadiene in particular have been examined extensively for several years. See T. Hirai, Y. Hatano and S. Nonogaki, J. Electrochem. Soc. **118**, 669, 1971. Although these resists appear to be very sensitive, it has also been reported that their contrast ratio is rather low, which can impair the resolution of very fine patterns. See E. D. Feit, R. D. Heidenreich and L. F. Thompson, 2nd Symp. on Scanning electron microscopy of polymers and coatings, Dallas, April 1973, p. 125.
High-molecular-weight silicone oils have been used to produce siliceous dielectric films. See H. Aoe, Y. Yatsui and T. Hayashida, Microelectronics and Reliability **9**, 267, 1970.

[10] G. H. Wagner, D. L. Bailey, A. N. Pines, M. L. Dunham and D. B. McIntire, Ind. Engng. Chem. **45**, 367, 1953.

[11] For more detailed information see E. D. Roberts, 3rd Int. Conf. on Electron and ion beam science and technology, Boston 1968, p. 571, and E. D. Roberts, J. Electrochem. Soc. **120**, 1716, 1973.

than the silanol condensation, and requires much higher exposures before it takes place extensively.

As noted earlier, the exposure curve shows that there is slight shrinkage of the film during exposure and development. It still contains many organic groups, but these can be almost completely eliminated by heating the developed pattern in wet oxygen at 650 °C. The infrared spectrum of the film is then almost identical with that of thermally grown silica films. A subsequent densification process at 1050 °C converts it to a form having the same refractive index as thermal silica and the same etch rate in hydrofluoric acid. During these heating processes, further shrinkage takes place, the thickness of the final silica film being about two-thirds that of the PMCS film initially applied [11].

Use of PMCS in device technology

Some experimental *N-P-N* silicon transistors have been made in which all the diffusion masks required have been produced directly on the silicon slice using PMCS exposed by a scanning electron beam. The final passivating layer has been produced in the same way.

For use as a diffusion barrier, the PMCS film is simply exposed and developed, though it is often beneficial to densify the film by heating before the diffusion process. For passivating layers, however, it is necessary to remove the organic residues. As mentioned above, this can be accomplished by heating the siliceous film pattern in wet oxygen at 650 °C. The film is subsequently densified by heating in nitrogen for 30 minutes at 900 °C.

When siliceous films are made in this way, no etching step is required to define the pattern, in contrast to conventional processing or to electron-beam processing with PMMA. It is desirable to remove the silica film from the slice at each stage, because any windows in the film would be likely to become covered with oxide. Windows for the next stage of diffusion would normally be formed in this oxide by an etching process. However, we did not in fact use etching for this since undercutting can reduce the resolution. To keep the resolution as high as possible, the complete oxide film is removed and a fresh mask prepared in a fresh layer of PMCS resist.

Transistors have been made having emitters either 50 μm wide or 5 μm wide. Photomicrographs of the two types are shown in *fig. 12*. The device with the 5-μm emitter is shown before the metallization stage has been carried out, allowing the PMCS passivating layer to be seen.

Devices with finer detail have not yet been made, as the purpose of these experiments was only to test the possibility of making devices by electron-beam techniques using the new material PMCS instead of conventional thermal oxide. The transistors had gain in the range 10-100 and were comparable in this respect to conventional transistors having similar dimensions. Their leakage currents, however, were in general much higher than those normally encountered, ranging between about 1 nA and 1 mA. The transistors with a 5-μm emitter were generally worse than those with a 50-μm emitter, though some of them had leakage currents less than 100 nA. It is not yet known whether the high leakage is due to pinholes in the PMCS films or to some other shortcoming in it, or if misalignment of succeeding masks due to difficulties in the electron-beam technique used at the time contributed to this. The fact that some good transistors were obtained

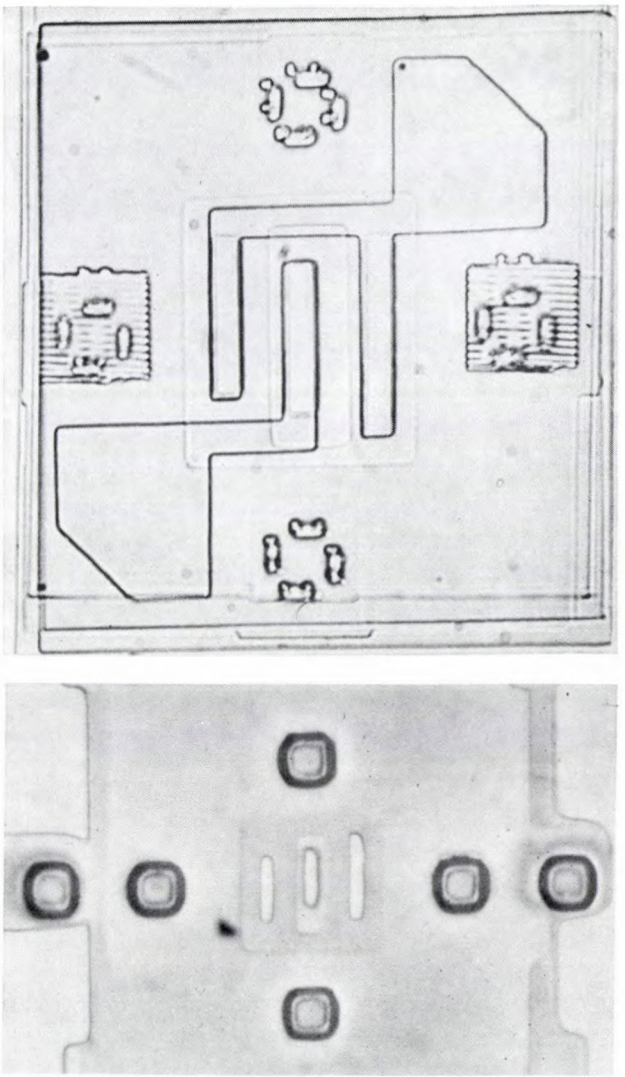


Fig. 12. Experimental *N-P-N* transistors made by electron exposure of PMCS resist, with emitter widths of 50 μm (*a*) and 5 μm (*b*). In (*a*) the central metal strip is the emitter connection, and the forked strip the base connection. The four surrounding patterns are location marks. The photograph (*b*) was made before metallization, and the PMCS passivating layer can be seen. Magnifications are 200× in (*a*), 675× in (*b*).

does suggest, however, that the proposed process is feasible, though further work is still required to perfect it.

One difference between the transistors produced by electron-beam exposure of PMCS films and those produced conventionally is that the passivating layer produced from PMCS has no phosphosilicate-glass layer. Conventional transistors normally do have such a layer, and it is known that if this phosphosilicate glass is not present, they too may show high leakage currents. Phosphosilicate glass is usually present as a result of conventional processing, though in certain devices a special process step may be required for its application. It is possible that as is the case with thermal silica, the properties of passivating layers prepared from PMCS would be improved if phosphosilicate glass were applied over it, and this may reduce the leakage current to values normally encountered. With this aim in view we have developed resists which allow the formation of phosphosilicate-glass patterns by electron-beam techniques [12].

the required P : Si ratio. It was necessary to match the electron sensitivities of the components so that both could be rendered insoluble by the same exposure to maintain the required P : Si ratio.

We have produced two mixtures suitable for use as electron resists: PMCS with tris-(dimethylsilylene)-

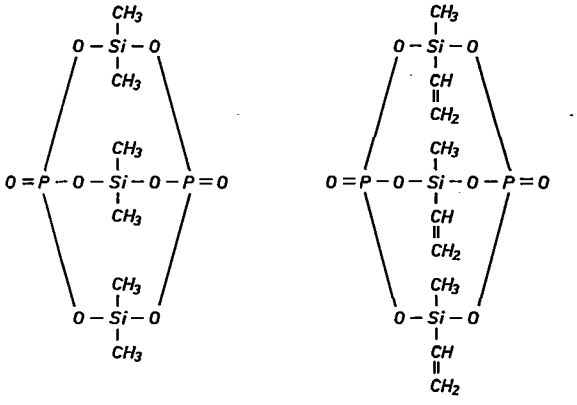


Fig. 13. Formulae of tris-(dimethylsilylene)diphosphate *(left)* and tris-(methylvinylsilylene)diphosphate.

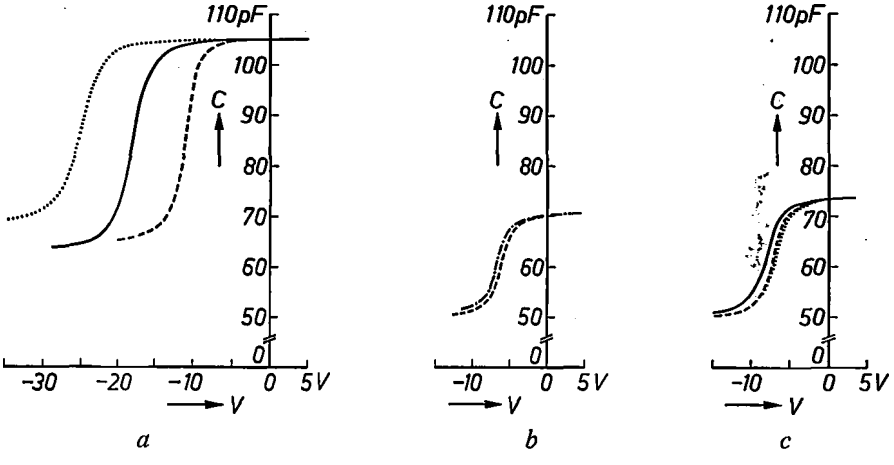


Fig. 14. The effect of electron-beam-defined phosphosilicate-glass coatings on the variation of capacitance of MOS capacitors with d.c. bias voltage. The capacitors all have PMCS dielectric layers. (*a*) No additional coating. (*b*) Coated with PMCS with 5% by weight of a silylene diphosphate with methyl substituents. (*c*) Coated with PMCS with 10% by weight of a silylene diphosphate with methyl and vinyl substituents. The exposures were made at 250, 800 and 5 $\mu C/cm^2$ respectively, and the capacitors were heated in wet oxygen at 650 °C for 15 minutes, and finally in dry nitrogen at 1050 °C for another 15 minutes. Solid lines: initial curve, dotted lines: after heating for 30 minutes at 180 °C with an applied voltage corresponding to +22.5 V/μm; dashed lines: after the same heat treatment at a field-strength of −22.5 V/μm.

Phosphorus-containing siliceous resists

The phosphosilicate glasses used in conventional processing cover a range of compositions, with the molecular ratio $P_2O_2 : SiO_2$ up to about 16 mol%. Known organic compounds containing both silicon and phosphorus and likely to be capable of being cross-linked by an electron beam (and so rendered developable) all contain much higher ratios of phosphorus to silicon than this, so the compounds were mixed with polysiloxanes (such as PMCS) to produce diphosphate, and polyvinylcyclosiloxane with tris-(methylvinylsilylene)diphosphate (*fig. 13*) (polyvinylcyclosiloxane is the vinyl analogue of PMCS). The first of these systems requires an electron exposure of about 800 $\mu C/cm^2$, while the second requires an exposure within the range 2-10 $\mu C/cm^2$, to produce developable images of the correct composition.

The stabilizing effect of the phosphorus-containing

[12] E. D. Roberts, 5th Int. Conf. on Electron and ion beam science and technology, Houston 1972, p. 102.

films on electron-beam-defined PMCS dielectric films has been demonstrated in MOS capacitors. Both the dielectric-oxide film and the protective coating were made using electron-beam techniques, starting from PMCS and the resists just mentioned, respectively. For both a series of heat treatments was necessary to eliminate organic residues and to densify the film. The effect of the coating film can be seen from the curves in *fig. 14*, which shows the relationship between the capacitance of the MOS capacitors and the applied voltage. Without a protective coating positive ions such as sodium can move within the passivating layer when a voltage is applied at high temperature. This causes a displacement of the curve along the voltage axis. When a protective coating is present the positive ions are trapped in the phosphosilicate glass and the curves remain in the same position. This effect of phosphosilicate glass is known in semiconductor technology, and is used for stabilizing thermally grown silica. Our electron-beam-defined phosphosilicate glass can also stabilize thermally grown silica dielectrics in MOS capacitors.

These resists for producing phosphosilicate glass have not yet been used on transistors made by electron-beam techniques, but investigations along these lines are already under way and we are convinced that this method will prove to be a fruitful one in the semiconductor technology of the future.

Summary. Electron-beam techniques for defining the patterns required in semiconductor technology offer the advantages, compared to the current optical methods, of a better definition, smaller details (0.1 μm) and greater amenability to automation. The already commercially available positive electron resist PMMA is discussed first. A number of its drawbacks, e.g. its thermoplasticity, are overcome in a modified methacrylate-type resist we have developed, in which cross-links are introduced in the unirradiated film by anhydride bridges. This makes the resist pre-eminently suited for the application of the 'lift-off' metallization technique. In our work on negative electron resists the emphasis was mainly on siliceous materials, which enable silica-like patterns to be produced directly by the electron beam, thus eliminating the need for etching steps. With a polymer (PMCS) prepared from methyltrichlorosilane we have made the diffusion masks and the final passivating layer of some experimental *N-P-N* transistors. As the leakage currents were still high, we have developed resists which also allow the direct formation of phosphosilicate-glass patterns. The stabilizing effect of these phosphorus-containing films has been shown by their influence on capacitance-voltage characteristics of MOS capacitors with electron-beam-defined PMCS dielectric layers.

Holographic strain analysis

C. H. F. Velzel

Holography has attracted wide general interest because of its ability to create 'three-dimensional images' which are so 'real' that they can be viewed from various directions just like the actual object; holography differs from the stereoscopic viewing of two images, by giving a true parallax. Holography is based on the fact that a light wave scattered by an object can be completely recorded in a hologram (e.g. on a photographic plate) and exactly re-created later. As a result, it is possible to allow a light wave to interfere with a second light wave that is produced at some other time. This leads to one of the most important applications of holography, the interferometric measurement of small displacements or strains. Measurement of displacements in the direction of the observer is the most obvious application. Displacements perpendicular to this direction can, however, also be measured. One method of doing this has been devised by the author. Both types of measurement are discussed here, preceded by an introduction to holography and followed by some examples.

When a smooth surface is covered by a thin transparent film not quite uniform in thickness, a pattern of interference fringes can be seen if the surface is illuminated by monochromatic light. The variations in thickness of the film can be read off directly from the pattern. The fringe pattern is produced by the interference of light rays reflected at the object-film interface and at the outer surface of the film.

An object that has been given a small deformation (e.g. as a result of heating or mechanical stress) is in a certain sense analogous to an object coated with a film, as discussed above. Here again we have two surfaces close together but now separated in time. There is now, however, no question of interference; the light rays reflected from the deformed and the undeformed surface are not simultaneously present.

To measure deformations by means of conventional interferometry, the incident light is split into two coherent beams. One is reflected by the deformed object whilst the other is reflected by a replica of the undeformed object or other optical reference surface (e.g. a flat); interference takes place when the two beams are brought together again. Both the reference surface and its positioning must be exact to within a fraction of a wavelength. This method is practicable only when the deformed object has a smooth surface.

Holography presents an elegant solution to this problem. In holography a light beam can be 'stored' in a hologram and regenerated later. In this way interference is possible between non-simultaneous light beams, so that deformations of an object can be measured in the same way as film thickness. The hologram provides in effect an exact and easily positioned replica of the surface. A strain measurement can now be made as follows. A hologram of a beam scattered by the undeformed object is made first; the wave field reconstructed from this hologram is then allowed to interfere with the beam scattered by the deformed object.

This article is concerned with the relationship between the deformation of an object and the resulting interference pattern. Particular attention will be given to the conditions that must be fulfilled to obtain high-contrast interference fringes and to the derivation of the deformation from the fringe pattern. Finally, a number of practical applications will be described.

Since it is an essential feature of the method a short explanation of holography will be given first.

Dr C. H. F. Velzel is with Philips Research Laboratories, Eindhoven.

Holography

When light waves forming a stationary wave field fall on a photographic plate, the *intensity* distribution of the field is recorded as a blackening of the photographic plate. In holography both the local *amplitudes* and *phases* of the wave field are recorded on the plate by allowing a coherent *reference wave* to fall on the plate as well as the original light. The resultant pattern of blackening on the plate, the hologram, can be used at any later time to produce a wave field having exactly the same amplitudes and phases in the plane of the hologram as the original field. In accordance with

Huygens's principle the original field in the space behind the hologram has then also been exactly reconstructed.

Fig. 1 illustrates schematically how holography works. To make the hologram, coherent light from a laser L is split into two beams, a reference beam A and a beam incident on the object V. Part of the light scattered from the object, the 'object beam' B is allowed to interfere with A. The resulting interferogram is recorded on a photographic plate and this is the hologram H. It is assumed for simplicity that A is a plane wave. If B were also a perfectly plane beam, H would consist of exactly parallel interference fringes. Because B is not a plane wave the blackening of H is modulated in a manner which depends on the local amplitudes and phases of B (see *fig. 2*).

To reconstruct the original field (fig. 1*b*), the hologram is illuminated with the same reference beam A. Diffracted beams appear behind H, one of zero order (G_1) and two of the first order (G_2, G_3). Because the interference pattern on H is not spatially exactly periodic, these beams exhibit local variations in amplitude and phase. In particular — and this will appear presently — the variations in G_2 are exactly the same as those in B so that G_2 is a *faithful copy* of B. The observer O in fig. 1*b* thus 'sees' the object V once more; a virtual image V' is formed of V. The beam G_3 forms a real but reversed and distorted image V'' of V.

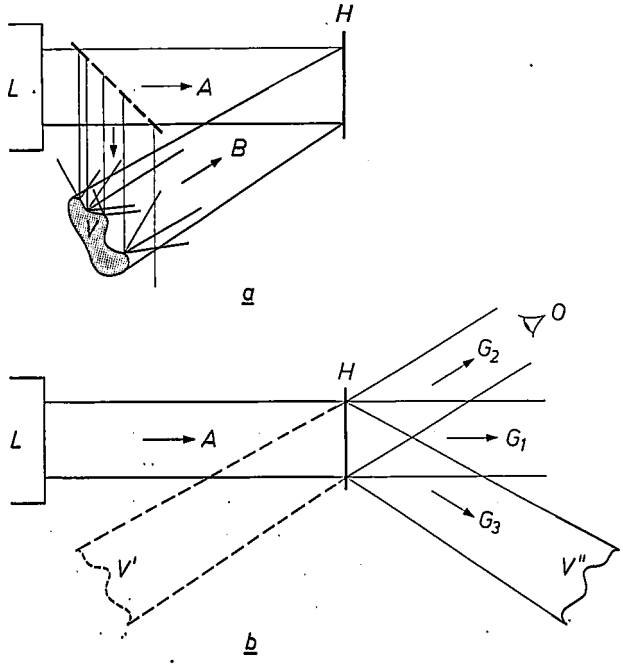


Fig. 1. Holographic recording (*a*) and reconstruction (*b*) of a light beam scattered by an object. L laser. To make the hologram, the laser beam A is divided into two beams by means of a beam splitter. The hologram is recorded on a photographic plate (H); it is the interferogram arising from the interference between the unmodified laser beam (the reference beam A) and the beam B scattered by the object V. If the same laser beam is allowed to fall on the hologram after development of the plate, then one of the diffracted beams (G_2) is identical to the object beam B, and an observer O sees a virtual image V' of the object V. The diffracted beam G_3 forms a real image V'' which is reversed and usually distorted. The 'efficiency' of the reconstruction (power of G_2 divided by that of A) is often very small because light is absorbed by H and also because most of the transmitted light resides in the zero-order beam G_1.

To obtain the above results the apparatus must strictly fulfil certain conditions. Firstly, the coherence of the light must be sufficient for a hologram to be formed. To achieve sufficient coherence the optical path between light source and hologram via the reference beam is made as nearly equal as possible to that via the object beam — in fact more mirrors than shown in fig. 1*a* are used — and a laser is used as the light source. In addition the apparatus must be rigid and unaffected by vibration: while the plate is being exposed to form the hologram, the interference fringes must remain stationary, i.e. all relative movements must be limited to less than a quarter of the wavelength of the light used. Finally, the photographic plate must have a high resolution: the spacing between the interference fringes is usually not much greater than a wavelength and it is necessary to record variations of detail in this fine pattern.

It is not essential for the reference beam to be a plane wave, nor is it essential for the beam to be normally incident on the hologram. However, it *is* essential that the reference beam used in reconstruction is exactly the same as the beam used to form the hologram and of the same orientation with respect to the hologram.

Holography was invented by D. Gabor [1] in 1947, long before the existence of lasers. In Gabor's first holographic experiments the object was small and transparent and situated on the axis between the source and the hologram. The light passing round the object formed the reference wave. The difference in optical path length between rays in the reference wave and those scattered by the object were then so small that the coherence of the 'non-coherent' sources which Gabor had perforce to use was sufficient to produce a hologram carrying interference fringes. This geometry, which corresponds to an angle of zero between the beams A and B in fig. 1, had the disadvantage that the virtual image could not be observed without unwanted effects due to the light from G_1 and G_3. Holography grew to fruition only after the advent of the laser. The laser is a source of coherent light, i.e. the light emitted has a long coherence length so that interference can be obtained even when the optical paths differ considerably in length. Holographic images of opaque and large objects can therefore be made. In addition, the reconstructed wave can easily be kept separate from the other beams emerging behind the hologram [2].

Relation between object beam, hologram and reconstructed beam

By making use of Huygens's principle, from which the structure of the object-beam field B in fig. 1 is determined by its amplitudes and phases in the plane of the hologram H, it will now be shown that H does indeed

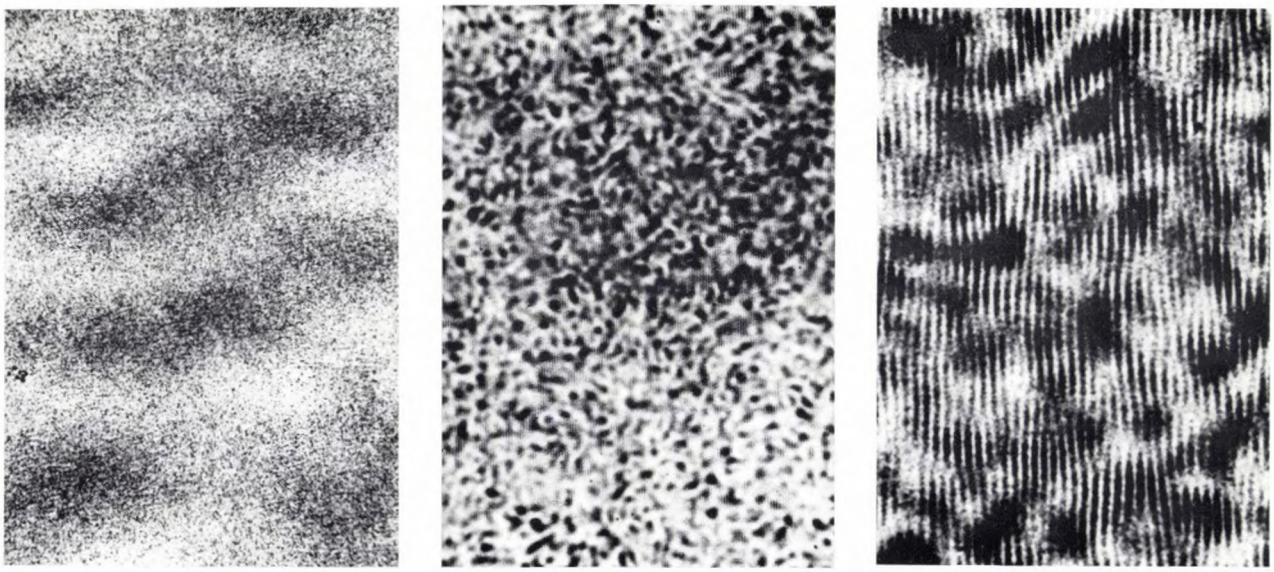


Fig. 2. Three enlargements of a part of a hologram. Magnifications (left to right): about 40×, 180× and 700×. As the right-hand photograph shows, the blackening forms a sinusoidal grating modulated in phase and amplitude; the spacing of the fringes, which can also be faintly distinguished in the central photograph, is 2 μm. The essential information is registered primarily in the phase modulation (the curvature of the fringes). The speckle pattern, which can also be seen in the following photographs, is not characteristic of the object, nor of the hologram, but of the use of light of high coherence. It is a kind of random distribution of diffraction discs. The discs increase in size as the angle subtended by the source at a point in the hologram decreases.

contain the necessary information concerning B and that G_2 is a faithful copy of B. We take a coordinate system x,y,z in which $z = 0$ is the plane of the hologram. We assume that the reference beam A is a plane wave propagating along the z-axis. This wave is described as a function of time and place by:

$$u_A = a \exp j(\omega t - kz), \qquad (1)$$

where u_A is the local instantaneous complex amplitude, ω is the angular frequency of the wave and k its wave number, i.e. 2π divided by the wavelength λ. The phase ϕ_a of the wave for $t = 0$, $z = 0$ and the peak amplitude a_0 are combined in the 'complex amplitude' a:

$$a = a_0 \exp -j\phi_a. \qquad (2)$$

Suppose that the object beam B propagates in the x,z-plane at an angle β to the z-axis. If this was a plane wave, we could express it as:

$$u_B = b_0 \exp j(\omega t - kx \sin\beta - kz \cos\beta - \phi_b), \qquad (3)$$

a wave whose wave vector has components $k \sin\beta$ in the x-direction and $k \cos\beta$ in the z-direction. The complex amplitude b of this wave in the plane $z = 0$ is a function of x:

$$b = b_0 \exp -j(kx \sin\beta + \phi_b).$$

When B is not exactly a plane wave, but very nearly — as will be the case if the angle subtended by the object at the hologram is not too large — it can still be represented by (3) although b_0 and ϕ_b are now weakly dependent on x, y and z (the derivatives of b_0 and ϕ_b with respect to x, y and z must be small compared with $k \sin\beta$). According to our assumptions, the detailed structure of the wave field B is entirely determined by the angle β and the functions $b_0(x,y)$ and $\phi_b(x,y)$ in the plane of the hologram $z = 0$.

While the plate is being exposed to form the hologram the total complex amplitude at every point x,y of the hologram is equal to $a + b$. The intensity is therefore given by:

$$I(x,y) = |a + b|^2 = a_0^2 + b_0^2 + a^*b + ab^*. \qquad (4)$$

If the photographic plate is exposed for a time t and developed, we get a hologram of transmittance $\tau(x,y)$ which is assumed here to be a linear function of I. In practice this can only be approximately true. *Fig. 3* shows qualitatively how τ in general depends on the exposure $H (= It)$. Only in a limited region near the point of inflexion H_0,τ_0 is the curve approximately straight. To keep within this region as far as possible the average intensity b_0^2 of the object wave is made much smaller than the intensity a_0^2 of the reference wave and $a_0^2 t$ is made approximately equal to H_0. In this linear approximation we have:

$$\tau = \tau_0 - \tau_1 b_0^2 - \tau_1(a^*b + ab^*) \qquad (5a)$$

$$= \tau_0 - \tau_1 b_0^2 - 2\tau_1 a_0 b_0 \cos\{kx \sin\beta + \phi_b - \phi_a\}. \qquad (5b)$$

[1] D. Gabor, Nature **161**, 777, 1948, and Proc. Roy. Soc. A **197**, 454, 1949.

[2] E. N. Leith and J. Upatnieks, J. Opt. Soc. Amer.: **52**, 1123, 1962; **53**, 1377, 1963; **54**, 1295, 1964.

If the small term $\tau_1 b_0^2$ is neglected, this shows that the variations in the phase $\phi_b(x,y)$ and the amplitude $b_0(x,y)$ of the object wave are contained in the hologram in the form of variations in the phase and amplitude of the modulation of the transmittance.

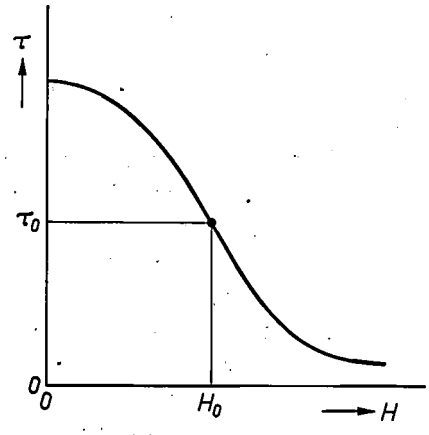


Fig. 3. Transmittance τ of a photographic plate as a function of the exposure H. The exposure is the product of the intensity of the light and its duration. It is preferable to work only on the linear part of the curve, near the point of inflexion (H_0,τ_0).

In the reconstruction of the original wave field, the hologram is illuminated by the reference beam A. Behind the hologram we then get a wave of complex amplitude τa. Using (5a) we find:

$$\tau a = (\tau_0 - \tau_1 b_0^2)a - \tau_1 a_0^2 b - \tau_1 a^2 b^*. \qquad (6)$$

The wave behind the hologram is thus a superposition of three waves with the complex amplitudes

$$g_1 = (\tau_0 - \tau_1 b_0^2)a,$$
$$g_2 = -\tau_1 a_0^2 b,$$
$$g_3 = -\tau_1 a^2 b^*.$$

These are the three waves of fig. 1*b*. The wave G_1 of complex amplitude g_1 has the same phase as the reference beam A in the plane of the hologram and is therefore propagated in the same direction (the z-direction). The factor $(\tau_0 - \tau_1 b_0^2)$ implies that G_1 exhibits small variations in amplitude in the plane of the hologram so that the beam behind the hologram diverges somewhat.

The wave G_2 of complex amplitude g_2 is identical, apart from the constant factor $-\tau_1 a_0^2$, to the object beam B. This is the result we sought.

Finally there is the wave G_3 of complex amplitude g_3. The argument of g_3 is $kx \sin\beta + \phi_b - 2\phi_a$. The wave G_3 is thus propagated in the x,z-plane at an angle $-\beta$ to the z-axis. More generally — even if A and G_1 are not propagated along the z-axis — G_3 and G_2 are symmetrical with respect to G_1. Thus if β is made large enough to separate G_2 from G_1, then G_2 is certainly separated from G_3.

The treatment given above refers to an *absorption hologram*: the transmittance (5) is real and the wave field behind the hologram differs only in amplitude and not in phase from the wave field incident on the hologram. In general, however, the transmittance is a complex quantity, so that the hologram changes the wave in both amplitude and phase. If the phase only is changed, we have a *phase hologram*. A phase hologram can be made by 'bleaching' away the blackening of a holographic pattern by rehalogenization [3]. The blackening vanishes but the places that were black remain slightly thicker; at these places the reference wave undergoes a larger phase change than in the non-thickened surroundings. A phase hologram has the advantage, often important in practice, that it transmits more power.

Three points will now be briefly discussed: the nature of the image formed by G_3, the nonlinearity of the transmittance characteristic and the efficiency of the hologram.

When we omit the constant factor $-\tau_1 a^2$, the complex amplitude of the wave G_3 is b^*. For brevity we write $kx \sin\beta + \phi_b$ as ψ, a function of x, y and z. We now restrict ourselves to the plane of the hologram and its close neighbourhood. Because ψ does not vary rapidly with z we can put ψ equal to its value $\psi(x,y)$ in the plane of the hologram. Again omitting the constant factor $-\tau_1 a^2$, the waves G_2 (or B) and G_3 are expressed respectively by:

$$b_0 \exp j(\omega t - kz\cos\beta - \psi),$$
$$b_0 \exp j(\omega t - kz\cos\beta + \psi).$$

The wavefronts — surfaces of constant phase — are thus given by the equations

$$kz\cos\beta = -\psi(x,y) + C \quad \text{for } G_2 \text{ and } B,$$
$$kz\cos\beta = +\psi(x,y) + C' \quad \text{for } G_3.$$

It follows that the wavefronts of G_3 are the mirror images of those of G_2 with respect to the plane $z = 0$ (*fig. 4*); the amplitudes at corresponding points are equal. Taking into account also the direction of propagation of the light rays, it can be seen from fig. 4 that rays in G_2 which diverge from (for example) the virtual image point P' correspond to rays in G_3 which converge towards the symmetry point P''. The virtual image V' thus corresponds to the real image V''. The latter is reversed: if the point Q' lies *in front* of P' for the observer O, then the corresponding point Q'' lies *behind* the point P'' for the observer O'. An essential assump-

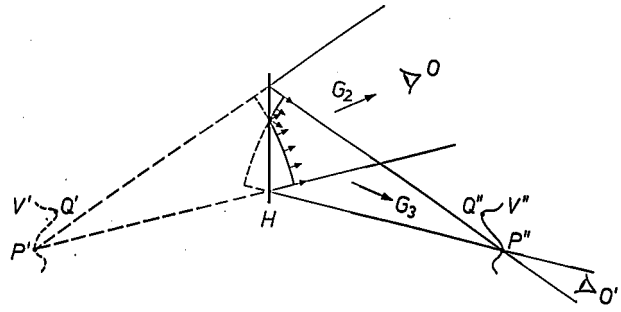


Fig. 4. Close to the hologram, the wavefronts in the beams B (or G_2) and G_3 of fig. 1 are mirror images of each other with respect to the plane of the hologram. Rays in G_2 diverging from the points P' and Q' correspond to rays in G_3 which converge to the corresponding points P'' and Q''; the virtual image V' thus corresponds to the real image V''.

tion in the foregoing is that the factor $-\tau_1 a^2$ is constant. This is the case only if the reference beam A is propagated along the z-axis. If A falls obliquely on the hologram, the image formed by G_3 is distorted.

Taking into account the nonlinearity of the transmittance, i.e. the terms of higher order in the expansion of τ in powers of H_0 (see fig. 3), then τ is not given by the expression of eq. (5b), but by a polynomial with terms of the form

$$b_0^{m+2n} \cos m(kx \sin \beta + \phi_b - \phi_a),$$

where m and n are integers. On reconstruction, the main term ($m = 1$, $n = 0$) discussed earlier gives the beams G_2 and G_3 in which G_2 reproduces B exactly. When the term with $m = 1$, $n = 1$ does not vanish, the reproduction is no longer perfect: the image has less contrast. The terms with $m > 1$ lead to diffracted beams of higher order which waste light. In order to limit these undesirable effects, b_0 must be made much smaller than a_0, as mentioned earlier.

Finally, the 'efficiency' of a hologram, which is in practice of great importance. This is the ratio of the power of the reconstructed wave G_2 to that of the reference wave during reconstruction. The transmittance of the absorption hologram of a *plane* 'object wave', registered on a photographic plate with a linear characteristic, is (see eq. 5b):

$$\tau = \tau_0 + \tau' \cos (k'x + \phi),$$

where τ_0, τ', k' and ϕ are constants. Since τ has everywhere a value between 0 and 1, the modulation depth τ' cannot exceed $\frac{1}{2}$. In this notation the amplitude of the reconstructed wave is $\frac{1}{2}\tau' a_0$; the efficiency is here, therefore, $\frac{1}{2}\tau'^2$ and this is at most equal to 1/16 or about 6%. In practice the hologram is not uniformly exposed and, as discussed earlier, τ' is preferably taken much smaller than $\frac{1}{2}$; the efficiency is therefore generally much smaller, usually less than 2%. A phase hologram transmits more energy and may have an efficiency of 10 to 15%. So far it has been tacitly assumed that the emulsion coating of the photographic plate that carries the hologram is very thin. In holograms which are thick compared with the wavelength, the efficiency can exceed 50% as a result of multiple reflections. Holograms with a thickness of several hundred microns have been made both in photographic emulsions and in crystalline materials such as $LiNbO_3$ and KBr [4].

Strain measurement by holography

There are two ways in which the 'interference of non-simultaneous beams of light' can be brought about in practice. In one method a hologram is made before the object is deformed. After development the plate is accurately replaced in its original position in the otherwise unchanged arrangement. The object is then deformed and viewed through the hologram. The viewer now sees both the reconstructed image of the undeformed object and the object itself in its deformed state, illuminated by the source. With this method the interference fringes and hence the deformation of the surface can be viewed as they occur. This is called the method of *time-dependent interference* [5]. In the other method two holograms are recorded on one photographic plate: one of the object wave before deformation and one after deformation. Since the only term relevant to the analysis of the holographic process is linear in the complex amplitude of the object wave (see eqs. 5 and 6), holograms can indeed be superposed and the reconstruction gives a superposition of the two object waves. The resulting interference pattern contains the required information. In this article the discussion will be limited to this *double-exposure* method [6].

If we are concerned with two essentially stationary states of an object such as a bar whose deformation is to be measured, there is no limitation to the choice of exposure time. The plate is exposed before the object is deformed and again afterwards. In other cases, however, e.g. in the analysis of fracture, the situations of interest are not stationary. The exposure then has to be so short that the deformation occurring during the exposure is small compared with the wavelength of the light. This condition can often be met by using a pulsed laser [7]. When the deformation is not only fast but also periodic, as in a vibrating loudspeaker, a stroboscopic illumination can be used [8]. Finally, information about vibrating objects can be obtained from a hologram recorded with an exposure much longer than the period of vibration — time-averaged holograms [9].

The most important problem in holographic strain analysis is the interpretation of the interference pattern and we shall concern ourselves mainly with this problem. From what has been said above it will be clear that the holographic process itself can be disregarded here, and the object can be considered to be simultaneously present in its two states, deformed and undeformed.

It is assumed that the microstructure of the surface of the object does not change during the deformation. The same point then scatters the light in the same way before and after the deformation — the only difference is that due to the actual deformation. It is also assumed that the object scatters the light *diffusely*, i.e. that the scattered light has lost most of its spatial coherence. Interference then occurs only between rays coming from the same point in the two situations; other interferences do not contribute effectively to the interference pattern.

[3] R. J. Collier, C. B. Burckhardt and L. H. Lin, Optical holography, Academic Press, New York 1971, p. 289.
[4] G. Kalman, in: Applications of lasers to photography and information handling (ed. R. D. Murray), Soc. Phot. Sci. and Engrs., 1968, p. 99.
[5] K. A. Stetson and R. L. Powell, J. Opt. Soc. Amer. 55, 1694, 1965.
[6] R. J. Collier, E. T. Doherty and K. S. Pennington, Appl. Phys. Lett. 7, 223, 1965.
[7] R. E. Brooks, L. O. Heflinger, R. F. Wuerker and R. A. Briones, Appl. Phys. Lett. 7, 92, 1965.
[8] P. Shajenko and C. D. Johnson, Appl. Phys. Lett. 13, 44, 1968.
[9] R. L. Powell and K. A. Stetson, J. Opt. Soc. Amer. 55, 1593, 1965.

Determination of the longitudinal component of the displacement vector

For an opaque object the light source and the observer must be situated on the same side of the object. In this configuration the change in optical path from source to observer via the object has its origin in the displacement component in the mean direction between source and observer. The information in the interference pattern thus concerns this 'longitudinal component'.

Fig. 5 shows how the difference in optical path depends on the displacement. If, as in fig. 5*a*, the source S and the observer O both have the same direction when viewed from an object point P and if P is also displaced in this same direction (to Q), the optical path difference $\Delta l = SPO - SQO$ is equal to twice the length of the displacement vector PQ. If PS and PO are at an angle χ to each other (fig. 5*b*) and the displacement vector $e = PQ$ lies along the bisector of the angle SPO, the path difference is modified by a factor $\cos \frac{1}{2}\chi$:

$$\Delta l = 2e \cos \tfrac{1}{2}\chi. \qquad (7)$$

Fig. 5*c*, finally, shows a displacement vector of arbitrary direction, obtained by locating Q at some arbitrary point in the plane perpendicular to the bisector (fig. 5*b*). The optical path SQO does not change since SQ becomes just as much shorter as QO becomes longer. Equation (7) therefore remains valid if we replace e by its component e_1 along the bisector (the longitudinal component):

$$\Delta l = 2e_1 \cos \tfrac{1}{2}\chi. \qquad (8)$$

In fig. 5 it is assumed that we have a point light source and a point observer and that they are both situated at infinity, i.e. that e is very much smaller than PS and PO.

The observer looking towards P sees light or darkness according to whether Δl is an even or odd multiple of $\lambda/2$. He thus sees light and dark fringes of equal longitudinal displacement; the interference pattern is a 'contour map' of the deformation. Examples of such patterns are shown in *figs. 6* and *7*. In a bright fringe of the pth order ($\Delta l = p\lambda$), we have

$$e_1 = p\lambda / 2\cos \tfrac{1}{2}\chi. \qquad (9)$$

If χ is not too large, $\cos \frac{1}{2}\chi$ is approximately unity; the difference in height of the surface between adjacent fringes is then approximately half a wavelength.

The fact that the observing instrument (the eye, a camera, etc.) has a pupil of finite and not zero diameter gives a certain practical limitation to this method. If the opening is so large that the optical path differences Δl for different parts of the pupil differ by more than

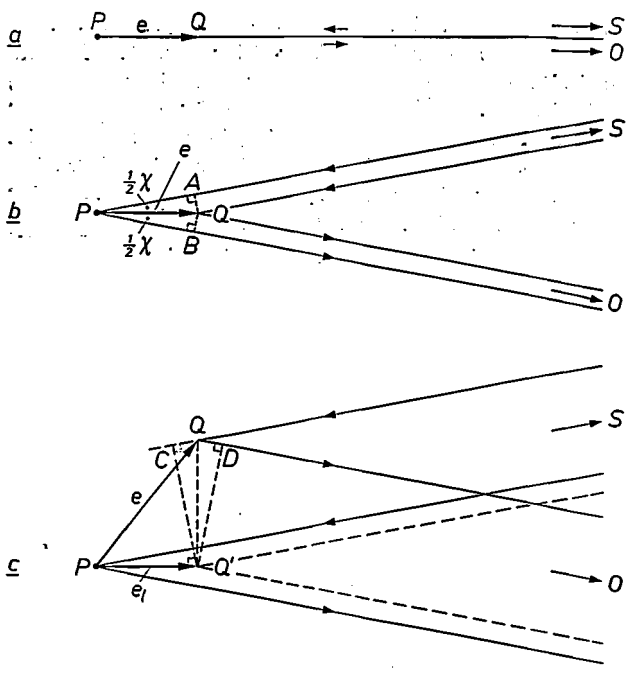


Fig. 5. Calculation of the change Δl in the optical path length between source S and observer O via a point in the surface of the object when this point is displaced from P to Q.
a) PS, PO and the displacement vector $e = PQ$ have the same direction; in this case, $\Delta l = 2e$.
b) PS and PO are inclined at an angle χ to each other and the displacement vector e is directed along the bisector of χ. In this case, $\Delta l = PA + PB = 2e \cos \frac{1}{2}\chi$.
c) When Q is displaced in the plane SPO perpendicular to the bisector PQ' of the angle SPO, the optical path length SQO does not change. The difference QC between SQ' and SQ is equal to the difference QD between QO and QO', so that SQO remains equal to $SQ'O$. This is still true if Q is displaced perpendicular to the plane SPO; SQ and QO do not then change at all. The difference Δl in optical path length therefore has the same value as in (*b*) except that e must be replaced by $e_1 = PQ'$, the longitudinal component of e.

a half-wavelength, the interference fringes fade. From *fig. 8* it can be seen that this is not the case provided the angle α subtended by the pupil at the object is less than $\lambda/2e_t$ where e_t is the component of the displacement perpendicular to PO. A more detailed calculation [10] shows that if this condition is fulfilled, the interference fringes have a contrast of more than 70%. Conversely, for a given aperture, the lateral displacement may not exceed the value $\lambda/2\alpha$. Suppose, for example, a helium-neon laser is used in a conventional arrangement and the interference pattern is observed with the naked eye; λ is then 0.6 μm and α is of the order of 0.01. The maximum permissible lateral displacement is then about 30 μm.

The maximum permissible value of the lateral displacement, $\lambda/2\alpha$, is simply the resolving power of the observing instrument. The foregoing is another way of saying that two points separated by a distance of more than $\lambda/2\alpha$ are separately imaged and that the rays coming from them no longer interfere with each other.

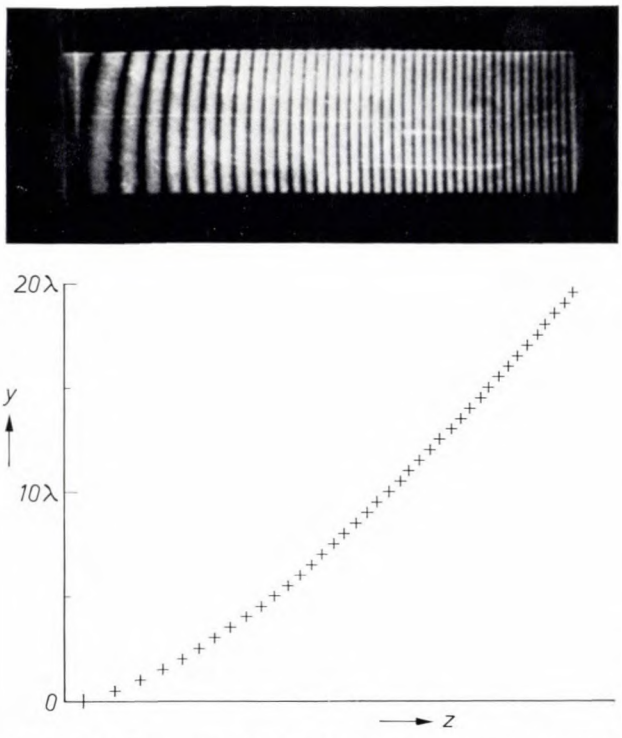


Fig. 6. *Above:* Example of an interference pattern on the image of a bar deformed under load (obtained from a double hologram). *Below:* Displacement y as a function of position z along the bar, derived from the interference pattern. Each fringe represents an increment of the displacement of $\lambda/2$.

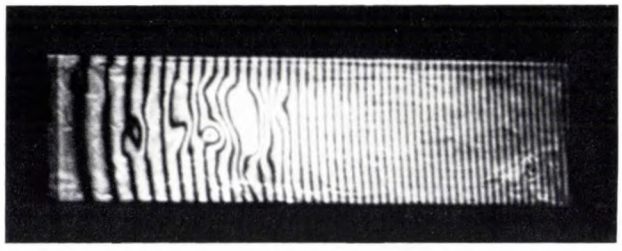


Fig. 7. Interference pattern on the double-hologram image of a bar with a weak spot. The bar is made up of two parts held together by adhesive, the plane of the joint lying in the plane of the paper. The irregularity in the interference fringes indicates a weak spot in the joint.

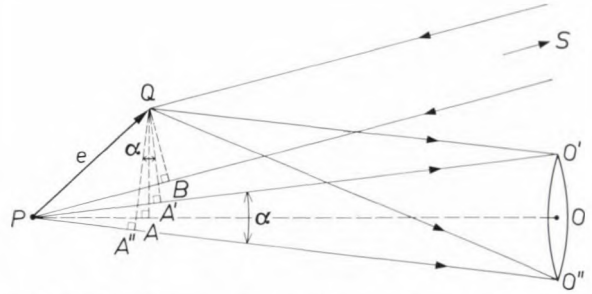


Fig. 8. Relation between the size of the pupil of the observer and the maximum permissible lateral displacement. The observed interference pattern tends to vanish if the pupil becomes so large that the differences in optical path lengths $SPO' - SQO'$ and $SPO'' - SQO''$ differ by more than $\lambda/2$. This difference is equal to $(PA' + PB) - (PA'' + PB) = PA' - PA'' = \alpha e_t$, where $e_t = QA$ is the lateral component of $\boldsymbol{e}$ and α is the angle subtended at P by the pupil. For a high-contrast interference pattern, α must therefore be less than $\lambda/2e_t$.

Determination of the transverse components of the displacement vector

In the method discussed above only the longitudinal component of the displacement vector is measured. In principle the complete displacement vector can be derived in this way by viewing the object from three directions or by making three exposures with different directions of the illuminating beam [11]. These procedures do, however, involve rather a lot of work with regard to both the experimental arrangement and the interpretation of the interferogram.

We have developed a method whereby the transverse components of the displacement can be measured directly. We look at the hologram from a point in a *real* holographic image. The required real image could be made by placing a positive lens in the beam G_2 in fig. 1*b*, but we wish to suggest here, with the help of *fig. 9*, a still simpler method. The hologram is recorded in the conventional way, using a *plane* wave as the reference beam (fig. 9*a*). In the convential method of reconstruction (fig. 9*b*) a good virtual image V' and a distorted real image V'' would be obtained. Now however the wave field is reconstructed with a reference beam A' identical to A as regards wavefronts but propagated in the *opposite* direction. In other words, the back of the hologram is illuminated with a plane wave from the opposite direction (fig. 9*c*). A wave G_2' is then generated at the front of the hologram, and this wave is in all respects a faithful copy of B except that it also propagates in the opposite direction. This can be easily checked from the mathematical description of holographic principles given on page 55/56; it is only necessary to replace ω by $-\omega$ in the description of the reconstruction. All waves are then propagated 'backwards', but otherwise the situation is unchanged. In this way a faithful, undistorted real image V_r of the object V is obtained. In fig. 9*d*, finally, it is shown schematically how this can best be done in practice; in the configuration of fig. 9*b* with a fixed reference beam and the conventional viewing direction, the hologram is simply rotated through 180° about an axis perpendicular to the paper. The situation so obtained is that of fig. 9*c*, assuming that A is indeed a plane wave.

A diaphragm is now placed with its centre at a point P in the real image and the hologram is viewed through it with the naked eye or a camera focused on infinity. Each point on the retina or on the photographic plate now corresponds to a viewing direction. If the hologram is a double hologram of a deformed object an interference pattern is again seen. In *fig. 10*, PQ is the displacement vector of a point in the real image. Since

[10] C. H. F. Velzel, J. Opt. Soc. Amer. **60**, 419, 1970.
[11] J. W. C. Gates, Optics Technol. **1**, 247, 1969.

G_2' is a faithful copy of B, apart from its direction of propagation, there is a fixed relation between the phases of the light at P and at Q which depends only on the positions of P and Q with respect to the light source. The interference pattern owes its existence to the fact that the optical path length $PQ' = e \cos \gamma$ in fig. 10 varies with the viewing direction PR. To put it more precisely: in the direction of a point R on the hologram the field appears light or dark according to whether

$$\Delta l = \Delta l_0 + e \cos \gamma$$

is an even or an odd multiple of $\lambda/2$, where Δl_0 is a constant independent of the viewing direction. Dark or light fringes in the field are lines of constant Δl, and thus also lines of constant γ; they are therefore intersections of the hologram with cones of apex P and axis e. The fringes are thus symmetrical with respect to the projection of e on the hologram. We now let R move along this symmetry axis and define its position by means of γ_0, the angle between PR and PN, the perpendicular to the hologram through P. We then have $\gamma = \varepsilon - \gamma_0$, where ε is the angle between e and PN. Differentiating Δl with respect to γ_0 gives:

$$\mathrm{d}\Delta l/\mathrm{d}\gamma_0 = e \sin (\varepsilon - \gamma_0).$$

For $\gamma_0 = 0$, this is simply equal to the magnitude e_t of the transverse component of e:

$$(\mathrm{d}\Delta l/\mathrm{d}\gamma_0)_{\gamma_0=0} = e \sin \varepsilon = e_t.$$

The optical path difference Δl between adjacent dark or light fringes increases with λ. Therefore the magnitude e_t of the transverse component is given by

$$e_t = \lambda/\Delta\gamma_0, \qquad (11)$$

where $\Delta\gamma_0$ is the angle subtended at P by two adjacent fringes near N. The transverse component is directed towards the centre of curvature of the fringes.

Fig. 11 shows an example of the measurement of a transverse displacement. Fig. 11*a* is the reconstruction via a double hologram of a bar clamped at the top and acted on at its lower end by a force directed to the right. Fig. 11*b* shows some photographs of what a viewer would see on the hologram through a diaphragm of 0.5 mm diameter in the real image. The positions in fig. 11*b* correspond to those in fig. 11*a*.

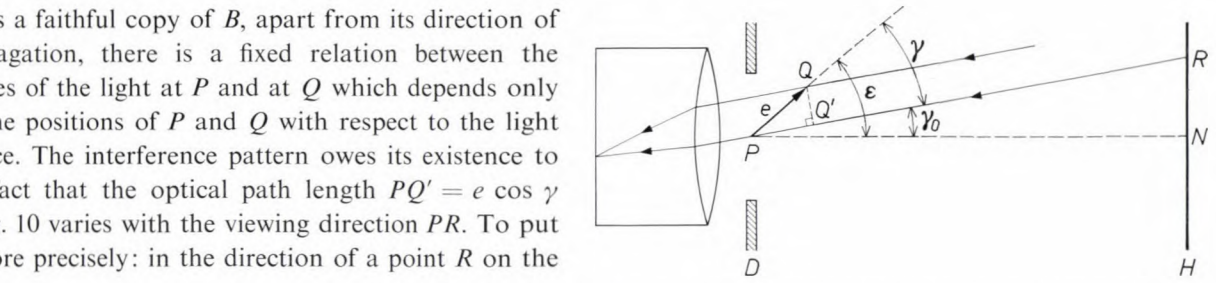


Fig. 10. Calculation of difference in optical path length from the interference pattern seen (fig. 9*c*) on the hologram H via the diaphragm D. P is the initial point and Q the final point of the displacement vector e at a certain location in the real image. An interference pattern is observed on the hologram because the difference PQ' in optical path length depends on the viewing direction PR.

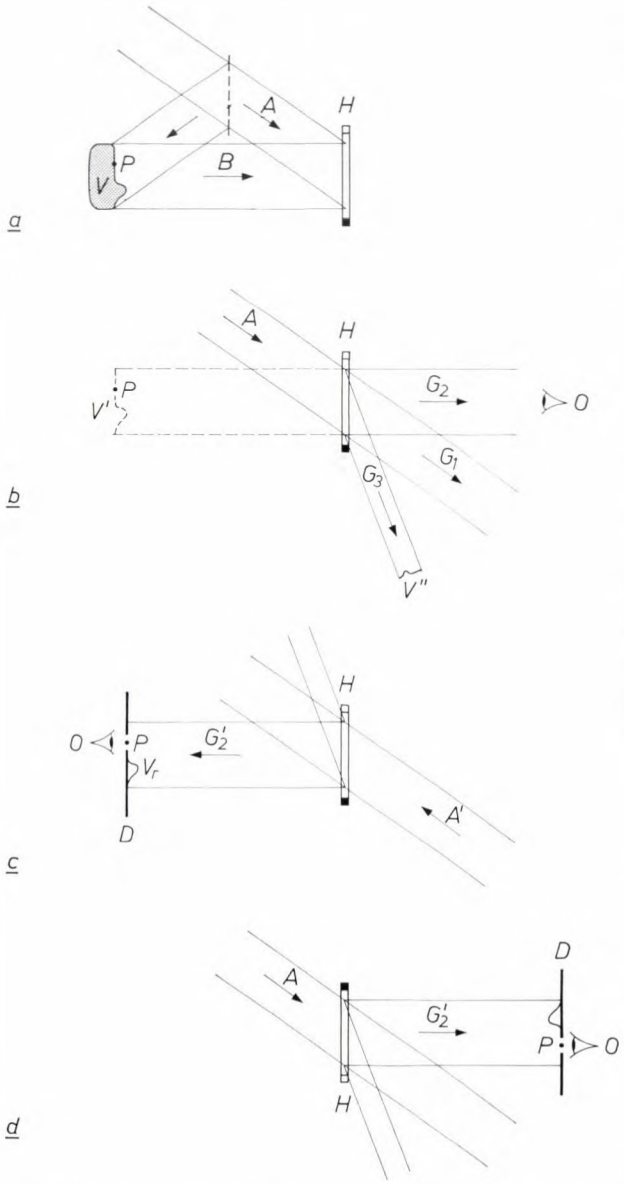


Fig. 9. Determination of the transverse component of the displacement vector. *a*) Recording the hologram. *b*) Usual configuration for reconstruction. *c*) Formation of a real image V_r by reversal of the direction of propagation. The reference beam A' has the same wave fronts as A but is propagated in the opposite direction. If, for example, A were divergent, then A' would have to be convergent. *d*) Practical version of the geometry of (*c*). Starting from the configuration (*b*), the hologram is rotated by exactly 180°; the beam A must be a *plane* wave. The configuration (*d*) is then equivalent to (*c*). The transverse component of the displacement at a point P on the object can then be derived from the interference pattern observed on the hologram through a diaphragm about the point P in the real image.

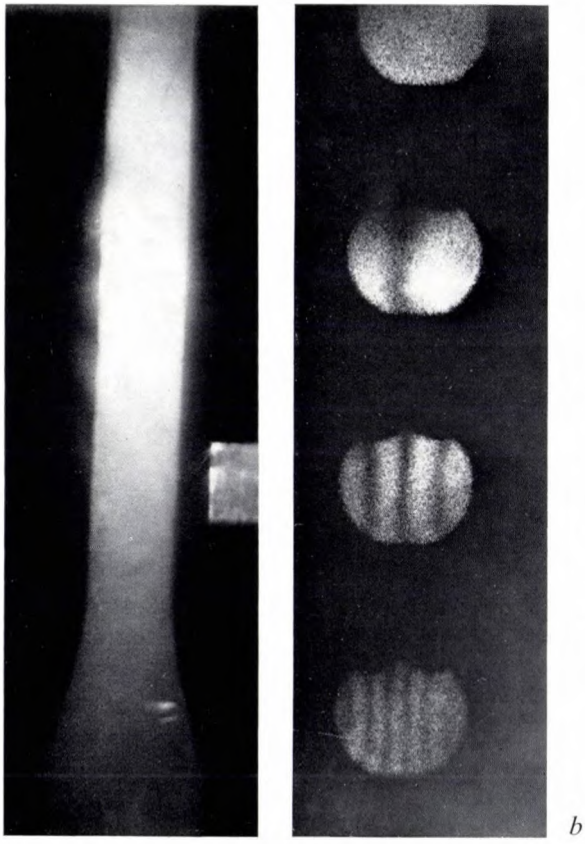


Fig. 11. Measurement of the lateral displacement of a bar subjected to a lateral force. *a*) Double-holographic image of the bar. The bar is clamped at the upper end and the force acts from left to right on the lower end. *b*) Interference patterns observed through a diaphragm of 0.5 mm diameter in the real image. The positions of the fields in (*b*) correspond to those in (*a*). The rather coarse speckle pattern in (*b*) is due to the small aperture (see caption to fig. 2).

To obtain a clearly defined interference pattern, a large number of point pairs such as PQ must interfere in the same way. The diaphragm should therefore be much larger than the transverse component of the displacement. For small deformations like those we are concerned with in this article this condition is always fulfilled. On the other hand, if the deformation varies over the object, the diaphragm may not be too large: the variation of the transverse component over the diaphragm must be so small that the optical path difference for a given viewing direction varies less than $\lambda/2$. If this latter criterion requires the diaphragm to be so small that, with an absorption hologram, the interference field is too faint, an improvement may be obtained by using a bleached hologram. A greater improvement can be obtained if, when recording the hologram, only those parts of the object are illuminated where it is desired to measure the displacement vector.

A disadvantage of the method described here is that a photograph must be made of the interference field on the plane of the hologram for each location where the displacement vector is to be determined. This disadvantage can be partly overcome by making the exposures for different locations simultaneously (see fig. 13, lower photograph).

Methods also exist for projecting contours of equal transverse displacement [12]; these will not be discussed here.

Comparison of the two methods

The two methods discussed above — determination of the longitudinal component of the displacement and determination of the transverse component — will now be briefly compared with the aid of fig. 9. In both cases a double hologram is made in the configuration of fig. 9*a*. To determine the longitudinal component e_l, the virtual image is viewed through the hologram in the configuration of fig. 9*b*; from the interference pattern seen, the distribution of e_l over the surface can be derived. The transverse component must be determined point by point; the magnitude (e_t) and direction are obtained from the interference pattern that a viewer sees in the configuration of fig. 9*d* (hologram rotated 180°) when he views the hologram through a diaphragm at the relevant part of the real iamge.

The two procedures have the following characteristic differences. In the first, only the *variations* of e_l over the surface are determined; e_l changes from fringe to fringe by an amount $\lambda/2$ (taking the factor $\cos \frac{1}{2}\chi$ in (9) equal to 1 for simplicity). Only if we know, for example, that e_l must be zero at a certain location, can the absolute values e_l be determined on the rest of the surface. In the other method, however, the actual values e_t are determined.

The conditions under which a high-contrast interference pattern is obtained are of a different nature in the two methods. In the determination of e_l, e_t must be less than $\lambda/2\alpha$. In the determination of e_t, the difference in optical path lengths and hence $2e_l$ must not vary more than $\lambda/4$ over the diaphragm. This means that the diaphragm in fig. 9*d* must be smaller than the fringe spacing on the object in fig. 9*b*.

Finally, the two procedures differ in sensitivity. The smallest detectable variation in e_l is obtained when (fig. 9*b*) only two fringes are visible at the object; the variation is then $\lambda/2$. The smallest detectable value of e_t occurs when (fig. 9*d*) only two fringes are visible on the hologram. This is the case when, in eq. (11), $\Delta\gamma_0$ is equal to the angle α_H subtended by the hologram at the object. The smallest detectable value of e_t is therefore λ/α_H. In general α_H is less than unity, so that e_l can be determined with greater sensitivity than e_t.

[12] K. A. Stetson, Optics Technol. **2**, 80, 1970.
J. N. Butters and J. A. Leendertz, J. Physics E (sci. Instr.) **4**, 277, 1971.

When all these conditions are fulfilled, both the longitudinal and the transverse components — and hence the complete displacement vector — can be determined from one double hologram.

Applications

Some applications of the double-hologram technique of holographic strain analysis will now be given. The examples may be divided into three areas of application: analysis of experimental models as an aid to design, the determination of material constants and fault-finding.

Strain analysis as an aid to design

For objects of complicated shape it is difficult to predict how they will deform under load or local heating. When the deformation can upset the correct functioning of the object, holographic strain analysis can be useful in analysing the deformation and applying appropriate corrections.

A casting (*fig. 12*) used as a mount for the pick-up tubes and the colour-separation prism in a Philips colour television camera was examined for thermoelastic distortion. In such a camera it is critically important that the relative positions of the pick-up tubes and the prism do not change under the conditions in a studio where strong lighting may cause local heating. *Fig. 13* shows some of the results of this investigation. A double hologram was made of the casting. During the first exposure the casting was uniformly at room temperature; during the second exposure the casting was heated underneath to 3 °C above room temperature. Fig. 13 (above) shows a reconstruction of the casting by means of this hologram with interference fringes which show the variations in the longitudinal displacement. In order to determine the lateral displacement, a real image was formed in the manner indicated in fig. 9*d*. A screen with a large number of holes of diameter 0.1 mm was then set up in the plane of this image. All the interference patterns visible through these holes were then simultaneously registered on a single photographic plate behind the screen. Interferograms recorded in this way are shown in fig. 13 (below); they correspond to a part of the upper photograph.

For a complete analysis of the deformation, double holograms of the casting must be made from various directions, but the discussion here will be limited to a few remarks relating to fig. 13, and in particular to the left-hand and right-hand parts of the wall of the cylinder pointed upwards and outwards towards the observer. The lower left-hand part is displaced longitudinally more than the lower right-hand part (in the upper photograph there are more fringes on the lower left half than on the lower right half); the reverse holds for the upper half. The fringes indicate locally the axes about which the surface (apart from a parallel displacement) tilts. On the right the surface tilts about a horizontal axis, on the left it tilts about an oblique axis. This indicates some distortion of the casting. This is also indicated by the lower photograph; the left-hand part of the cylinder wall has undergone a nearly horizontal displacement while in the right-hand part the transverse displacement is in an oblique direction.

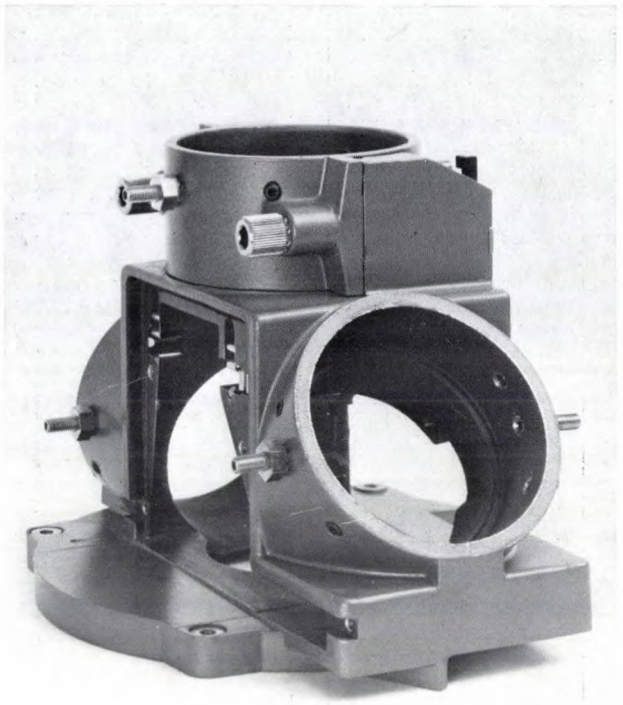


Fig. 12. Mount for pick-up tubes and colour-separation prism in a Philips television camera. The pick-up tubes are clamped in the cylindrical openings directed obliquely forwards, upwards and backwards in the photograph. The casting is shown with its flange on the table. This flange is often subject to local heating from the powerfull lights in the studio.

Determination of material constants

Expansion coefficients and elastic constants of a material can be determined by measuring the deformation taking place as a result of heating or mechanical stress [13]. A simple example of this was given in fig. 6: the determination of the elastic modulus E from the bending of a bar. The bar is clamped at one end (the left in fig. 6) and at the other end subjected to a force F which in fig. 6 is perpendicular to the paper. In fig. 6 (below) the deflection y is plotted as a function of the distance z to the clamping point. No account is taken of the angle χ between the direction of the illu-

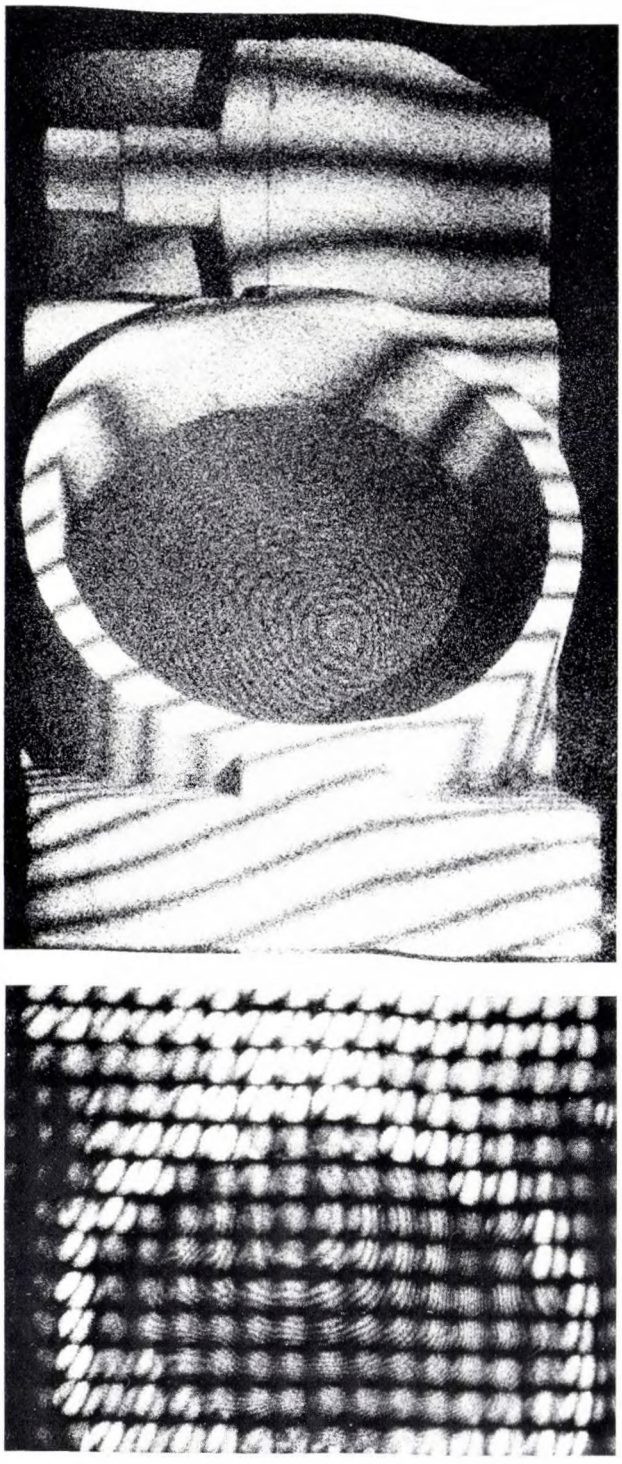


Fig. 13. Strain analysis of the casting shown in fig. 12. After the first exposure of the (double) hologram, the flange (pointing downwards in the photographs) was locally heated three degrees above the ambient; the second exposure was then made. *Above:* determination of the longitudinal strain. *Below:* Determination of the transverse strain. For the lower photograph, the plate was set up close behind a screen with a large number of 0.1-mm holes and situated in the plane of the real image. The interference patterns in the hole of the cylinder have no significance (to prevent undesirable reflections this hole was filled with black paper).

minating beam and the viewing direction, which was about 20°, so that the difference in height between adjacent fringes is not exactly $\lambda/2$ but $\lambda/(2\cos\frac{1}{2}\chi) = \lambda/(2\cos 10°) = \lambda/1.97$ (see eq. 9). The theoretical relation between y and z is:

$$y = Fz^3/4EI, \qquad (12)$$

where I is the second moment of area of the cross-section of the bar. The value of E is found by fitting (12) to the experimental points. It has to be remembered that the zero point of y in fig. 6 is not known because the fringes near the clamp cannot be seen properly. Not only can E be determined very accurately, because of the large number of fringes, but there is also a good check on whether the use of eq. (12) is legitimate. For example if the bar were not uniform in thickness or did not have everywhere the same modulus of elasticity, this would be evident from the impossibility of fitting (12) to the experimental points.

Detection of defects

A weak spot in a workpiece usually deforms excessively or non-uniformly when under load. In a carefully chosen double hologram this is immediately evident. A simple example was shown earlier in fig. 7 where a weak spot in a bar consisting of two parts held together by adhesive shows up in the interference pattern. *Fig. 14* shows the result of a check on the adhesion of the screen of a cathode-ray tube to the rest of the tube. At the upper edge the adhesion is good, on the lower edge there is a weak spot.

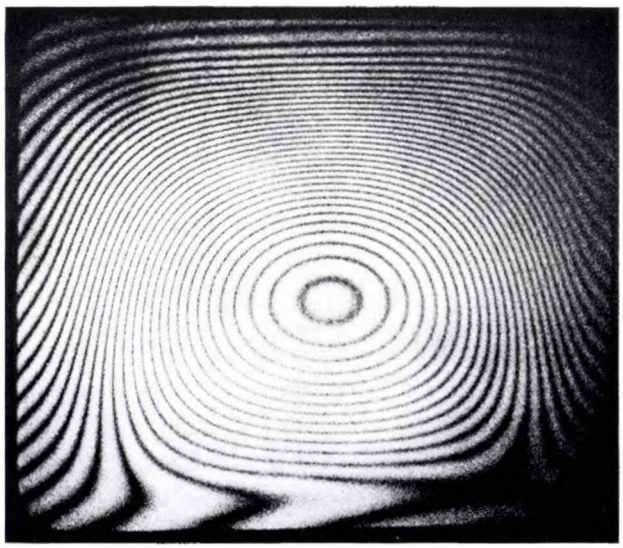


Fig. 14. Tests on a cathode-ray tube. The screen of the tube is cemented to the body of the tube. A double-holographic image can show whether a good adhesion has been achieved. The tube shown here has a weak spot on the lower edge. The mechanical stress consists of a small underpressure in the tube.

[13] See the article by K. A. Stetson referred to in [12].

This method is widely used for the detection of all sorts of faults, e.g. in car tyres, in plastic bottles, in reinforced materials used in the aircraft industry, etc.[14]. The great advantage of holography for such purposes is its non-destructive nature. The stress that has to be applied to the object to detect faults is generally so small that the deformation occurring is completely reversible and very much smaller than any stress applied in normal use.

[14] See for example H. Rottenkolber, Z. Werkstatt u. Betrieb 103, 189 and 245, 1970.

Summary. In holography the amplitude and phase of laser light scattered by an object are registered in a hologram, which is the photographic record of an interference pattern formed by the scattered light (the 'object wave') and a reference beam from the same source. When the hologram is illuminated later by the same reference beam, a number of beams are produced behind the hologram, one of which is a faithful reconstruction of the object wave. In holographic strain analysis, in particular the technique using double exposure, to which the author restricts himself, two holograms are superimposed on one another on the same photographic plate, one before the deformation of the object and one after. The reconstruction procedure then yields a superposition of the recorded object waves corresponding to these two states; the deformation can be derived from the interference between these two waves. The interference fringes seen on the 'reconstructed' object form a contour map of the deformation. For the displacement in the source-observer (longitudinal) direction adjacent fringes represent a difference in strain of $\lambda/2$. The transverse displacement is derived from the interference pattern seen when the hologram is viewed through a diaphragm placed at a point in a real image of the object. This real image is obtained after rotating the hologram through 180°. Holographic strain analysis can be used in the determination of material constants, in the detection of weak spots and as an aid to design. Some examples of these applications are given.

A simplified method for the isostatic hot pressing of ceramics

K. H. Härdtl

The investigations described in this article pave the way for the cheap mass production of pore-free ceramic material, which is increasingly in demand for technological applications. Pore-free materials are already used in applications that include envelopes for high-pressure sodium lamps, cutting tools, acoustic surface-wave filters, and read and write heads for magnetic recording.

Introduction

In the applications of ceramic materials in electrical technology there is an increasing requirement for materials that are completely free of pores. Some examples here are the transparent Al_2O_3 for the envelopes of sodium lamps, (Pb,La)(Zr,Ti)O_3 — known as PLZT — which is also transparent and used in electro-optical applications, Al_2O_3 for cutting tools, piezoelectric ceramics based on Pb(Zr,Ti)O_3 for surface-wave filters, $MoSi_2$ for electric heating elements, and ferromagnetic ceramics (ferrites and garnets) for microwave components and read and write heads in magnetic recording systems.

The conventional sintering process (i.e. sintering at atmospheric pressure) will only yield completely dense products in a few particular cases and then only if special measures such as the addition of dopes are taken. Examples already known include Al_2O_3, to which a small quantity of Mg is added, Y_2O_3 doped with Th, ferrites, and PLZT with an excess of PbO [1].

A process which is commonly used to obtain high densities is 'hot pressing'. Nowadays two methods of hot pressing are in use: *uniaxial hot pressing*, both piece by piece [2] and continuous [3], with a punch and a die of refractory material, and *isostatic hot pressing*, in which the starting material, in powder form, is coated with a thin metallic layer and an inert gas is used to transfer the pressure [4]. Some disadvantages of these processes are that a corrosive reaction often occurs between the die (or metallic coating) and the ceramic material to be compressed, and that only pieces with simple shapes can be manufactured. Moreover, these processes are expensive and not very suitable for mass production.

In this article a method of isostatic hot pressing is described in which no metallic layer is necessary and which does not suffer from the previously mentioned disadvantages [5]. Different technically important ceramic materials such as (Pb,La)(Zr,Ti)O_3, $BaTiO_3$, $SrTiO_3$, Al_2O_3, ferrites and garnets have been used in the studies we have made to find the optimum conditions for the application of this method.

Dr K. H. Härdtl is with Philips Forschungslaboratorium Aachen GmbH, Aachen, West Germany.

The new method

In the second method of hot pressing mentioned above, a thin metallic encapsulating layer is necessary to prevent the working gas from penetrating into the open pores of the starting material. Increasing the pressure would not then cause the pores to disappear.

If the starting material is not a powder, but a ceramic product with pores which are not connected to the surface (closed pores), it can then be pressed without the metallic layer, since the working gas cannot now penetrate into the pores.

The new procedure is therefore performed in two steps: firstly, a conventional sintering process, from which an 'intermediate product' with closed pores is obtained, followed by a second sintering at high gas pressure.

The conventional sintering process which is used for ceramic materials in the electronics industry, such as

[1] For Al_2O_3 doped with Mg see: R. L. Coble, J. appl. Phys. 32, 787 and 793, 1961; for Y_2O_3 doped with Th see R. C. Anderson, U.S. Patent 3,545,987, 8 December 1970; for ferrites see A. L. Stuijts, Proc. Brit. Ceram. Soc. 2, 73, 1964; for PLZT with an excess of PbO see G. S. Snow, J. Amer. Ceram. Soc. 56, 91, 1973.

[2] See for example: P. Murray, D. T. Livey and J. Williams, The hot pressing of ceramics, in: W. D. Kingery (ed.), Ceramic fabrication processes, Technology Press M.I.T., Cambridge, Mass., U.S.A., 1958, pp. 147-171.

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

[4] See for example: H. Bumm, F. Thümmler and P. Weimar, Ber. Dtsch. Keram. Ges. 45, 406, 1968.

[5] Similar work, parallel to ours but independent, has been carried out by two other investigators: see E. A. Bush, U.S. Patent 3,562,371, 9 February 1971, and S. E. Isaksson, British Patent 1,300,864, 20 December 1972.

perovskites, ferrites or garnets, normally yields products with a density of 95 to 99% of the theoretical value. The pores of ceramics with a density of more than 93% are generally closed. This is in agreement with a theoretical result according to which the pores in all ceramic materials are closed when the total porosity is reduced to 9% [6].

To illustrate this point, *fig. 1* shows a photomicrograph of a wafer of $(Pb,La)(Zr,Ti)O_3$ ceramic, which has been sintered conventionally for four hours in an oxygen atmosphere at a temperature of 1170 °C. The residual porosity lies between 1 and 2%. The grain boundaries and also the residual pores are clearly visible after thermal etching of the polished surface. It can be seen that the pores are situated exclusively on the grain boundaries.

At first sight one might suppose that the dark patches are not pores but holes resulting from grains breaking loose from the surface during the polishing. However, evidence to the contrary is supplied in *fig. 2*. These two photographs show the same segment of a conventionally sintered wafer of PLZT, which has been polished on both sides to make it translucent. While the photograph in fig. 2*a* has been taken with reflected light, the one in fig. 2*b* has been made using a combination of reflected and transmitted light. Only surface pores are revealed in fig. 2*a*, all of which can be found in fig. 2*b* (two pores are ringed). In fig. 2*b* many other pores which lie deeper in the ceramic material are made visible by the transmitted light. It can be established that the concentration and dimensions of these deeper pores agree with those of the pores on the surface shown in fig. 2*a*.

It can also be clearly seen in fig. 1 that the residual pores are not directly connected to the surface. We are therefore dealing with closed pores.

A second condition which must be satisfied by our method of isostatic hot pressing is that, at the temperature at which hot pressing occurs, the gas diffusion velocity in the ceramic material must be negligibly small, so that no noticeable quantity of gas can penetrate the pores. In the literature there are no quantitative data on the diffusion constants of the usual working gases (N_2, noble gases, CO_2 etc.) in the ceramic materials which are considered here. Research has however shown that, in the perovskite materials which we have investigated, the diffusion velocity of nitrogen at the sintering temperature is many orders of magnitude lower than that of oxygen [7]. The absolute magnitude of the first diffusion velocity is unknown but it appears to be low enough for our purposes. The same is also valid for argon.

In a later phase of our investigations we were able to confirm that the diffusion velocity of nitrogen in the

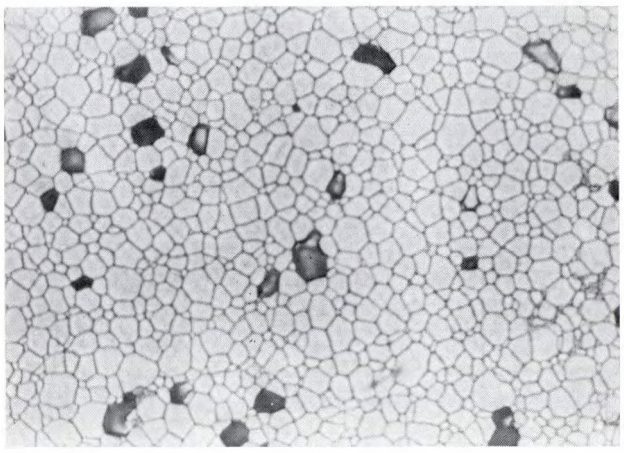
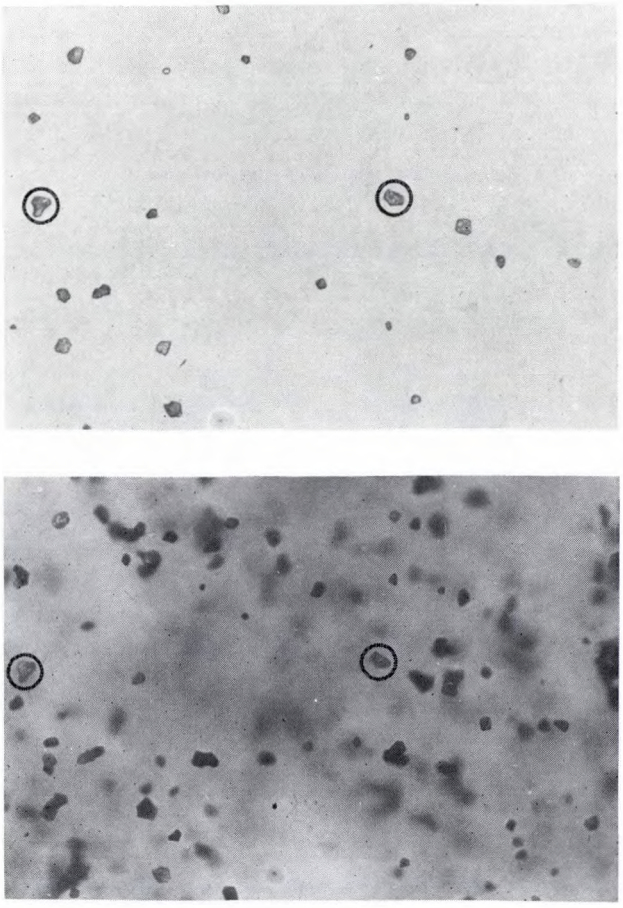


Fig. 1. Photomicrograph of a wafer of $(Pb,La)(Zr,Ti)O_3$ ceramic (molecular proportion of La/Zr/Ti : 11/65/35) which has been sintered conventionally for four hours in an atmosphere of oxygen at a temperature of 1170 °C. The sample has been polished and then thermally etched (1050 °C, one hour in air). The residual porosity is between 1 and 2%. It can clearly be seen that the residual pores (darker) are not linked with each other or the surface. Magnification 800×.



a

b

Fig. 2. Photomicrograph in (*a*) reflected and (*b*) transmitted light of the same wafer of PLZT (11/65/35), which has been polished on both sides. The pores on the surface (two of which have been circled) are visible in both exposures; in (*b*), moreover, many more which are situated in the interior of the plate can be seen. Magnification 400×.

material which we have studied is indeed rather low. This explains not only why nitrogen is a good working gas, but also why, in some applications, nitrogen should not be present in the atmosphere in which the first sintering process (leading to closed pores) takes place. This sintering step must not therefore be performed in air, but in oxygen, for example. We shall come back to this later at the end of the article.

A diagram of the apparatus which we have used for our method of hot pressing is shown in *fig. 3*. The most important components are a water-cooled pressure vessel, which can withstand a pressure of 200 bars, and an electrical furnace with platinum heating elements suitable for a maximum temperature of 1400 °C.

The following procedure is used: the conventionally sintered material is placed in the furnace at an atmosphere of air. After closing the pressure vessel, the working gas (nitrogen or argon) is pumped in until a pressure of approximately 60% of the final value is obtained; the furnace is then brought to the desired temperature in about 30 minutes. Owing to the heating, the pressure in the vessel rises to the desired value, at which it is maintained by means of a regulating valve. The cooling-down period is approximately 30 minutes.

Because the pressure vessel is closed in air, there is always a partial pressure of oxygen present (0.3 to 1 bar), depending on the temperature during pressing. The presence of oxygen prevents a possible reduction of the material at the high temperature during hot pressing. Should it be necessary, in certain circumstances, to have a better control of the oxygen pressure than is described here, a solution is readily found.

Preliminary experiments

(Pb,La)(Zr,Ti)O₃ ceramic

For certain applications $Pb(Zr,Ti)O_3$ ceramic is used in which a proportion of the Pb^{2+} ions are replaced by La^{3+} ions. This increases the dielectric constant, the coupling factor and the transparency. These changes in the properties are connected with the defect structure caused by the substitution of divalent ions by trivalent ions. The surplus charge of the trivalent ions is compensated by vacancies in the Pb or the (Zr,Ti) sublattice.

The preparation of this type of ceramic is impaired by the tendency of the volatile PbO to escape during the sintering process. A simple way of reducing the loss of PbO is to add an extra quantity of PbO to the starting material and to maintain a high PbO partial pressure during the sintering. It has been shown that a comparatively high density is obtained after the first sintering step when these methods are used [1].

In our first experiments, which were carried out with $(Pb,La)(Zi,Ti)O_3$ ceramic (PLZT), the starting materials PbO, TiO_2 and La_2O_3 were weighed out in the required proportions and carefully mixed by dry-milling. We selected the appropriate proportions in

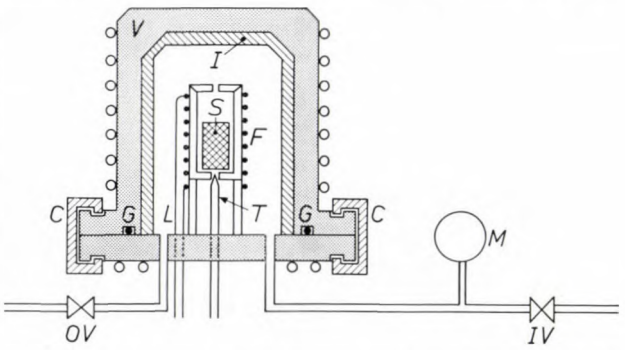


Fig. 3. Schematic cross-section of our apparatus for isostatic hot pressing. *V* water-cooled pressure vessel with sealing clamps *C* and gasket *G*. *I* heat insulation. *F* electric furnace. *S* ceramic sample. *T* thermocouple. *L* supply leads for the heating elements. *IV* gas inlet valve. *M* manometer. *OV* gas outlet valve.

accordance with the formula $Pb_{1-\alpha y}La_yZr_xTi_{1-x}O_3$, in which the parameter α indicates the amount to which lead had been substituted [8]. The mixed powders were pre-sintered for 10 hours at 800 °C, and then carefully dry-milled and compressed isostatically (4 kbars) without a binding agent, to a prismatic or cylindrical shape. The conventional sintering took place in an atmosphere of oxygen at temperatures between 1100 and 1300 °C. $PbZrO_3$ was added in order to maintain a high partial pressure of PbO. A value of the parameter α in the end product of 1.2 to 1.3 was obtained from the weight loss during sintering due to the loss of PbO.

The conventionally sintered intermediate products exhibit a residual porosity of 1 to 0.1%. Since a low porosity had already been obtained at this stage, we were faced with the difficulty of how to determine the further reduction which can be achieved with our method of isostatic hot pressing. We could not use conventional methods to measure the increase in the density, such as immersion in water, or the determination of the change in dimensions, since these methods are not sufficiently accurate. With such low porosities, information concerning the elimination of pores can only be obtained using optical methods. The best information, even if it is only qualitative, is obtained from photomicrographs, taken with either reflected or transmitted light. The latter method is only possible if thin

[6] D. W. Budworth, Trans. Brit. Ceram. Soc. **69**, 29, 1970.
[7] T. F. Murray and R. H. Dungan, Ceram. Ind. **83**, No. 6, 74, 1964.
[8] D. Hennings and K. H. Härdtl, Phys. Stat. sol. (a) **3**, 465, 1970.
K. H. Härdtl and D. Hennings, J. Amer. Ceram. Soc. **55**, 230, 1972.

Fig. 4. Photomicrograph in reflected light of a wafer of the same material as in fig. 1 after isostatic hot pressing for two hours at 1170 °C in nitrogen at a pressure of 200 bars. The wafer has been thermally etched and polished. Magnification 800×.

wafers of a material which does not strongly absorb are used.

Fig. 1 shows the result of conventional sintering of a sample of PLZT ($x = 0.65$, $y = 0.11$), photographed with reflected light. In *fig. 4* a sample of the same material is shown after isostatic hot pressing by the method described here (2 hours at 1170 °C in N_2 at a pressure of 200 bars). It can be seen that the pores have disappeared, while the grain size has not changed during the treatment.

We have also investigated the densification of this material with transmitted light. *Fig. 5a* shows the result of conventional sintering and in fig. 5*b* and *c* the effect of conventional sintering followed by isostatic hot pressing using our method (respectively after 2 and after 16 hours at 1170 °C in N_2 at a pressure of 200 bars) is seen.

The majority of the pores had already disappeared after 2 hours and the porosity was almost completely eliminated after 16 hours. We do not know whether the small dark specks which are visible in fig. 5*c* represent remaining pores, or whether we are concerned here with a separated second phase (possibly PbO). Electron microprobe studies have left this question unanswered.

The results of similar experiments with related materials are reproduced in *Table I*. The tabulated theoretical densities have been calculated from the molecular weight obtained from the composition, the number of molecules per unit cell and the unit-cell dimensions, as determined by X-ray diffraction. Since the distribution of the vacancies at Pb or (Zr,Ti) sites is not known with certainty in PLZT [8], the molecular weight, and therefore the theoretical density, cannot be accurately given. It is apparent from the table that our method of isostatic hot pressing leads to the theoretical density for all the materials studied, as is confirmed by photomicrographs.

In order to obtain an impression of the increase in transparency which is obtained with our method, we have measured the attenuation coefficient for light with a He-Ne laser; for PLZT with $x = 0.65$ and $y = 0.11$ we found a value of 4.5 cm^{-1}. This means that in a ceramic wafer with a thickness of 100 μm, as is used in electro-optical applications, the light loss due to absorption and scattering is smaller than 5%.

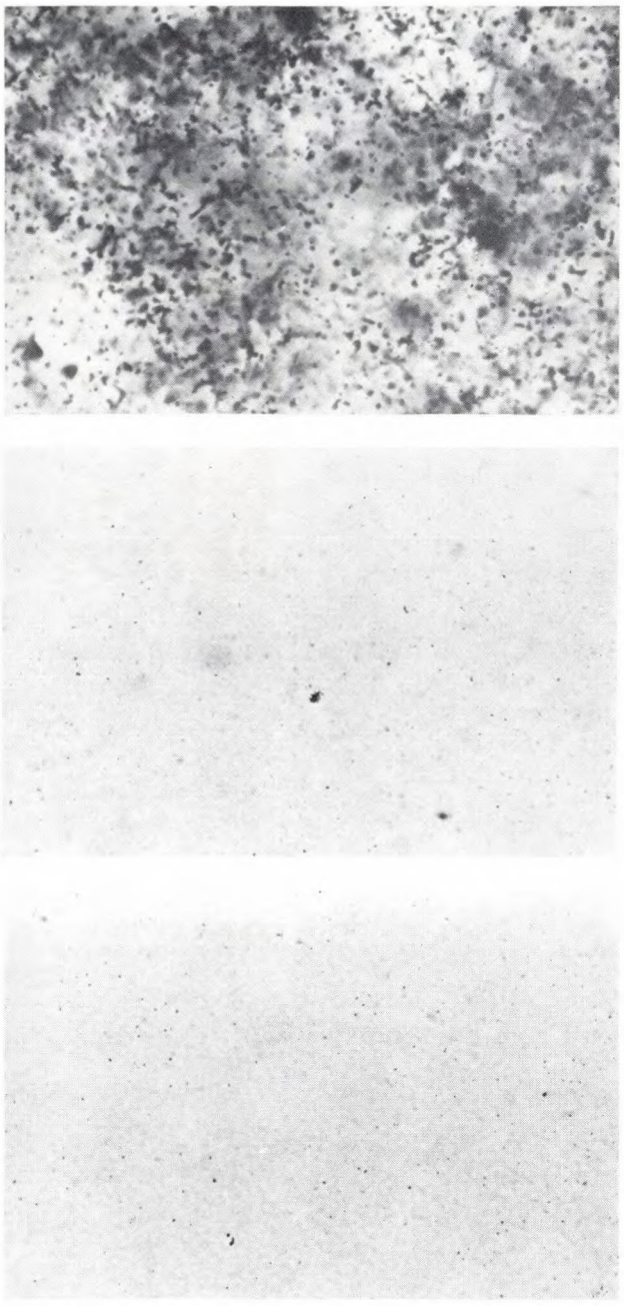
The influence of pressure, temperature and sintering time

$BaTiO_3$ and $SrTiO_3$ ceramics

In order to obtain a better idea of the effect of pressure, temperature and duration of our method of hot pressing on the shrinkage behaviour of ceramic objects, we have also studied the ceramic materials barium

Table I. Experimental data obtained with the isostatic hot-pressing method without an encapsulating metallic layer, carried out with $Pb_{1-y}La_yZr_xTi_{1-x}O_3$ and other materials with a perovskite structure.

Composition	Pressure (bars)	Time (hrs)	Temperature (°C)	Initial density (g/cm³)	Final density (g/cm³)	Theoretical density (g/cm³)
$Pb_{1-y}La_yZr_xTi_{1-x}O_3$						
$x = 0.65$ $y = 0.025$	200	0.5	1250	7.88	7.95	7.91-7.95
$x = 0.65$ $y = 0.06$	200	2	1150	7.73	7.86	7.83-7.86
$x = 0.65$ $y = 0.075$	200	0.5	1250	7.68	7.79	7.78-7.81
$x = 0.65$ $y = 0.09$	200	2	1170	7.77	7.80	7.75-7.79
$x = 0.65$ $y = 0.11$	200	0.67	1170	7.63	7.72	7.70-7.74
$x = 0.65$ $y = 0.14$	200	0.5	1220	7.57	7.68	7.64-7.69
$x = 0.58$ $y = 0.075$	200	0.5	1250	7.65	7.78	7.78-7.81
$x = 0.62$ $y = 0.075$	200	0.5	1250	7.69	7.79	7.78-7.81
$x = 0.72$ $y = 0.075$	200	0.5	1250	7.67	7.79	7.78-7.81
$x = 0.76$ $y = 0.075$	200	0.5	1250	7.67	7.77	7.78-7.81
$x = 0.80$ $y = 0.075$	200	0.5	1250	7.63	7.79	7.78-7.81
$Pb(Ti_{0.5}Zr_{0.5})_{0.985}Fe_{0.015}O_3$	200	2	1250	7.95	8.01	8.01
$PbTi_{0.42}Zr_{0.58}O_3$ + 1 mol % MnO	200	2	1250	7.98	8.01	8.01



a

b

c

Fig. 5. Photomicrograph of a sample of PLZT (11/65/35), taken with transmitted light. *a*) Initial condition after conventional sintering. *b*) After hot pressing by our method (two hours, 1170 °C, in N_2 at 200 bars). *c*) After hot pressing under the same conditions, but now for 16 hours. Magnification 150×.

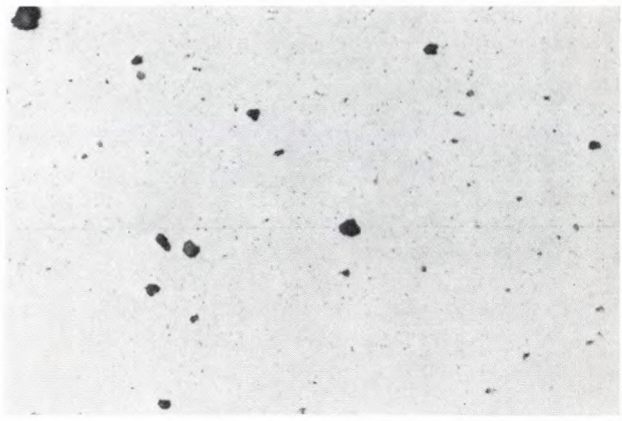


Fig. 6. Photomicrograph of a conventionally sintered wafer of $BaTiO_3$ ceramic. Magnification 150×.

titanate and strontium titanate. These materials are chosen rather than PLZT for two reasons. Firstly, there is no risk during the hot pressing of a volatile component evaporating, as is the case to a certain extent with PbO in PLZT. Secondly, the theoretical density of $BaTiO_3$ and $SrTiO_3$ is known to a much greater accuracy than that of PLZT.

In the case of $BaTiO_3$ we used $BaCO_3$ and TiO_2 as the starting materials, weighing them out in accordance with the formula $BaTi_{1.005}O_3$. The small excess of titanium is necessary to ensure good sintering properties. The powders were mixed in a dry state, presintered for 10 hours at 1100 °C and, after dry-milling, were sintered for 4 hours at 1315 °C in oxygen at a pressure of 1 atmosphere. For $SrTiO_3$ ceramic the starting material was commercially available $SrTiO_3$ powder containing an excess of 1.6 mol% TiO_2. This powder was carefully dry-milled and then sintered for 8 hours at 1340 °C in O_2 at a pressure of 1 bar.

After conventional sintering we obtained densities of between 5.83 and 5.86 g/cm³ for $BaTiO_3$ ceramic and between 5.02 and 5.05 g/cm³ for $SrTiO_3$ ceramic. Assuming a perfect crystal lattice, we have calculated from the lattice parameters that the theoretical density of $BaTiO_3$ is 5.996 g/cm³ and that of $SrTiO_3$ is 5.117 g/cm³. In conventionally sintered samples of $BaTiO_3$ and $SrTiO_3$ ceramics this corresponds to a residual porosity of approximately 2.5% and 1.5% respectively, which is confirmed in the photomicrographs (*fig. 6*). Our method of isostatic hot pressing has been carried out at temperatures of 1100, 1150, 1200, 1250 and 1300 °C at a nitrogen pressure of 25, 50, 100, 150 and 200 bars. In order to follow the densification process we have measured the linear shrinkages of a prism-shaped sample (5×5×17 mm) in its longest direction. The change in density is then calculated from the equation:

$$d_i/d_f = 1 - 3s + 3s^2 - s^3,$$

where d_i and d_f are respectively the initial and final densities. Since the initial densities amounted to 95% of the theoretical values and the shrinkage was of the order of 0.01, the second- and third-order terms in the above equation could be ignored.

Fig. 7 and *8* clearly show how the density of $BaTiO_3$ and $SrTiO_3$ ceramic samples asymptotically approaches the theoretical density, as a function of time, temperature and pressure. The samples have become

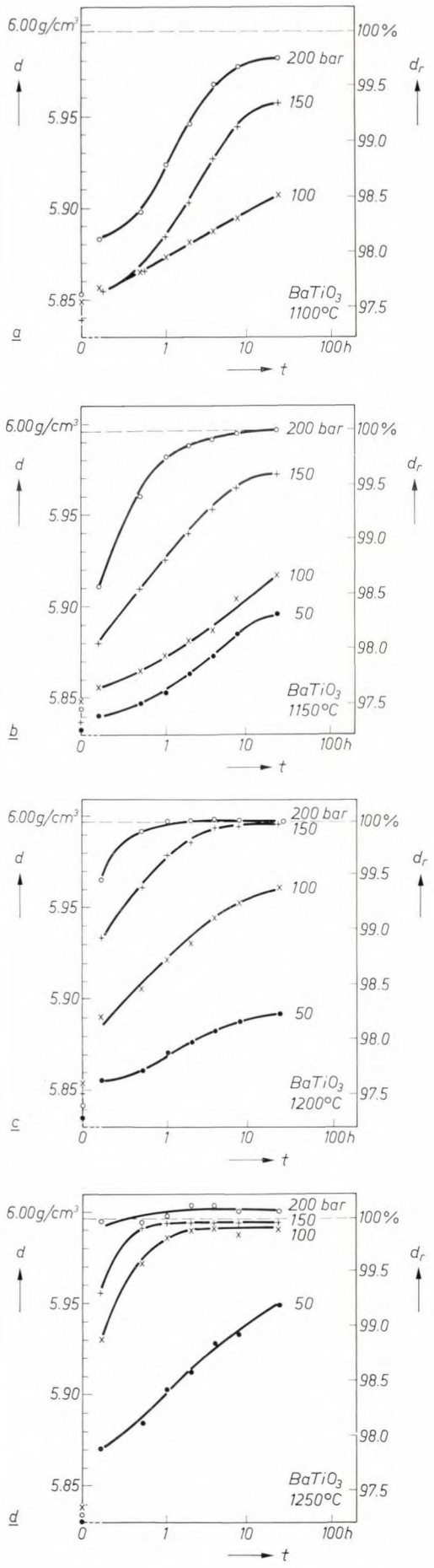


Fig. 7. Shrinkage behaviour (density d and relative density d_r plotted against time t) of conventionally sintered $BaTiO_3$ which has been hot pressed by our method at 1100 °C (*a*), 1150 °C (*b*), 1200 °C (*c*) and 1250 °C (*d*) in N_2 with a pressure which varies from 50 to 200 bars. The samples have been periodically cooled in order to carry out shrinkage measurements. The theoretical density ($d_r = 100\%$) has been indicated by a dashed line; the points plotted on the ordinate are the initial densities (after conventional sintering).

completely dense at temperatures 100 °C below the conventional sintering temperature and at a pressure of 200 bars. If the temperature during hot pressing approaches the conventional sintering temperature, complete densification can be achieved after only ten minutes.

The densities which are measured after 24 hours of hot pressing agree well with those determined by shrinkage measurements. Photomicrographs show that samples which have attained the theoretical density asymptotically have become pore-free. This may be seen in *fig. 9*, for example, in which the same sample as in fig. 6 is shown, but after isostatic hot pressing for 16 hours at 1200 °C in N_2 at 200 bars.

In general we can say that at the pressure under consideration, doubling the pressure has approximately the same effect as increasing the temperature by 50 °C. It is interesting to note that the increase in density varies logarithmically as a function of time at low pressures and temperatures.

Fig. 9. Photomicrograph of the $BaTiO_3$ ceramic of fig. 6, after hot pressing for 16 hours at 1200 °C in a nitrogen pressure of 200 bars. Magnification 150×.

[9] Al_2O_3 ceramic has also been studied by E. A. Bush [5]. He had to use considerably higher values of the process parameters (1650 °C, 1000 bars and 1 hour) than we did in order to obtain a dense Al_2O_3 ceramic. We think that the fact that we were able to use much lower values is connected with our addition of a small quantity of MgO to the Al_2O_3 (the defects thus introduced increase the sintering velocity).

[10] 'Cermets' are combinations of ceramics and metals.

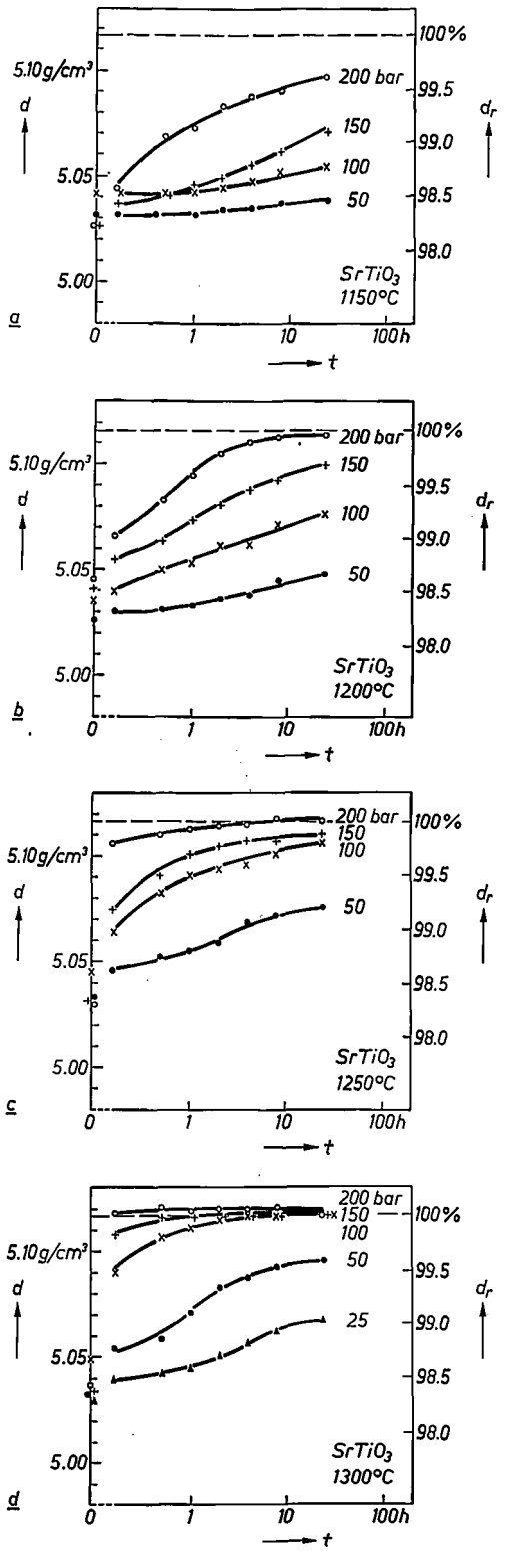


Fig. 8. Shrinkage behaviour of conventionally sintered $SrTiO_3$ ceramic which has been hot pressed at 1150 °C (*a*), 1200 °C (*b*), 1250 °C (*c*) and 1300 °C (*d*) in N_2 at a pressure which varies from 25 to 200 bars.

Other materials

Our method of isostatic hot pressing can be successfully applied to other ceramic materials. We have used commercially available ceramic materials such as Al_2O_3 [9], $Y_3Fe_5O_{12}$ (YIG), Mn-Zn ferrite and Ni-Zn ferrite. The results are presented in *Table II*. S. E. Isaksson has also reported good results with 'hard metals' such as WC and TaC (with Co, Ni or Fe as binder), with 'cermets' [10] based on Al_2O_3-Ag and with $MoSi_2$ and UO_2 [5]. Difficulties or failures with the application of our method have only arisen with coarse-grained $BaTiO_3$ and with $Pb(Zr,Ti)O_3$ ceramic in which traces of a liquid second phase appeared at the sintering temperature. Studies are in progress to obtain an insight into the nature of these difficulties.

The gas atmosphere at the first sintering step

As stated earlier, the first sintering step which is used to obtain closed pores should preferably be carried out in oxygen and not in air. We came to this conclusion when we tried to find out whether the diffusion velocity of nitrogen is indeed much lower than that of oxygen. We subjected two samples of the same PLZT ceramic to similar treatments, the only difference being that the first sintering step in one experiment took place in nitrogen and in the other experiment in oxygen. We assumed that in both cases the residual pores would contain the gas from the atmosphere in which the sintering occurred. The gas pressure in such pores enclosed by grains is not known with certainty, but probably lies somewhere between 1 and 10 bars; this pressure is approximately equal to $2\gamma/r$, where γ is the surface tension and r is the radius of the pores. During isostatic hot pressing these pores

Table II. Experimental data obtained with our method of isostatic hot pressing, carried out with non-perovskite material.

Composition	Pressure (bars)	Time (hrs)	Temperature (°C)	Initial density (g/cm³)	Final density (g/cm³)	Theoretical density (g/cm³)
Al_2O_3 + 0.3 wt% MgO	200	22	1300	3.93	3.98	3.99
$Y_3Fe_5O_{12}$ (garnet)	200	4	1300	4.82	5.16	5.17
Ni-Zn-ferrite	200	4	1200	5.31	5.33	5.33
Mn-Zn-ferrite	200	4.5	1200	4.94	5.10	5.12

will shrink to such a small volume that the internal pressure will become equal to the sum of the external pressure and the pressure due to surface tension. It will make little difference to the shrinkage behaviour of these samples whether they are sintered in N_2 or O_2. However, if the samples are heated again for 2 hours at atmospheric pressure to the sintering temperature, we may expect that, in the first case, the pores will swell, resulting in a decrease in density, whilst in the second case the oxygen will just escape because of its higher diffusion velocity and no swelling of the pores will occur. As *Table III* shows, the experimental results completely agree with these expectations.

Table III. Shrinkage and expansion behaviour of samples of PLZT ($x = 0.65, y = 0.11$) where the first sintering process has been carried out in either an oxygen or a nitrogen atmosphere.

d_i initial density;
d_t density after two hours of hot pressing at 1170 °C in argon at a pressure of 200 bars;
Δl shrinkage during the hot pressing. The original length was 16.6 mm;
Δl_a increase in length after two hours annealing at 1170 °C in air at a pressure of 1 bar;
d_a density after annealing.

	d_i (g/cm³)	d_t (g/cm³)	Δl (μm)	Δl_a (μm)	d_a (g/cm³)
In O_2 (4 hrs, 1170 °C)	7.60	7.77	116	0	7.77
In N_2, with a partial pressure of oxygen of 10^{-4} bar (4 hrs, 1170 °C)	7.52	7.76	179	47	7.67

For certain applications, such as envelopes for lamps, which reach a high temperature, it goes without saying that such a swelling of the pores must be avoided. In these cases the first sintering step must be performed in a gas such as oxygen, and in any case not in air. The first sintering step for Al_2O_3 used for the envelope of sodium lamps is performed in hydrogen, which also diffuses very rapidly.

The advantages of our method of isostatic hot pressing in comparison with earlier methods can be summarized as follows. The technique is simpler and appears to lend itself to mass production. Because no die or metallic encapsulating layer is necessary, there are no problems of corrosive reactions between the materials from which these are constructed and the ceramic. The ceramic samples can easily be pressed into any desired shape. Comparatively low pressures (50-200 bars) and temperatures (100-200 °C below the conventional sintering temperatures) are sufficient.

Summary. In earlier methods for isostatic hot pressing in the manufacture of ceramics a metallic layer has always been used in order to avoid the penetration of the working gas into the starting material. In the new process described here the starting material in powder form is first sintered at atmospheric pressure to a ceramic body with closed pores, after which hot pressing without the encapsulating metallic layer can be carried out. Suitable working gases are nitrogen or argon, both of which have a sufficiently low diffusion velocity. The sintering at atmospheric pressure should preferably be performed in oxygen. The new method is especially simple to carry out and is suitable for the mass production of ceramic components of any desired shape. The process requires a relatively low pressure (50-200 bars) and temperature (approximately 100 °C below the conventional sintering temperature). Because no die or encapsulating layer is used, problems relating to the contact between the ceramic and the materials from which the die or encapsulating layer are made do not arise. When these methods are used, the ceramic materials Pb(Zr,Ti)O_3, (Pb,La)(Zr,Ti)O_3, $BaTiO_3$, $SrTiO_3$, Al_2O_3, $Y_3Fe_5O_{12}$, Mn-Zn ferrite and Ni-Zn ferrite can be densified to almost the theoretical density.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or cooperate 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, 3 avenue Descartes, 94450 Limeil-Brévannes, 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*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.) *N*

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. Bäuerle: Optical phonons in small crystals of ABO_3 perovskites.
Phys. Stat. sol. (b) **63**, 177-182, 1974 (No. 1). *A*

J. W. M. Biesterbos: Structure and dc electrical properties of a Au-Rh-glass thick-film system.
J. appl. Phys. **45**, 153-160, 1974 (No. 1). *E*

S. J. Bless: Drop transfer in short-circuit welding.
J. Physics D **7**, 526-539, 1974 (No. 4). *E*

P. Blood, G. Dearnaley (AERE, Harwell) **& M. A. Wilkins** (AERE, Harwell): The depth distribution of phosphorus ions implanted into silicon crystals.
Radiation Effects **21**, 245-251, 1974 (No. 4). (*Also published in* Ion implantation in semiconductors and other materials, editor B. L. Crowder, Plenum, New York 1974, pp. 75-85.) *M*

P. F. Bongers & G. W. Rathenau: E. W. Gorter (1912-1972). Aspects of some fields of magnetism to which he has contributed.
Int. J. Magnetism **4**, 282-289, 1973 (No. 4). *E*

J. Borne, R. Goudin, G. J. Naaijer & J. Ratovonar: Console de visualisation et de dialogue pour enseignement assisté par ordinateur.
Acta Electronica **17**, 329-342, 1974 (No. 3). *L*

H. Bouma & A. H. de Voogd (Institute for Perception Research, Eindhoven): On the control of eye saccades in reading.
Vision Res. **14**, 273-284, 1974 (No. 4).

P. Branquart, J. P. Cardinael & J. Lewi: Optimized translation process, application to ALGOL 68.
Proc. Int. Computing Symp. 1973, Davos, pp. 101-107; 1974. *B*

J. W. Broer: Comment on the paper 'Systematic techniques for solving nonlinear communications'.
IEEE Trans. **PC-16**, 215, 1973 (No. 4). *E*

J. Brokken-Zijp & H. van de Bogaert: CIDNP evidence for radical intermediates in the thermal decomposition of Z-alkylaryldiazo sulphides.
Tetrahedron Letters 1974, 249-252 (No. 3). *E*

H. H. Brongersma: Surface structure analysis by ion scattering.
J. Vac. Sci. Technol. **11**, 231-235, 1974 (No. 1). *E*

E. Bruninx: The accurate determination of Ga in Ga_xSe_y by means of neutron activation: a comparison with X-ray fluorescence and wet chemical methods.
J. radioanal. Chem. **19**, 47-53, 1974 (No. 1). *E*

K. H. J. Buschow, A. M. van Diepen & H. W. de Wijn (State University of Utrecht): Moment reduction in magnetically ordered samarium intermetallics.
Phys. Rev. B **8**, 5134-5138, 1973 (No. 11). *E*

V. Chalmeton: Intensificateur à galette de microcanaux pour neutronographie.
La neutrographie, publ. Kodak-Pathé, Paris 1973, pp. 48-50. *L*

E. Cuppens & C. Ducot: Future education and new technical means.
Acta Electronica **17**, 241-270, 1974 (No. 3). (*Also in French.*) *E, L*

N. H. Dekkers: A universal deflection unit generating a field of any order and azimuth.
J. Physics D **7**, 805-814, 1974 (No. 6). *E*

J. P. Deschamps: Asynchronous automata and asynchronous languages.
Information and Control **24**, 122-143, 1974 (No. 2). *B*

P. A. Devijver: On a new class of bounds on Bayes risk in multihypothesis pattern recognition.
IEEE Trans. **C-23**, 70-80, 1974 (No. 1). *B*

H. J. A. van Dijk, C. M. G. Jochem, G. J. Scholl & P. van der Werf: Diameter control of LEC grown GaP crystals.
J. Crystal Growth **21**, 310-312, 1974 (No. 2). *E*

J. Donjon, G. Marie, B. Monod & R. Le Pape: Utilisation du tube TITUS pour la projection d'images de télévision sur grand écran.
Onde électr. **54**, 121-127, 1974 (No. 3). *L*

C. Z. van Doorn & J. L. A. M. Heldens: Angular dependent optical transmission of twisted nematic liquid-crystal layers.
Physics Letters **47A**, 135-136, 1974 (No. 2). *E*

D. den Engelsen & G. Frens: Ellipsometric investigation of black soap films.
J. Chem. Soc., Faraday Trans. I **70**, 237-248, 1974 (No. 2). *E*

L. J. M. Esser: The peristaltic charge coupled device.
Proc. Charge Coupled Device Appl. Conf., San Diego 1973, pp. 269-277. *E*

L. J. M. Esser: The peristaltic charge-coupled device for high-speed charge transfer.
1974 IEEE Int. Solid-State Circuits Conf. Digest tech. Papers, pp. 28-29 & 219. *E*

E. Fabre & R. N. Bhargava: Thermally stimulated current measurements and their correlation with efficiency and degradation in GaP LED's.
Appl. Phys. Letters **24**, 322-324, 1974 (No. 7). *L, N*

A. J. Fox: The grating guide — a component for integrated optics.
Proc. IEEE **62**, 644-645, 1974 (No. 5). *M*

G. Goutelle & M. Martin: Dispositifs d'apprentissage pour processus opératoires linéaires.
Acta Electronica **17**, 283-296, 1974 (No. 3). *L*

W. S. C. Gurney & J. W. Orton: New techniques for the study of Gunn diode contacts.
Solid-State Electronics **17**, 743-750, 1974 (No. 7). *M*

J. Haisma, G. Bartels & W. Tolksdorf: Preparation of the magnetic-bubble materials $Y_{3-x}Sm_xFe_{5-y}Ga_yO_{12}$ and $Y_{3-x}La_xFe_{5-y}Ga_yO_{12}$ by liquid-phase epitaxy, and their relevant physical properties.
Philips Res. Repts. **29**, 493-516, 1974 (No. 6). *E, H*

J. Haisma, H. J. Prins & K. L. L. van Mierloo: Magneto-optic determination of the refractive indices of ferric borate.
J. Physics D **7**, 162-168, 1974 (No. 1). *E*

J. C. M. Henning & J. H. den Boef: Mechanisms of flexural-strain-modulated electron spin resonance.
Solid State Comm. **14**, 993-996, 1974 (No. 10). *E*

F. Heutink: Widerange variable-pulse generator.
Electronics Letters **10**, 57-58, 1974 (No. 5). *E*

E. E. Hoefer, H. Grimmert & B. Kieslich: Computer-controlled synthesis of tomograms by means of a TV storage tube.
IEEE Trans. **BME-21**, 243-244, 1974 (No. 3). *H*

W. K. Hofker, H. W. Werner, D. P. Oosthoek & H. A. M. de Grefte: Experimental analysis of concentration profiles of boron implanted in silicon.
Ion implantation in semiconductors and other materials, editor B. L. Crowder, Plenum, New York 1974, pp. 133-145. *E*

B. B. van Iperen: Efficiency limitation by transverse instability in Si IMPATT diodes.
Proc. IEEE **62**, 284-285, 1974 (No. 2). *E*

R. N. Jackson & K. E. Johnson: Gas discharge displays: a critical review.
Adv. in Electronics & Electron Phys. **35**, 191-267, 1974. *M*

W. H. de Jeu, Th. W. Lathouwers & P. Bordewijk (University of Leiden): Dielectric properties of di-*n*-heptyl azoxybenzene in the nematic and in the smectic-*A* phases.
Phys. Rev. Letters **32**, 40-43, 1974 (No. 2). *E*

H. Klamet: Computer-aided design of the layout of integrated circuits (CADLIC).
Proc. Int. Computing Symp. 1973, Davos, pp. 451-458; 1974. *H*

A. J. R. de Kock, P. J. Severin & P. J. Roksnoer: On the relation between growth striations and resistivity variations in silicon crystals.
Phys. Stat. sol. (a) **22**, 163-166, 1974 (No. 1). *E*

G. Krekow & J. Schramm: Electrostatic gray-scale facsimile recording.
IEEE Trans. **ED-21**, 189-192, 1974 (No. 3). *H*

F. A. Kuijpers & B. O. Loopstra (Reactor Centrum Nederland, Petten): A neutron-diffraction study on the structural relationships of RCo_5 hydrides.
J. Phys. Chem. Solids **35**, 301-306, 1974 (No. 3). *E*

A. van Lamsweerde: Deadlock prevention in real time systems.
Proc. Int. Computing Symp. 1973, Davos, pp. 135-142; 1974. *B*

S. Leblanc: Une expérience de recyclage dans l'industrie au moyen de l'enseignement assisté par ordinateur.
Acta Electronica **17**, 297-318, 1974 (No. 3). *L*

D. B. Lee (Mullard Southampton Works): The push-out effect in silicon *n-p-n* diffused transistors.
Thesis, Southampton 1973. (Philips Res. Repts. Suppl. 1974, No. 5.)

J. Lemmrich: Winkelregelungen von Gleichstrommaschinen mit frequenzanaloger Signalverarbeitung.
Thesis, Darmstadt 1974. (Philips Res. Repts. Suppl. 1974. No. 4.) *H*

J. Lohstroh: The JFET as a photosensitive cell in image sensor arrays.
1974 IEEE Int. Solid-State Circuits Conf. Digest tech. Papers, pp. 34-35 & 221. *E*

M. H. van Maaren, H. J. van Daal, K. H. J. Buschow & C. J. Schinkel (University of Amsterdam): High electronic specific heat of some cubic UX_3 intermetallic compounds.
Solid State Comm. **14**, 145-147, 1974 (No. 2). *E*

J. Magarshack: On synchronizing signatures and oscillator optimization.
Microwave J. **17**, No. 3, pp. 70 & 87, March 1974. *L*

H. H. van Mal, K. H. J. Buschow & A. R. Miedema: Hydrogen absorption in $LaNi_5$ and related compounds: experimental observations and their explanation.
J. less-common Met. **35**, 65-76, 1974 (No. 1). *E*

G. Marie: Large area display materials and devices.
Electronic materials, editors N. B. Hannay & U. Colombo, Plenum Press, New York 1973, pp. 317-369. *L*

D. Meyer-Ebrecht: Schnelle Amplitudenregelung harmonischer Oszillatoren.
Thesis, Braunschweig 1974. (Philips Res. Repts. Suppl. 1974, No. 6.) *H*

A. R. Miedema: Superconductivity in transition metal alloys: an alternative to the rigid band model II.
J. Physics F **4**, 120-135, 1974 (No. 1). *E*

J. Monin (Conservatoire National des Arts et Métiers, Paris) **& G.-A. Boutry:** Optical and photoelectric properties of alkali metals.
Phys. Rev. B **9**, 1309-1327, 1974 (No. 4). *L*

K. Mouthaan & R. M. Snoeren: Characterisation of avalanche photodiodes having nearly-unilateral-carrier multiplication.
Electronics Letters **10**, 118-120, 1974 (No. 8). *E*

J. M. van Nieuwland & H. L. Hagedoorn (Eindhoven University of Technology): On the effect of very small irregularities at the pole faces of a cyclotron magnet on the quality of the ion beam.
Nucl. Instr. Meth. **116**, 13-16, 1974 (No. 1). *E*

J. M. van Nieuwland & N. Hazewindus: Some aspects of the design of a cyclotron central region.
Philips Res. Repts. **29**, 528-559, 1974 (No. 6). *E*

T. G. J. van Oirschot: Liquid-phase epitaxial growth of (AlGa)As on polished and roughened GaP substrates for transmission photocathodes.
Appl. Phys. Letters **24**, 211-213, 1974 (No. 5). *E*

C. van Opdorp, R. C. Peters & M. Klerk: Use of Schottky-diode collectors for SEM determination of bulk diffusion lengths.
Appl. Phys. Letters **24**, 125-126, 1974 (No. 3). *E*

E. M. van der Ouderaa: Structure and application of computer program ICAN: integrated circuit ac analysis.
AGARD Conf. Proc. No. 130, 20.1-20.15, 1973. *E*

R. I. Pedroso & G. A. Domoto (Columbia University, New York): Technical note on planar solidification with fixed wall temperature and variable thermal properties.
Trans. ASME C (J. Heat Transfer) **95**, 553-555, 1973 (No. 4). *N*

J. G. J. Peelen & R. Metselaar: Light scattering by pores in polycrystalline materials: transmission properties of alumina.
J. appl. Phys. **45**, 216-220, 1974 (No. 1). *E*

G. Piétri: Progress in PMT for high energy physics instrumentation.
1973 Int. Conf. on Instrumentation for high energy physics, Frascati, pp. 586-591. *L*

T. J. A. Popma, A. M. van Diepen & P. F. Bongers: Magnetic properties of the garnets
$$Bi_{0.8}Ca_{2x}Y_{2.2-2x}Fe_{5-x}V_xO_{12}.$$
J. Phys. Chem. Solids **35**, 201-205, 1974 (No. 2). *E*

W. Puschert: Optical detection of amplitude and phase of mechanical displacements in the angstrom range.
Optics Comm. **10**, 357-361, 1974 (No. 4). *A*

E. D. Roberts: A modified methacrylate positive electron resist.
Appl. Polymer Symp. No. 23, 87-98, 1974. *M*

P. Schagen: Image tubes with channel electron multiplication.
Adv. in Image Pickup and Display **1**, 1-69, 1974. *M*

C. Schiller & D. Bois: Observation des défauts dans les semiconducteurs par microscopie à balayage en cathodoluminescence.
Rev. Phys. appl. **9**, 361-371, 1974 (No. 2). *L*

E. Schröder: Korrektur der Farbfehler doppelbrechender Prismen in digitalen Laserstrahlablenkern.
Optik **39**, 489-498, 1974 (No. 5). *H*

A. M. J. H. Seuter (Philips Lighting Division, Eindhoven): Defect chemistry and electrical transport properties of barium titanate.
Thesis, Twente 1974. (Philips Res. Repts. Suppl. 1974, No. 3.)

P. J. Severin: The influence of the phase shift on thickness measurements of silicon epitaxial layers with a Fourier transform spectrometer.
J. Electrochem. Soc. **121**, 150-153, 1974 (No. 1). *E*

J. G. Siekman & P. J. van Vugt (Philips Centre for Factory Technologies, Eindhoven): Precision electron beam welding of vacuum chambers applied to a titanium large area proportional counter.
DVS-Berichte No. 28, 105-108, 1973.

G. Simpson: The dielectric constants of a ferroelectric ceramic.
Ferroelectrics **6**, 283-288, 1974 (No. 3/4). *M*

A. J. Smets: The fine solar sensor of the Astronomical Netherlands Satellite.
Proc. 1st European Electro-Optics Markets and Technology Conf., Geneva 1972, pp. 57-63; 1973. *E*

M. Snelling: General context-free parsing in time n^2.
Proc. Int. Computing Symp. 1973, Davos, pp. 19-24; 1974. *B*

J. L. Sommerdijk, A. Bril & A. W. de Jager: Two photon luminescence with ultraviolet excitation of trivalent praseodymium.
J. Luminescence **8**, 341-343, 1974 (No. 4). *E*

J. L. Sommerdijk, J. M. P. J. Verstegen (Philips Lighting Division, Eindhoven) **& A. Bril:** Luminescence of $6s^2$ ions in magnetoplumbite-like aluminates.
Philips Res. Repts. **29**, 517-527, 1974 (No. 6). *E*

F. A. Staas, K. Weiss & A. P. Severijns: Surface efficiency of various types of heat exchangers in an He^3-He^4 dilution refrigerator.
Cryogenics **14**, 253-263, 1974 (No. 5). *E*

W. T. Stacy, R. Metselaar, P. K. Larsen, A. Bril & J. A. Pistorius: Crystal defect mapping via impurity-activated luminescence in gadolinium gallium garnet.
Appl. Phys. Letters **24**, 254-256, 1974 (No. 6). *E*

W. T. Stacy, J. A. Pistorius & M. M. Janssen: Helical growth defects in gadolinium gallium garnet.
J. Crystal Growth **22**, 37-43, 1974 (No. 1). *E*

A. L. Stuijts: Synthesis of materials from powders by sintering.
Ann. Rev. Mat. Sci. **3**, 363-395, 1973. *E*

A. Thayse: Differential calculus for functions from $(GF(p))^n$ into $GF(p)$.
Philips Res. Repts. **29**, 560-586, 1974 (No. 6). *B*

G. E. Thomas & E. E. de Kluizenaar: SIMS with a standard quadrupole residual gas analyser.
Rev. sci. Instr. **45**, 457-458, 1974 (No. 3). *E*

H. van Tongeren: Positive column of the Cs-Ar low-pressure discharge.
J. appl. Phys. **45**, 89-96, 1974 (No. 1). *E*

N. C. de Troye: Integrated injection logic — a new approach to LSI.
1974 IEEE Int. Solid-State Circuits Conf. Digest tech. Papers, pp. 12-13 & 214. *E*

J. D. B. Veldkamp & W. F. Knippenberg: Anomalous Young's modulus behaviour of SiC at elevated temperatures.
J. Physics D **7**, 407-411, 1974 (No. 3). *E*

L. Vriens: Two- and three-electron group models for low-pressure gas discharges.
J. appl. Phys. **45**, 1191-1195, 1974 (No. 3). *E*

H. W. Werner, H. A. M. de Grefte & J. van den Berg: Fingerprint spectra in secondary ion mass spectrometry.
Adv. in Mass Spectrom. **6**, 673-682, 1974. *E*

A. de Winter & H. Blok (Delft University of Technology): Fling-off cooling of gear teeth.
Trans. ASME B (J. Engng. Ind.) **96**, 60-70, 1974 (No. 1). *E*

H. Zijlstra: Eigenschaften und Anwendungen von Wasserstoff absorbierenden Verbindungen der Übergangsmetalle mit seltenen Erden.
Fette - Seifen - Anstrichmittel **76**, 49-52, 1974 (No. 2). *E*

Contents of Electronic Applications Bulletin **32**, No. 3, 1974:

L. E. Jansson: Converter circuits for switched-mode power supplies (pp. 83-92).
G. van Aller & Th. G. Schut: Combined image-intensifier and 'Plumbicon' tube for studio colour cameras (pp. 94-100).
J. Mulder: Input network design for a high-frequency wideband power amplifier (pp. 101-109).
A. A. Los, D. C. L. Vangheluwe & A. C. Smaal: Ultrasonic delay lines for chrominance signal decoding (pp. 111-120).
Nomogram for single-layer air-cored coils (p. 124).

Contents of Mullard Technical Communications **13**, No. **124**, 1974:

Titanium-gold: high-reliability transistor metallisation (pp. 142-152).
J. Merrett: Summary of factors affecting power supply choice and design (pp. 153-156).
W. Hetterscheid: Base circuit design for high-voltage switching transistors in power converters (pp. 157-169).
C. Barrow: Cores for switched-mode power supply transformers (pp. 170-181).
CL8960 X-band doppler radar module (pp. 182-184).

Contents of Valvo Berichte **19**, No. 1, 1974:

P. Bockelmann & W. Westendorf: Die elektronischen NF-Einsteller TCA 730 und TCA 740 (pp. 1-18).
F. Pötzl: VHF-Breitbandzirkulatoren in konzentrierter Bauweise (pp. 19-28).

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 4

Small electric motors III

In this last issue of our group of three on small electric motors [*] we give pride of place to an article on the early history of the electric motor, from Oersted's discovery of the movement of a magnetic needle in the neighbourhood of a current-carrying conductor until about the year 1900, by which time various types of electric motor were already in industrial production. This issue also includes a treatment of the optimum dimensioning of a small d.c. commutator motor; the results lead to dimensional ratios that are different from those usually encountered. Next, there is a theoretical treatment of the brushless d.c. motor; the author shows how to construct diagrams from which some of the important characteristics of the motor can be read off directly. The issue concludes with a comparison of the various types of motor with particular reference to running costs; by using figures of merit and their dependence on the dimensions it can be shown that in general terms some types are more suitable as small motors, whereas others are more suitable as motors of larger dimensions.

The early history of the electric motor

Brian Bowers

The modern electric motor is the product of a century and a half of development. In this article the early stages of that development are followed from the first ideas for producing motion by electric action up to the point where the main kinds of motor had been conceived and put into production.

The first section deals with the very early period, and the second describes some of the many 'electromagnetic engines' made between 1832 and 1860. The third section considers some of the ideas which lay behind these machines. The fourth section shows how the practical d.c. motor evolved not from the electromagnetic engine but from the electric generator. The final section is concerned with alternating-current machines, which had to be invented when electricity-supply companies adopted alternating-current systems.

Motion produced by electricity

A continuous electric current, as opposed to a spark or static charge, became available in 1800 as a result of the work of the Italian Alessandro Volta. Volta was following up the work of his fellow countryman, Galvani, who had shown in 1786 that the muscle in the leg of a dead frog could be made to convulse by touching it with two dissimilar metals. Galvani thought that the source of the action was in the muscle; Volta thought it was at the junction of the two dissimilar metals. Volta wrote a letter from Italy to Sir Joseph Banks, the President of the Royal Society of London, in which he announced his 'Pile' and his 'Crown of Cups'. Volta's pile consisted of a stack of pairs of silver and zinc discs, each pair being separated by paper or cloth spacers soaked in salt water. The crown of cups consisted of a number of cups arranged in a circle and each filled with salt water into which strips of copper and zinc were dipped. Volta's discovery aroused great interest, and in November 1801 he demonstrated it before an assembly of scientists in Paris. Napoleon Bonaparte was present, and he decided to establish prizes to encourage new research on 'galvanism', as it was called, and remarked 'Galvanism, in my opinion, will lead to great discoveries' [1] [**].

Among the scientists who studied galvanism, or current electricity, was the Dane H. C. Oersted. In July 1820 he published a four-page paper in Latin announ-

Brian Bowers, Ph.D., C.Eng., M.I.E.E., is Deputy Keeper of the Department of Electrical Engineering and Communications at the Science Museum, London.

[*] The previous issues were Volume 33, No. 8/9, 1973, and 34, No. 7, 1974.

[**] The references are listed at the end of the article.

cing his discovery that an electric current in a wire could cause a compass needle nearby to be deflected [2]. Oersted's discovery created great interest because it established for the first time a connection between electricity and magnetism. Its significance in the history of the electric motor is that it was the first production of mechanical motion by an electric current.

Oersted's discovery was studied further by the French physicist Arago. He had noticed that the oscillations of a magnetic needle were damped if a metal plate was nearby. Arago constructed a device in which a copper disc was mounted on a vertical spindle and rotated beneath a pivoted magnetic needle. He found that the needle also rotated, albeit at a lower speed, in the same direction as the disc. The device has since been known as 'Arago's disc'[3].

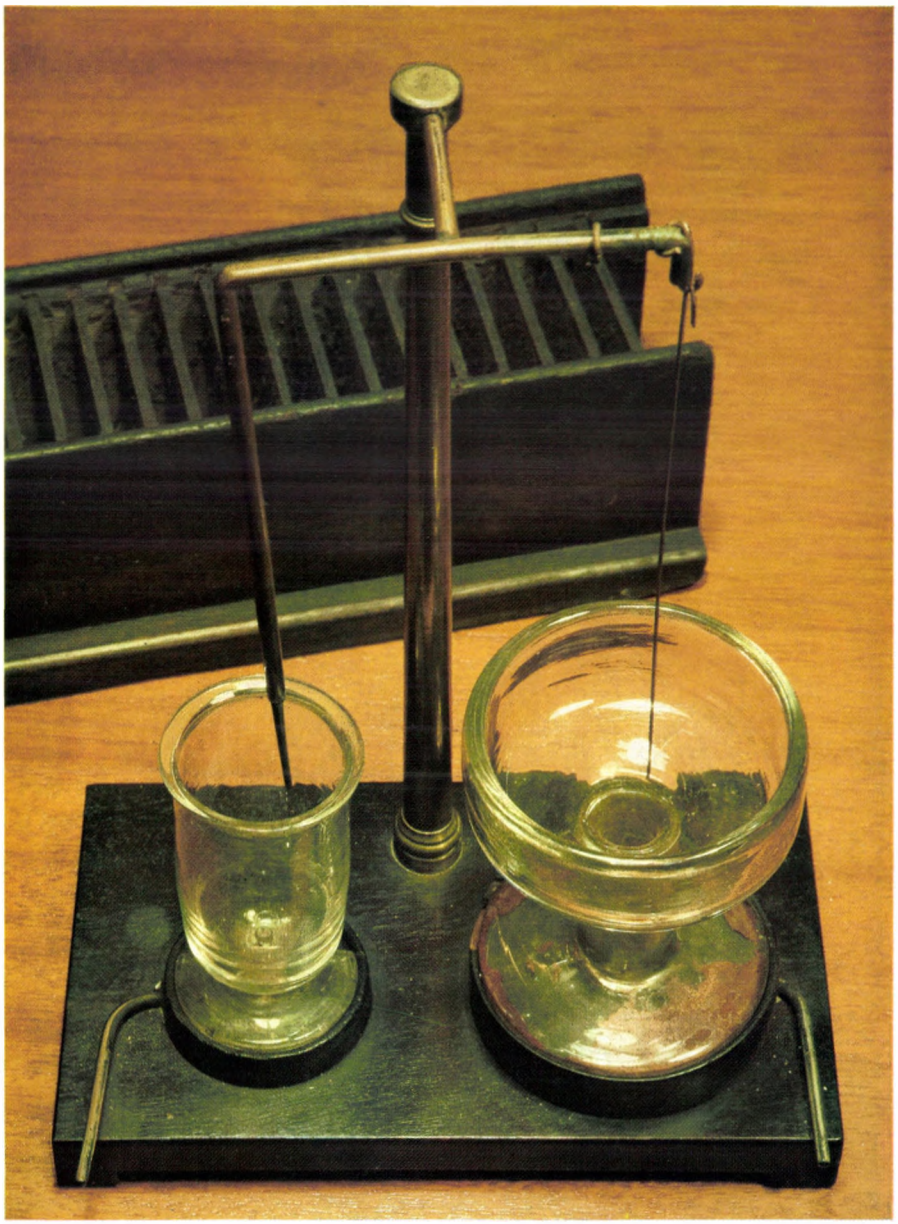


Fig. 1. The apparatus used by Faraday in 1821 to demonstrate 'electromagnetic rotation'. The two basins were filled with mercury; in the left-hand one a bar magnet floated vertically, with the lower end freely attached to the bottom of the basin. When a current flowed, the upper end of the magnet described circles around the fixed electrode dipping into the mercury. In the right-hand basin a bar magnet was fixed to the bottom. The loose wire dipped into the mercury and described circles around the upper end of the magnet when a current was passed. A wooden battery trough, divided into cells, can be seen behind the apparatus. (Photo Royal Institution, London.)

The English scientist Michael Faraday, then only the little-known assistant to Sir Humphry Davy, took up the

subject of electromagnetism in the summer of 1821. He had been invited to write an historical account of electromagnetism for the *Annals of Philosophy* edited by his friend Phillips. With his usual thoroughness Faraday repeated all the important experiments of other scientists while writing his account. As he studied the subject he became convinced that it ought to be possible to produce continuous circular movement by making use of the circular magnetic force around a wire carrying a current. In September 1821 he found how to do it; he made two devices, which were really the first electric motors, to show 'electromagnetic rotation'. In the first he fixed a bar magnet vertically in a basin with a blob of wax. The basin was then nearly filled with mercury and a wire with a cork on its end was loosely fixed to a point above the basin. A battery was connected between the first wire and another wire connected to the mercury, and the first wire moved in a circular path round the magnet. 'Very satisfactory', wrote Faraday in his notebook. In the second arrangement the magnet was fixed at the bottom of the basin but the upper end was free to move. The wire from above was fixed, dipping into the middle of the mercury. With this arrangement the upper end of the magnet moved in circles around the fixed wire. Faraday then had a special demonstration apparatus made which combined the rotating wire arrangement and the rotating magnet (*fig. 1*), and he published an article on electromagnetic rotations in October 1821. The apparatus was copied and the experiments repeated all over Europe [4].

Peter Barlow, who lectured at the Woolwich Academy, repeated Faraday's rotation experiment and then replaced the rotating wire with a star-shaped wheel rotating between the poles of a horseshoe magnet. The wheel dipped into mercury maintaining continuous contact and when a current flowed between the axis of the wheel and the mercury then the wheel rotated (*fig. 2*) [5].

Faraday's electromagnetic-rotation device and Barlow's wheel both demonstrated that it was possible to produce continuous motion by electrical means, but they had the fundamental limitation that only a single current-carrying conductor was passing through the magnetic field. Barlow's wheel was basically the same machine as Faraday's disc generator, although that was not made until 1831.

The critical advance necessary before electrical machines — motors or generators — could progress further was first made in a generator. This was the concept of making a coil of wire pass through the magnetic field, rather than a single conductor. In 1832 the French instrument maker Hippolyte Pixii made a generator (though he did not call it that) in which there was relative rotation between a permanent magnet and a coil of wire, and a crude, cam-operated reversing switch performed the function of a commutator and provided a direct-current output (*fig. 3*).

The way was then clear for the development of early forms of motor, or 'electromagnetic engines', and many were made during the 1830s. One suggestion from the 1820s is worth mentioning, however: in 1824 the *Mechanics' Magazine* described a device to show the 'mechanical Effects of Electricity' [6]. A light wheel with vanes cut out of paper was mounted on an axle beneath two point electrodes in such a way that the vanes could move in the space between the electrodes from one electrode to the other. When an electric discharge from a frictional electric machine or a Leyden jar was passed between the electrodes then the wheel turned, with the vanes moving from the positive to the negative electrode (*fig. 4*).

Electromagnetic engines

The first person to appreciate that electromagnetism might be used to provide mechanical power was probably the American Joseph Henry.

Electromagnets were studied by several people, one of the first being Sturgeon in England. Gerard Moll, Professor of Natural Philosophy in the University of Utrecht, was in England in 1828 and saw some of Sturgeon's experiments (*fig. 5*). Moll obtained an electromagnet and made experiments relating the weight supported by the electromagnet to the active area of zinc in the battery supplying the current. He

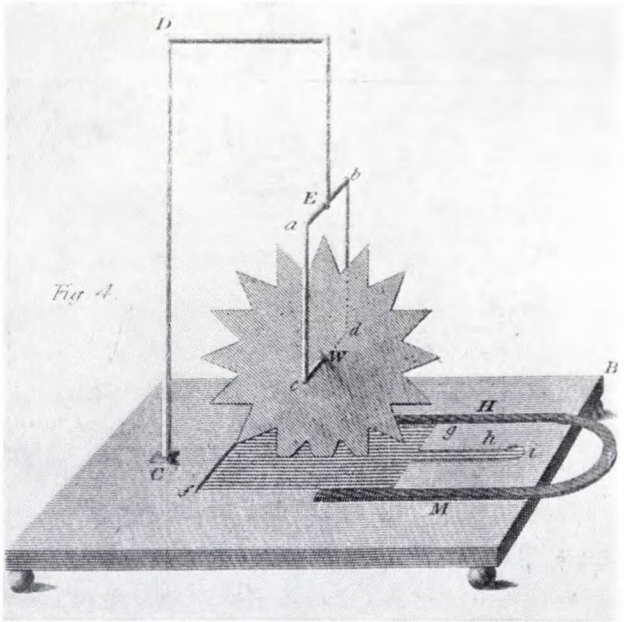


Fig. 2. Barlow's wheel. The teeth of the wheel *W* dip into mercury between the poles of a horseshoe magnet *HM*. When a current flows the wheel turns.

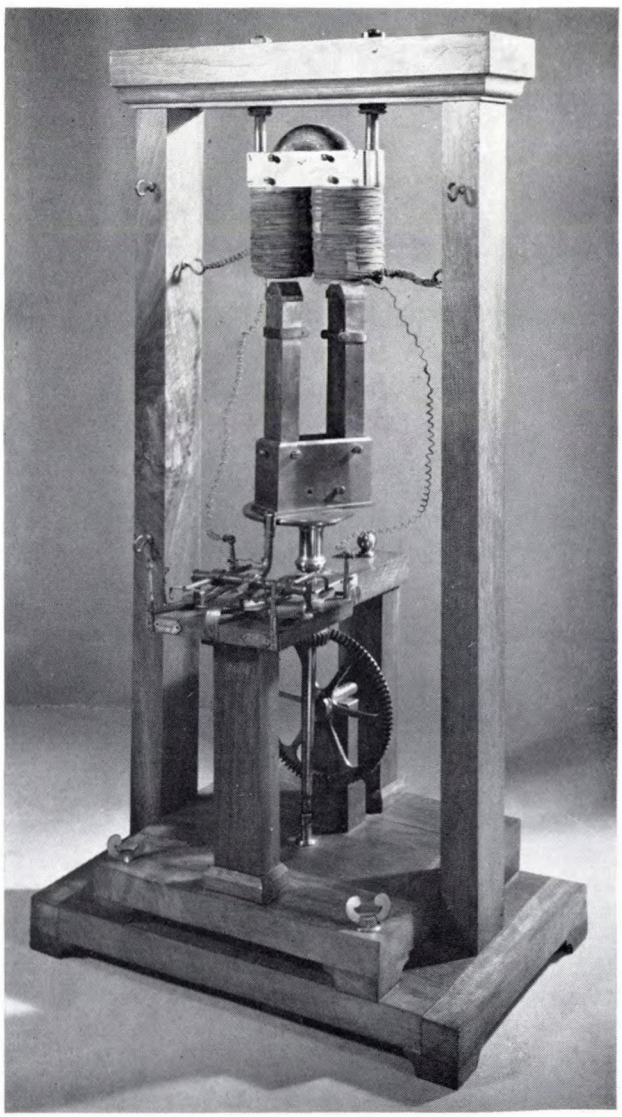
also tried to observe the speed with which the magnetism could be created and destroyed when the circuit was made and broken, and the rapidity with which the polarity could be reversed [7].

In 1830 Professor Salvatore del Negro, of Padua University in Italy, obtained rotary motion from an electric current by means of an electromagnet. A permanent magnet hanging from a pivot was attracted to the electromagnet and the electromagnet was energized through a contact which was broken when the permanent magnet moved from its rest position. In this way the permanent magnet was made to perform an oscillatory motion which was converted into rotary motion by a pawl and ratchet [8]. This has been called the first electric motor [9], but the title seems rather undeserved.

Henry made a reciprocating electric engine which he described in 1831 as follows [10]:

'I have lately succeeded in producing motion in a little machine by a power, which, I believe, has never before been applied in mechanics — by magnetic attraction and repulsion. Not much importance, however, is attached to the invention, since the article, in its present state, can only be considered a philosophical toy; although, in the progress of discovery and invention, it is not impossible that the same principle, or some modification of it on a more extended scale, may hereafter be applied to some useful purpose.'

This engine consisted of a straight electromagnet mounted horizontally and supported on knife-edges at its centre (*fig. 6*). A bar permanent magnet was placed vertically below each end of the electromagnet with the north poles facing in the same direction. The electromagnet coil was connected between two pairs of stiff wires, one pair at each end, and these wires dipped into



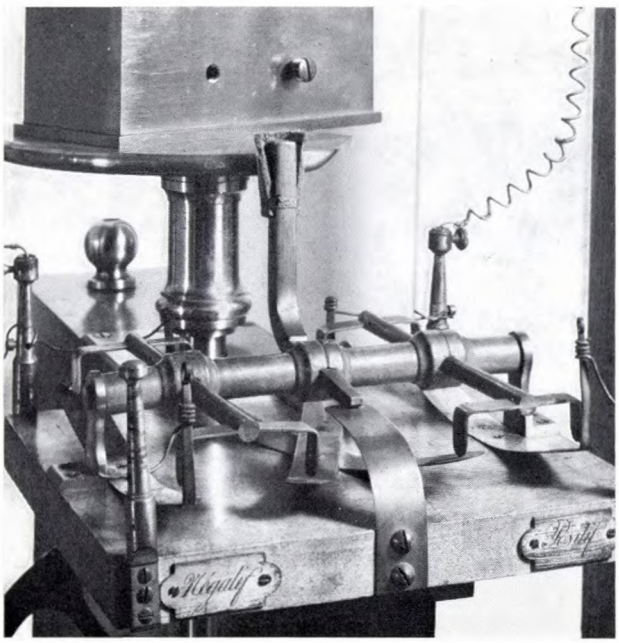


Fig. 3. Pixii's generator (1832). This was the first device in which a multi-turn coil was used to increase the electromagnetic effect. *Right:* The two bar magnets are rapidly rotated below the two coils on an iron core by the handwheel and gearing. *Below:* A cam operates a reversing switch, consisting of leaf springs, twice in each revolution, thus providing a simple commutator action. (Photos Deutsches Museum, Munich.)

MECHANICAL EFFECT OF ELECTRICITY.

The mechanical Effects of Electricity are exhibited in its power of impelling and dispersing light bodies ; of perforating, expanding, compressing, tearing, and breaking to pieces, all conducting substances through which it is sufficiently powerful to force its passage.

If a light wheel, having its vanes made of card paper, be made to turn freely upon a centre, it will be put in motion when it is presented to an electrified point. The wheel will always move from the electrified point, whether its electricity is positive or negative. In this experiment the current seems to be produced by the recession of the similarly electrified air in contact with the point, and therefore the circumstance of the wheel turning in the same direction when the electricity is negative, cannot as Mr. Singer has remarked, be considered as any proof of the existence of a double current of the electric fluid. As an illustration take the following experiment :—

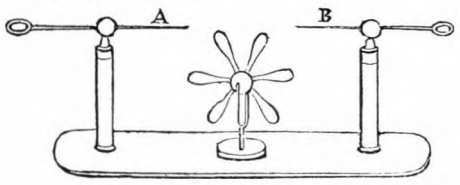


Fig. 4. Rotation of a paper wheel under the influence of an electric discharge. The vanes move from the positive to the negative electrode. (Mechanics' Magazine, 1824.)

Place upon an insulating stem a light wheel of card paper, properly suspended upon pivots, as represented in our Plate, and introduce it between the pointed wires (AB) of the universal discharger, placed exactly opposite to each other, and at the distance of little more than an inch from the upper vanes. Then having connected the wire A with the positive conductor, and the wire B with the negative conductor, of an electrical machine, the little wheel will revolve in the direction AB ; and if the wire B is connected with the positive end, and A with the negative end, the motion of the wheel will be from B to A. The transmission of a small charge through the wires, by an insulated jar, will produce the same effect.

The preceding experiment, imagined by Mr. Singer, is considered by him as a proof that there is only one electric fluid, and that it passes from the positive to the negative wire ; for, if there were two electric fluids, he concludes, "that the wheel being equally acted upon by each, will obey neither, and remain stationary."—*Chemist.*

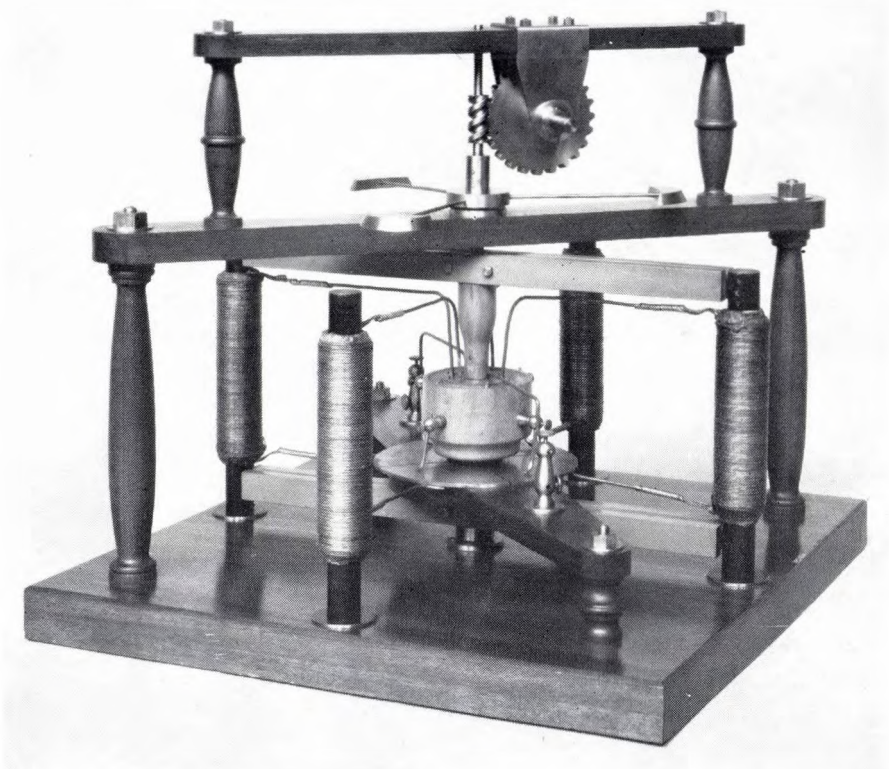


Fig. 5. Reconstruction of one of Sturgeon's motors. The rotor consists of two horizontal bar magnets, which rotate past the poles of four electromagnets. These are connected through a commutator, which consists of two concentric jars of mercury, four sliding contacts and a contact disc divided into four quadrants. (Crown Copyright. Science Museum, London.)

mercury cups when their end was down. A battery was connected to each pair of mercury cups. When one end of the bar was depressed the electromagnet was energized in such a way that the magnetic reactions between the poles of the electromagnet and those of the two permanent magnets caused an attraction at one end of the bar and a repulsion at the other so that the bar rocked in the other direction. This movement shifted the connections of the coil from one battery to the other and the direction of current in the coil was reversed.

The forces acting on the bar were then reversed and the bar returned to its original position. In this way a reciprocating motion was established.

Henry's dipping contacts performed the function of the commutator in a conventional motor. According to Henry's own account the horizontal electromagnet was seven inches (18 cm) long and wound with three parallel strands of 'copper bell-wire' each twenty-five feet (7.5 m) long. (By 'bell-wire' Henry meant wire used for pulling mechanical bells, not of course the modern cotton-covered 'bell-wire' used for electric bells.) The machine ran uniformly at about seventy-five vibrations per minute.

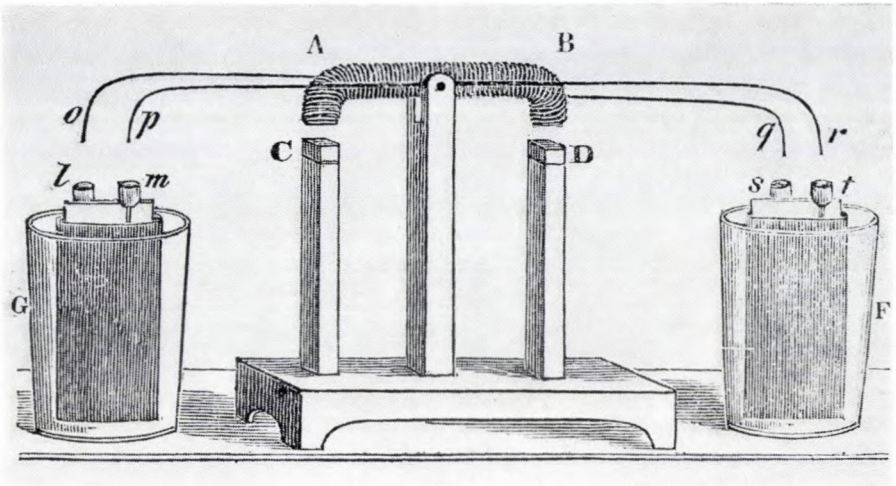


Fig. 6. Henry's reciprocating electric engine (1831). The reciprocating electromagnet *AB* is alternately connected to the batteries *F* and *G*. Attraction and repulsion between the electromagnet and the permanent magnets *C* and *D* maintain a continuous motion.

In 1833 Ritchie described a rotary motor with a commutator. The Rev. William Ritchie was Professor of Natural Philosophy at the Royal Institution. His motor consisted of a straight horizontal electromagnet pivoted about a vertical axis (*fig. 7*). The commutator was a circular groove in the baseboard concentric with the axis and filled with mercury to above the level of the baseboard. The groove was divided into two semicircular sections by partitions and the mercury in each section was connected to one pole of the battery. The ends of the electromagnet coil pointed downwards into the mercury. Initially Ritchie showed that this device would rotate in the Earth's magnetic field, but he found that it worked better when permanent field magnets were added, and he was able to raise a weight of several ounces (several tens of grams) [11].

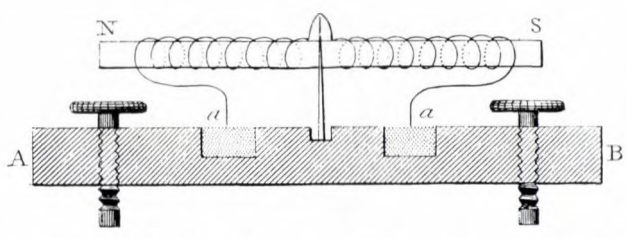


Fig. 7. Ritchie's rotary motor (1833). Current is supplied to the electromagnet *NS* through the mercury in two semicircular grooves *a* in the baseboard. The motor would rotate in the Earth's magnetic field.

In 1835 the London instrument maker Francis Watkins designed a motor consisting of a group of stationary coils facing a bar magnet mounted on a shaft. The shaft also carried contacts which were arranged to send a succession of pulses of current from a battery to the coils at the appropriate intervals. Watkins stated that his machine was based on the generator recently made by Saxton, which suggests that he may have recognized the interchangeability of motors and generators. If he did, he did not make it very clear, and the fact that motors and generators are essentially interchangeable was not generally appreciated until much later.

Watkins wrote a paper for the Royal Society of London in which he described Henry's reciprocating engine, which he said was the first true electromagnetic motor, and some of his own machines including one which incorporated a speed-reducing gear. He illustrated some of the ways in which motors might be used as a source of power by using his machine to drive small models of hammers, pumps and dredging machines [12].

Also in 1835 Sibrandus Stratingh and his assistant Christopher Becker, an instrument maker, made an electric vehicle which ran on a table until the battery supplying it was run down. Stratingh was a doctor of medicine and a leading scientist in Groningen, Holland,

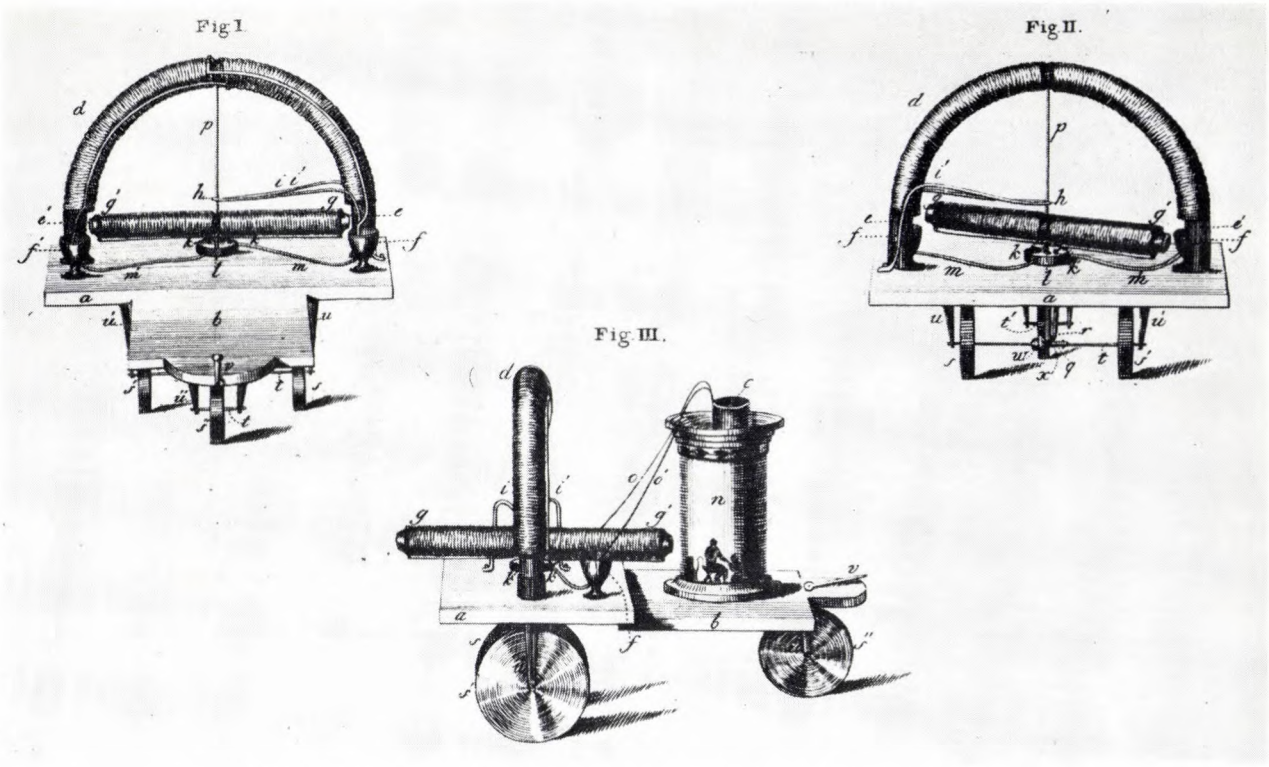


Fig. 8. Stratingh and Becker's electrically propelled model vehicle (1835). The horizontal bar magnet is the rotor; its motion is transferred to the wheels by gearing. The jar in *Fig. III* is a galvanic battery.

where in 1824 he became Professor of Chemistry at the University. In 1801 he is said to have driven in a steam car from Groningen to Zuidlaren, a few miles to the South. He read of the earlier work of Jacobi in Russia, and wanted to make an electric car. He never made a full-size car but is said to have had an electric boat in 1840 in which he could sail with his family.

An account of their work by Stratingh and Becker was published in Dutch in 1835, but is little known outside Holland. Their drawings (*fig. 8*) show a three-wheeled vehicle carrying a single voltaic cell and a rotating horizontal electromagnet with a commutator beneath it very like Ritchie's motor. The axis of the rotating magnet is geared to the running wheels, and the field excitation is provided by a semicircular electromagnet arranged in a vertical plane over the rotor [13].

The earliest patent for electromagnetic engines was obtained by Thomas Davenport. He had suggested using electromagnetism to produce motion in 1833, and he patented a machine in the United States of America in 1837 [14]. In the same year an English patent agent, Miles Berry, acting on behalf of Davenport, obtained a similar patent in England [15]. Davenport's patent specification gives a remarkably well developed description for the first patent in the field. The American Patent Office required a model, and that model is now in the Smithsonian Institution (*fig. 9*).

The rotor consists of four coils on a cruciform frame on a vertical shaft. Opposite coils are connected in series and there are four ends brought out below the coils as rudimentary brushes, which turn with the rotor. The fixed part of the commutator is made from two semicircular copper pieces, connected to the battery. The stator consists of two semicircular permanent magnets, with their like poles adjacent. The specification envisages the possibility of a four- (or more) pole stator, in which case the fixed part of the commutator requires correspondingly more segments. The stator magnets could be electromagnets rather than permanent ones.

Davenport was apparently a blacksmith [16] in Rutland, Vermont, though his patent agent just described him as a 'gentleman' [15]. He seems to have been the first person to use an electric motor to do useful work. In 1837 he employed a motor weighing fifty pounds (23 kg) and running at 450 revolutions per minute to drill holes a quarter of an inch (about 6 mm) in diameter in steel [17]. Davenport had great hopes for future electrical developments which he expressed thus [18]:

> 'I hope not to be considered an enthusiast, when I venture to predict that soon engines capable of propelling the largest machinery will be produced by the simple action of two galvanic magnets, and working with much less expense than steam!'

The two pioneers of electric motors whose names are best known today are probably Jacobi and Davidson.

The Russian professor M. H. Jacobi of St Petersburg obtained a grant from the Czar to enable him to conduct research into electric power, probably the first example of a government grant for electrical-engineering research. Although some of Jacobi's earlier work was apparently known to Stratingh in 1835, the first detailed account published in Western Europe was contained in a letter from Jacobi to Faraday which was published in 1839. In this letter Jacobi described his work with an electrically driven boat. The boat was a ten-oared shallop fitted with paddle wheels driven by an electromagnetic engine. He 'travelled for whole days', usually with ten or twelve people aboard, on the River Neva. Jacobi had a motor consisting of two wooden frames, each carrying a series of horseshoe electromagnets with pole pieces facing one another but spaced apart. The rotor moved in the space between the pole pieces. It had a six-legged spider with a straight permanent magnet on each leg. A commutating arrangement with four contact discs and four wipers periodically reversed the connections between the battery and the coils so that the rotor would move always in the same direction. When supplied through a battery of 128 Grove cells the vessel travelled at just over four kilometres per hour. Jacobi reported that the motor could provide 'a force of one horse' from a battery with twenty square feet (1.9 m^2) of platina [19].

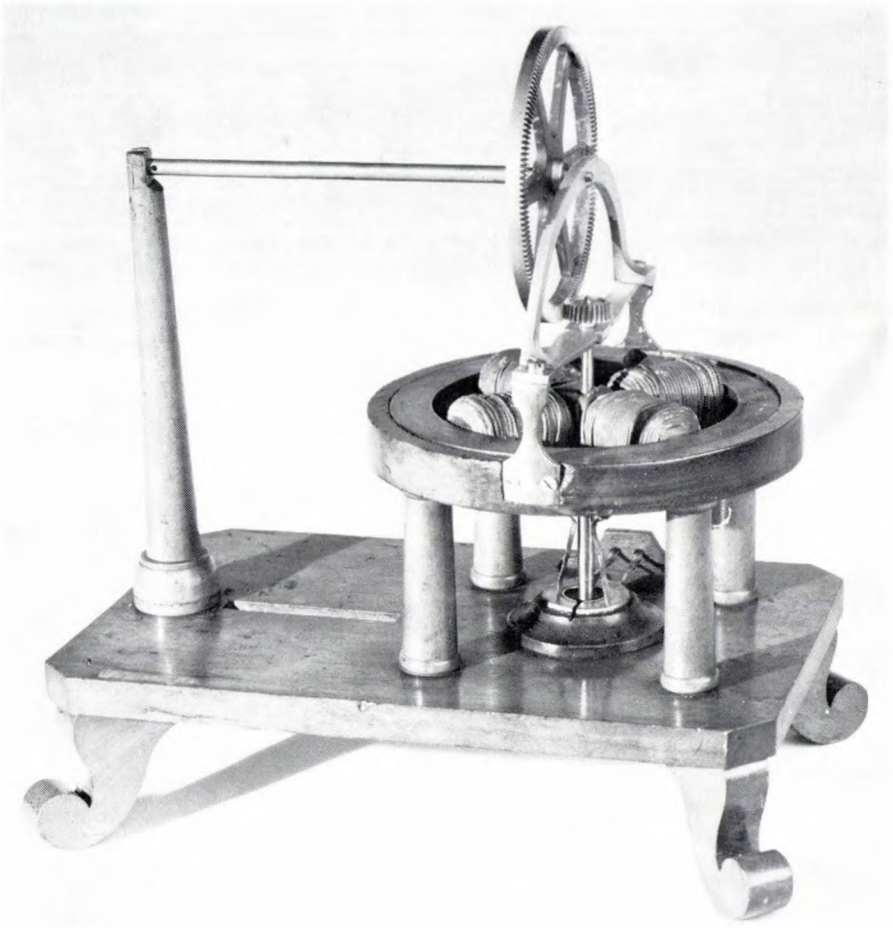


Fig. 9. Davenport's motor (1837). The rotor carries four coils on a cruciform frame, and the stator consists of two semicircular permanent magnets, with their like poles adjacent. The model was made for the American Patent Office. (Photo Smithsonian Institution, Washington.)

In a later motor Jacobi dispensed with permanent magnets and had two frames carrying straight electromagnets. One frame was fixed and the other rotated (*fig. 10*).

Robert Davidson was a Scotsman. The first account of his work was published in 1840 by Professor P. Forbes of King's College Aberdeen, who had seen an account of Jacobi's experiments in the Philosophical Magazine and wrote to say that 'A countryman of our own, Mr Robert Davidson of this place, has been eminently successful in his labors in the same field of

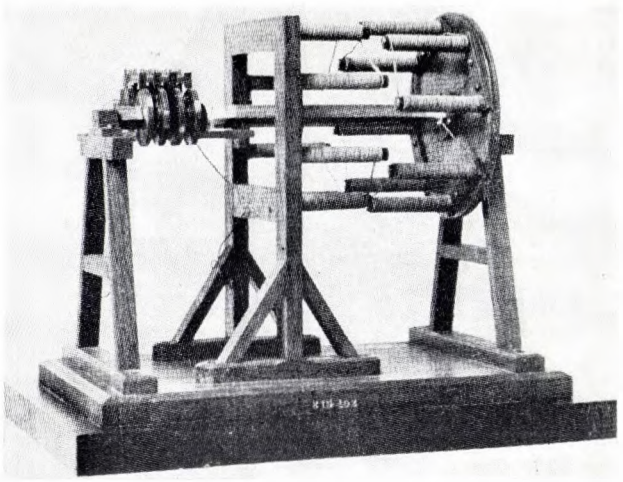


Fig. 10. Jacobi's motor. A motor rather similar to this one (but with permanent magnets) was used by Jacobi to propel a boat on the River Neva at St Petersburg. (Photo Science Museum, London.)

discovery'. With a machine which Forbes did not describe except to state that it had only two electromagnets, Davidson could drive a lathe and turn small articles using a battery with less than one square foot (9 square decimetres) of zinc surface. Another machine powered by the same battery would propel a small carriage with two people along the wooden floor of the room [20]. In the winter of 1841-42 the Royal Scottish Society of Arts gave Davidson financial help to continue his experiments. In September 1842 he tried an electrically driven carriage on the Edinburgh and Glasgow Railway; the four-wheeled carriage was sixteen feet (4.8 metres) long and weighed over 5 tons (*fig. 11*). The motors consisted of wooden cylinders on each axle with iron strips fixed in grooves in the cylindrical surface. Horse-shoe electromagnets on either side of the cylinder were energized alternately through a simple commutator on the axles. As shown in the drawing the batteries were arranged at each end of the carriage; they had a total of forty cells, each with a zinc plate between two iron plates just over one foot square (about 10 dm^2). The plates could be raised out of the wooden troughs containing the electrolyte by a simple windlass arrangement. These batteries proved insufficient and more were added on each side of the carriage, roughly doubling the power. The carriage then ran at about 4 miles per hour (about 6.5 km/h) on level track. Although the contemporary account says the experiments were carried out on the Edinburgh and Glasgow Railway, it does not say the carriage ran from Edinburgh to Glasgow. No indication is given of the actual distance travelled.

An odd feature of the construction was that the cores of the electromagnets were hollow. Each limb of the cores consisted of four iron plates arranged to form a box. According to a contemporary account this construction was adopted to save weight. Clearly Davidson did not appreciate that it was the total cross-section of iron which was important.

In the year of the great California gold rush, 1849, the United States Commissioner of Patents, Thomas Ewbank, included in his annual report some thoughts on the subject of electric motors [21]:

'. . . the belief is a growing one that electricity, in

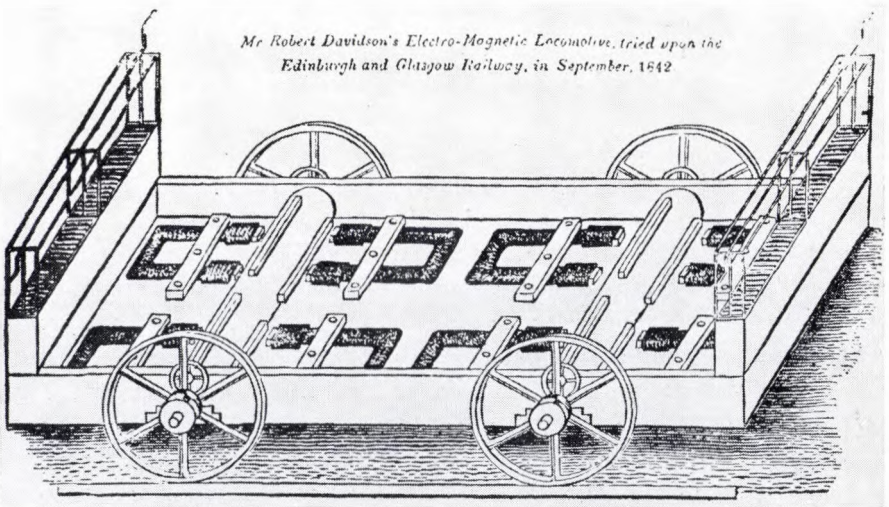


Fig. 11. Davidson's electric lomocotive, operated experimentally in 1842 on the Edinburgh and Glasgow Railway. The drive was provided by the attractive force between horseshoe magnets and iron strips fixed in grooves in wooden cylinders around the axles. The batteries were arranged at each end of the carriage.

Electron resists for the manufacture of integrated circuits

E. D. Roberts

In semiconductor technology the patterns on the wafers are usually produced by an optical method. However, in recent years an alternative method, in which the wafer is irradiated by electrons, has been investigated for cases in which high resolution is desired. This method requires the use of electron resists that behave like the well known photoresists when they are irradiated by electrons. Both positive and negative working electron resists are available to serve in various ways — etch resists, diffusion barriers, dielectrics, etc. It seems likely that in the early stages methacrylate-type resists will be favoured, since they are very versatile, necessitate no changes to other processes used in semiconductor technology, and lend themselves readily to the formation of metal patterns in the elegant 'lift-off' technique. At Mullard Research Laboratories a new methacrylate resist has been developed that is exceptionally suitable for the lift-off technique. Resists such as polymethylcyclosiloxane (PMCS) have also been developed at MRL for the direct application of silicon-dioxide patterns to silicon without etching. The use of this method in the manufacture of transistors and integrated circuits will allow the full resolution of electron-beam techniques to be obtained.

Introduction

One of the procedures used in making monolithic integrated circuits is photolithography. A coating of a photosensitive lacquer (photoresist) is deposited on a wafer of silicon oxidized at its surface, and is exposed to the light of a mercury UV lamp through a photomask. Depending on the type of photoresist (positive or negative) either the exposed or the non-exposed areas are dissolved in a liquid developer. At the places no longer covered by photoresist the layer of SiO_2 can be dissolved in a selective etching bath and the remaining photoresist is then removed, leaving a pattern of SiO_2.

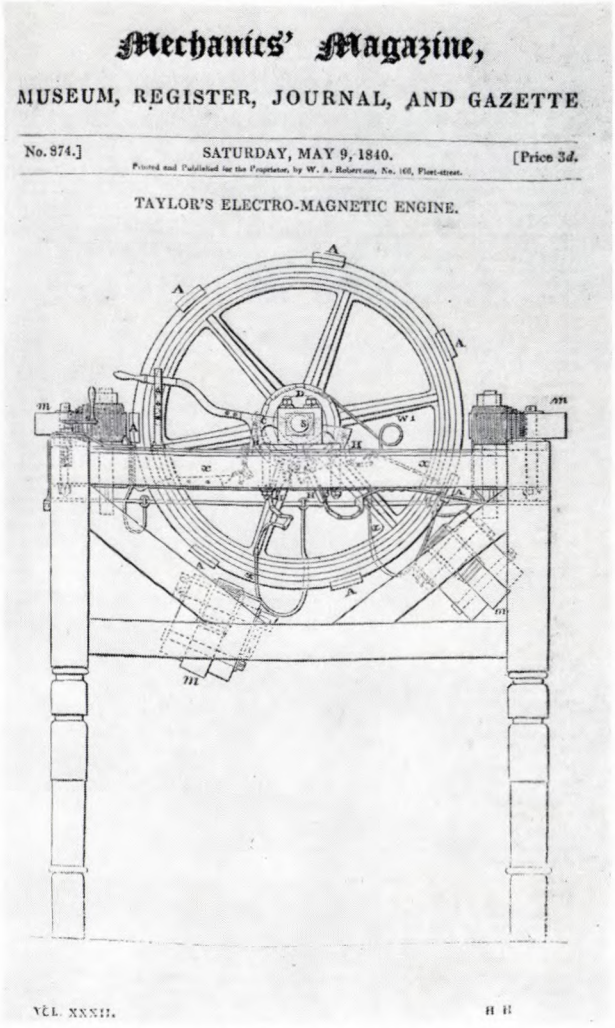
Electron-beam techniques for defining the pattern required in semiconductor technology have been investigated for a number of years. Used as an alternative for the mask and the mercury lamp in photolithographic procedures, they offer the following advantages.

The beams used have high electron energies (5-30 keV) so that diffraction effects are negligible. This allows better definition of details in the range 0.5-2 μm than is possible with photolithographic procedures, and it is expected that the yield of devices with such detail should be substantially better than is obtainable by photolithography. It is also feasible to focus electron beams to very small diameters (3 nm upwards) and the use of finely focused beams can allow patterns to be made with detail much smaller than is possible by photolithography. It has been shown that aluminium conductors only 0.1 μm wide can be made by electron-beam techniques [1]. The ability to make finer detail can lead to an increase in the packing density of the components of integrated circuits with consequent reduction in their size and cost. Smaller components will also reduce power requirements and increase the speed of operation. The ease and precision with which electron beams can be deflected by electric or magnetic fields makes electron-beam techniques relatively amenable to automation. The energy of the electrons is much greater than that of the UV photons, which will not exceed a few electron volts. This broadens the range of materials from which a choice of a resist can be made.

After a brief description of electron-beam equipment now in use and the modes of operation, this article discusses a number of resist systems which are particularly applicable to electron-beam techniques, and explains the mechanisms by which they operate. Several applications of electron-beam techniques will be mentioned. Special attention will be given to the possibility

E. D. Roberts, B.Sc., is with Mullard Research Laboratories, Redhill, Surrey, England.

[1] M. Hatzakis and A. N. Broers, Record 11th Symp. on Electron, ion, and laser beam technology, Boulder 1971, p. 337.


Mechanics' Magazine,

MUSEUM, REGISTER, JOURNAL, AND GAZETTE

No. 874.] SATURDAY, MAY 9, 1840. [Price 3d.

TAYLOR'S ELECTRO-MAGNETIC ENGINE.

VOL. XXXII. H H

Fig. 12. Taylor's motor (1840). The rotor is a wooden wheel with seven soft-iron armatures *A*. Four electromagnets *m* are connected and disconnected in turn.

> that the means taken to retain the magnets and iron in their assigned positions were insufficient. They required to be more firmly secured, and their distances had to be somewhat increased, by which considerable power was lost.'

So it was purely mechanical considerations which limited the power of Davidson's locomotive.

We have enough data to estimate the efficiency of this machine. The whole locomotive weighed over five tons and went at four miles (6.5 km) per hour on the level. A horse could pull such a vehicle [24], and a likely power is about 500 watts. The batteries were of a kind capable of giving about 0.2 ampère per square inch (about 3 A per dm^2) of surface of zinc plate [25]. Each cell had a zinc area of 360 square inches (23 dm^2) and ought therefore to have given about 70 ampères at about one volt. In the original form with forty cells the theoretical power available was therefore about $40 \times 70 \times 1 = 2800$ watts. However, that proved inadequate and the battery capacity was almost doubled. The power available then would have been about 5 kilowatts. The overall efficiency of Davidson's locomotive is therefore in the range 500/5000 to 500/3000, or from 10% to 16%.

The efficiency of electromagnetic engines was a matter of some interest in the 1840s. In 1843 Charles Wheatstone described his rheostat, or variable resistance, in a paper to the Royal Society on Electrical Measurements [26]. The rheostat was developed initially as a measuring device, but Wheatstone stated that it could be used for controlling the speed of a motor or keeping it constant as the battery varies. He also stated:

> 'Since the consumption of materials in a voltaic battery . . . decreases in the same proportion as the increase of the resistance in the circuit, this method of altering the velocity has an advantage which no other possesses, the effective force is always strictly proportional to the quantity of materials consumed in producing the power, a point which, if further improvements should ever render the electromagnetic engine an available source of mechanical power, will be of considerable importance.'

This, of course, is not correct, since it ignores the power wasted in the resistance. But it suggests that he had conducted some experiments designed to relate the power obtained from an engine and the consumption of materials in the battery.

In 1850 the Philosophical Magazine reported that one grain (about 0.07 g) of coal consumed in the furnace of a Cornish engine would lift 143 pounds (about 55 kg) one foot (about 30 cm), whereas one grain of zinc consumed in a battery would lift only 80 pounds through the same distance. The cost of zinc was much more than the cost of coal, and therefore under the most favourable conditions 'the magnetic power must be nearly 25 times more expensive than steam power' [27].

The problem which designers of electromagnetic engines had to solve was that of obtaining a useful output stroke from a power source which gave great power but through only a short distance.

One approach was that adopted by Wheatstone who, in 1841, patented a number of rotary engines in which he sought to increase the output stroke in relation to the distance between the electromagnet and the armature. His idea was to make the armature move in a direction inclined to the direction of the magnetic force. He was not expecting to obtain more work from his machine (the relationship 'work = force × distance moved in the direction of the force' was already well known) but he wanted it in a longer output stroke. He made at least three machines of different designs to exploit the idea. In the first design a number of horse-

shoe electromagnets are arranged in a circle around the stator with their poles pointing inward. The armature is a soft-iron ring whose outside diameter is a little less than the diameter of the circle which bounds the surfaces of the electromagnet poles. The armature is carried on a crank on the main shaft and the electromagnets are energized sequentially so that the armature rolls round inside the stator, pulling the crank with it. According to the patent drawings (*fig. 13a*) the armature ring acted as a commutator by pressing on contact strips as it rolled around the stator. The machine now in the Science Museum (fig. 13*b*) has a simple fixed commutator with a contact for each electromagnet and a single wiper carried on the main axis.

Wheatstone called these machines 'eccentric electromagnetic engines'. In another one the armature is an iron disc which performs a wobbling motion as a number of coils are energized sequentially. The patent drawing shows electromagnet coils in milled slots in an iron stator disc, though the actual machine in the Science Museum has four separate horseshoe electromagnets to act on the disc.

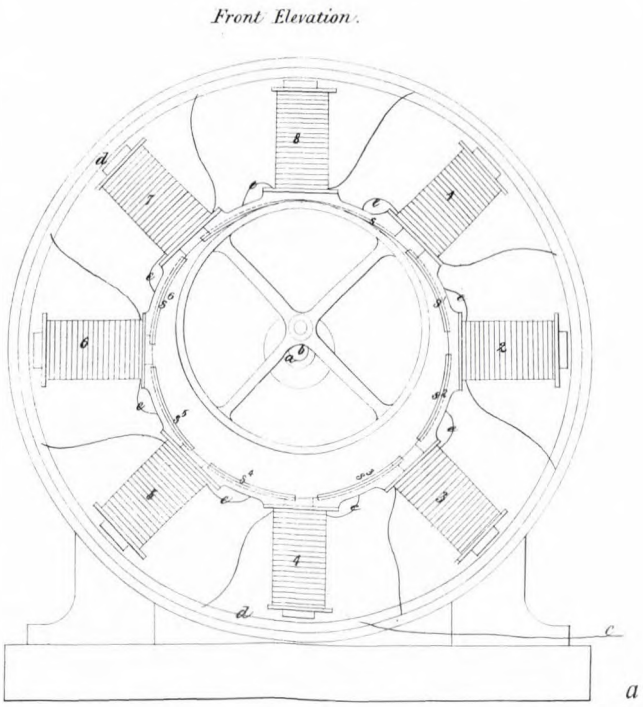
In both these machines the armature actually moves a distance equal to the spacing between two adjacent electromagnets, but, as can be seen in the illustrations, the magnetic pull has to be effective over only a much smaller distance.

Froment was the only other person to make machines of this kind. He made a machine similar to Wheatstone's first in about 1847. It may be seen working at the Musée du Conservatoire des Arts et Métiers in Paris. Froment called it the 'électro-moteur epicycloïdal'.

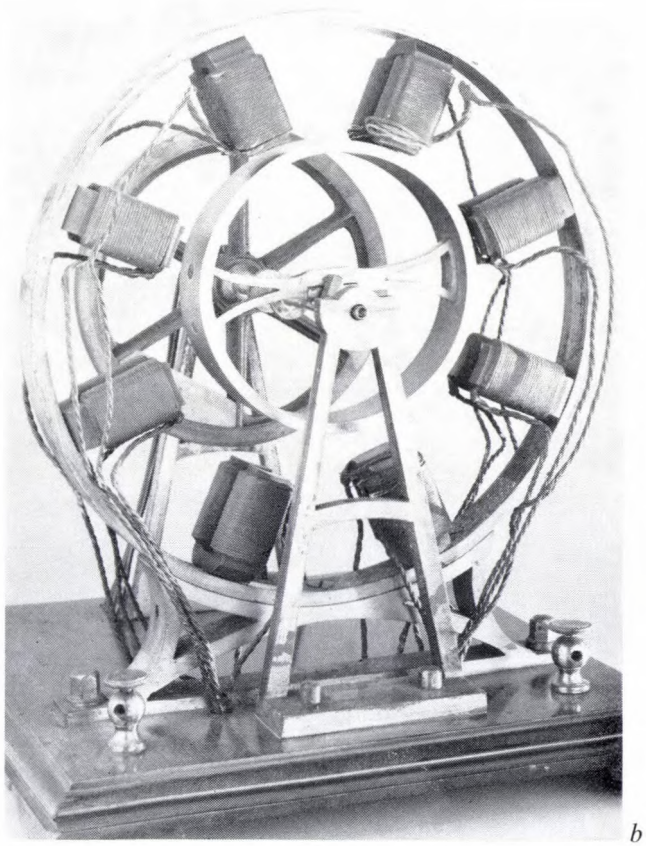
Several methods of increasing the output stroke were tried in reciprocating electromagnetic engines. One of the first men to do this was Uriah Clarke of Leicester, who exhibited an 'electromagnetic carriage' at the Leicester Exhibition in 1840. Clarke's carriage weighed only 60 pounds (27 kg), so it must have been fairly small. The motor had a single electromagnet which acted on an armature on a pivoted lever (*fig. 14*). The lever was coupled by a chain to a crank, and Clarke's idea was that the electromagnet was energized to give the armature and hence the chain one brief, sharp pull during each revolution of the crank [28].

Later in the same year Thomas Wright published his idea for increasing the working stroke of a reciprocating electromagnetic engine. His proposal was that one edge of the armature should be fixed by a hinge to the electromagnet (*fig. 15*). In that way he thought he could increase the useful working stroke from $\frac{1}{4}$ inch to $1\frac{1}{4}$ inches (from 6 mm to 3 cm) [29].

The most ingenious of all electromagnetic engines must surely be the machine constructed by Thomas



a



b

Fig. 13. Wheatstone's electromagnetic engines. The armature rolls round inside the stator; it is carried on a crank on the main shaft. The movement of the armature is therefore larger than the distance through which the attractive force operates. *a*) Drawing from an 1841 patent specification. The armature ring acts as a commutator by pressing on contact strips as it rolls around the stator. *b*) Model with a single wiper commutator carried on the main axis (at the rear). (Lent to the Science Museum, London, by Kings College, London.)

Allan in 1852. This is a reciprocating engine with four cranks and four 'piston rods' (*fig. 16*). Each piston rod carries four armatures which press on collars on the rod, but are not otherwise fixed to it. Sixteen sets

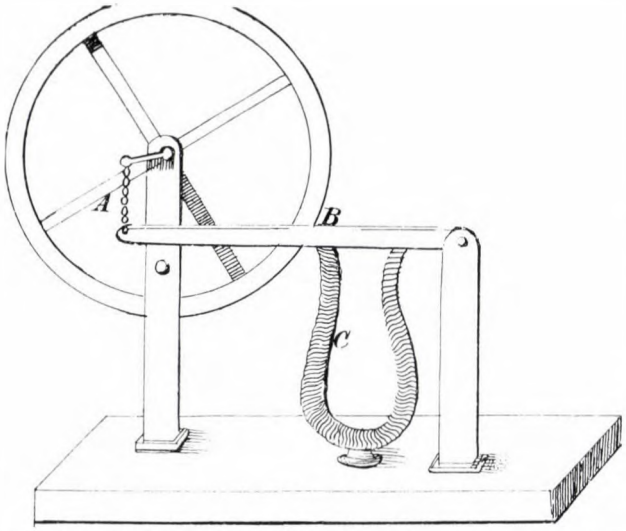


Fig. 14. Clarke's motor (1840). The output stroke of the electromagnet *C* was magnified by the lever *B*, which gives the chain *A* a brief, sharp pull during each revolution of the crank.

of coils, one set for each armature, are energized one at a time by a commutator. Each piston rod is active for one quarter of a revolution of the output shaft, and each armature is active for only one quarter of the working stroke of its piston rod. When an armature reaches its electromagnets it is stopped by them but the piston rod continues its travel, driven by the next armature in sequence [30].

None of these machines were successful, in the sense of being adopted commercially. They failed for two reasons. Firstly, the only electricity supply was from batteries and, as we have seen, they were much more expensive than steam power. But if that were the whole story then electromagnetic engines would have flourished in the 1880s when generators became available for electricity supply. However, the French writer on electrical engineering, Comte Th. du Moncel, wrote in 1878 that [31]

> 'Attempts have been made . . . to employ the attractive effect of electromagnets . . . as a motive force . . . But it can almost be predicted that those motors which are most successful when of small size are precisely those which give worst results when of large size, should they still give any, which does not always happen.'

When practical generators were developed in the 1870s it was soon found that an identical machine would function as a motor and that larger ones were more efficient than smaller. The development of electromagnetic engines came to an abrupt halt, except in two applications. Until the widespread adoption of the telephone thousands of ABC telegraph receivers were in daily use, and each contained a tiny electromagnetic engine to turn the pointer. More recently battery-powered electric clocks, driven by an electromagnetic engine, have become popular.

Practical electric motors

The electric generator was developed by several independent inventors in 1866. This invention made large-scale electricity supply a practical possibility. At first the new electrical industry supplied only electric lighting, but it was soon found that a machine identical to the generator could be used as a motor (*fig. 17*). It is not known whether such a machine was first tried as a motor because theoretical reasoning suggested the interchangeability of motors and generators, or whether the motor action was discovered accidentally by someone who tried to operate two generators in parallel.

The first public demonstration of the electrical transmission of power was given during the Vienna Exhibition of 1873. The motor and generator were identical Gramme machines with permanent-magnet excitation. The generating machine was driven by a steam engine and supplied current to the second machine, 550 yards (500 metres) away, which drove a pump [32].

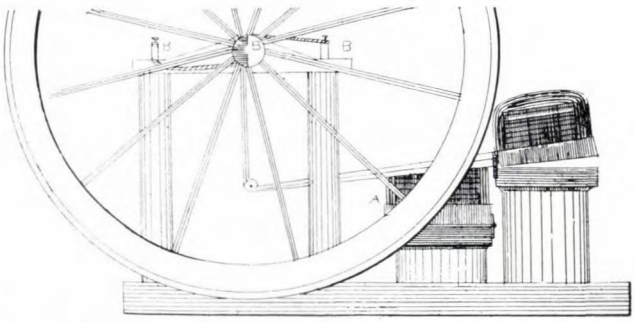


Fig. 15. Wright's motor (1840). The output stroke is again magnified by a lever; the armatures are hinged to the edge of the electromagnet.

Similar demonstrations were given at the Loan Collection of Scientific Apparatus in London and at the American Centennial Exposition, both in 1876. The people responsible for these demonstrations seem to have been very eager to transmit power over long distances, and they appreciated the advantages of high-voltage transmission. In 1880 at the Munich Exhibition Marcel Deprez transmitted electric power over a distance of 34 miles (55 km) from Miesbach to Munich through a double line of telegraph wire. The machines

used were ordinary Gramme generators, with wound fields (*fig. 18*), but the usual windings were replaced by coils with more turns of finer wire. The machines had a resistance of 470 ohms each and the line 950 ohms. The terminal voltage was 2400 V at the generator and 800 V at the motor, and the overall efficiency of the power transmission was 33% [32].

At the International Electrical Exhibition in Paris in 1881 Deprez demonstrated several motors, including some driving sewing machines and lathes.

Compound winding of generators to give a constant output voltage was developed in the early 1880s. Constant voltage was very important for the newly invented incandescent-filament lamps. It was soon shown that motors could be made to run at constant speed over a range of loads by adopting compound winding there also [21] [32].

Also in the 1880s electric power was used for ploughing in France. The arrangement was similar to that used for ploughing by steam power. Two Gramme machines were required, mounted on wheeled trolleys at opposite ends of the field, and the plough was then drawn along by a rope which was pulled first to one motor and then to the other. Current was supplied from a steam engine and generator installed under cover [32].

The first electric motor to be mass produced was probably the small motor used by Edison in his 'electric pen' (*fig. 19*). The electric pen was a stencil cutter in

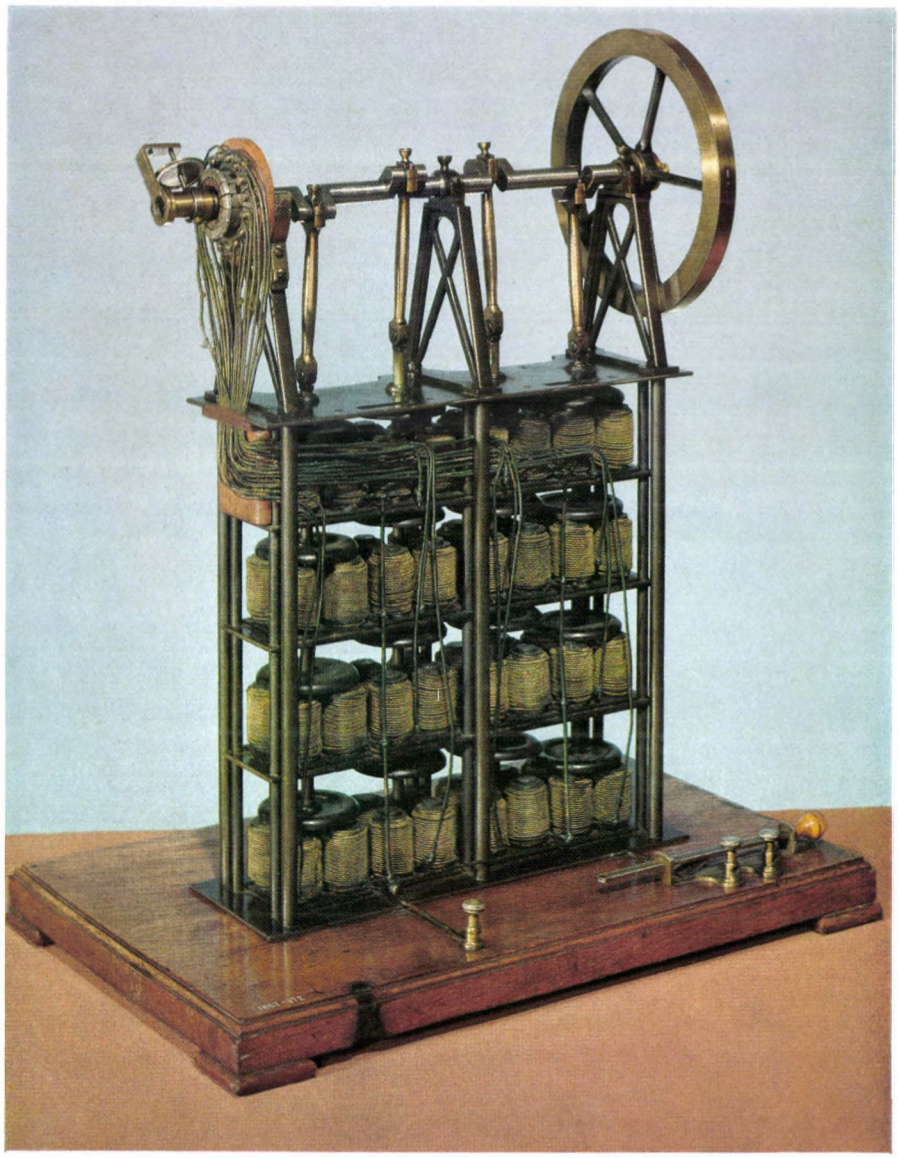


Fig. 16. An ingenious electromagnetic engine due to Allan, with four 'piston rods' (1852). Each piston rod is pulled downwards by four groups of electromagnets; these do not operate simultaneously but in sequence, thus magnifying the stroke. (Photo Science Museum, London.)

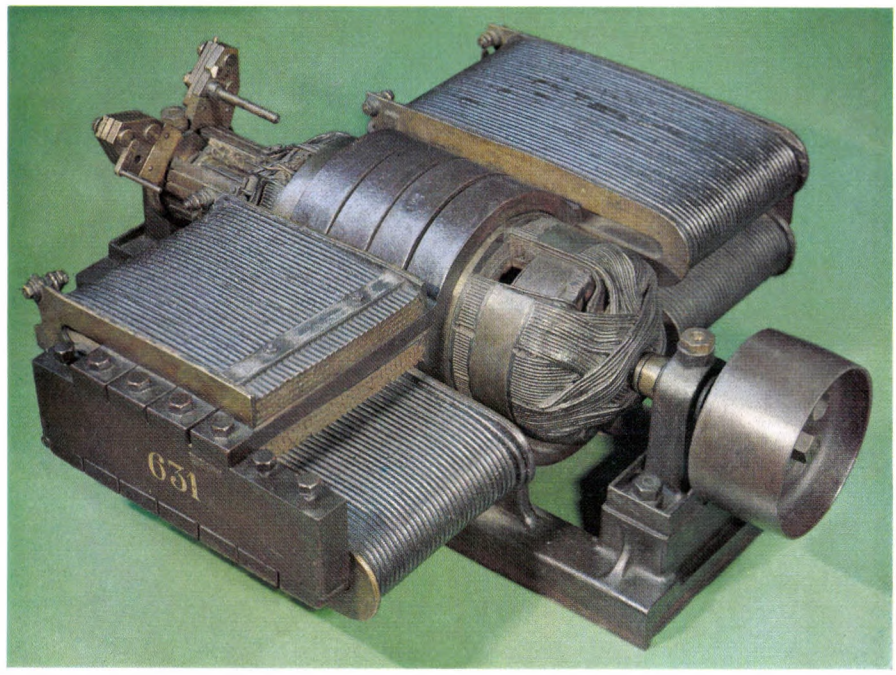


Fig. 17. Cut-away view of an early Siemens generator that was used as a motor. (Photo Science Museum, London.)

which a motor running at about 65 revolutions per second drove an oscillating needle to perforate the stencil. The motor was only 1.6 inches (4 cm) high and 0.8 inches (2 cm) wide. It had a fixed electromagnet, a rotor consisting of a straight permanent magnet, and a commutator which reversed the connections of the electromagnet twice per revolution [33]. Tens of thousands of electric pens were sold [9].

In 1879 Edison began marketing a small motor intended for running sewing machines and other light machinery from the electric-lighting mains. The construction was very similar to Edison's early generators, with their characteristic long field windings [33].

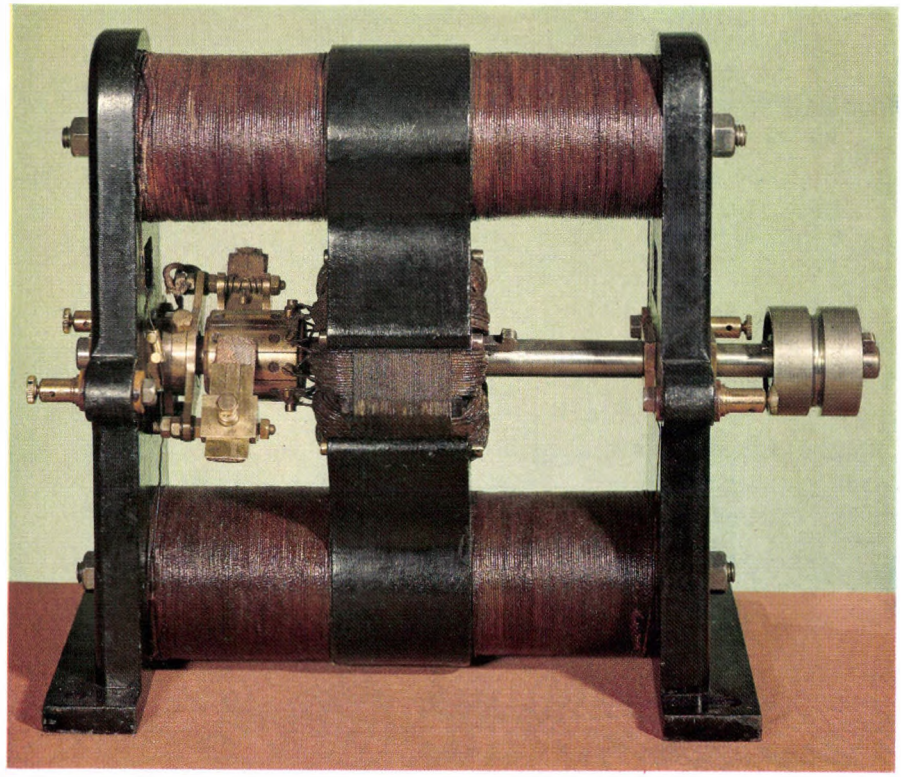


Fig. 18. Gramme generator with wound fields. (Photo Science Museum, London.)

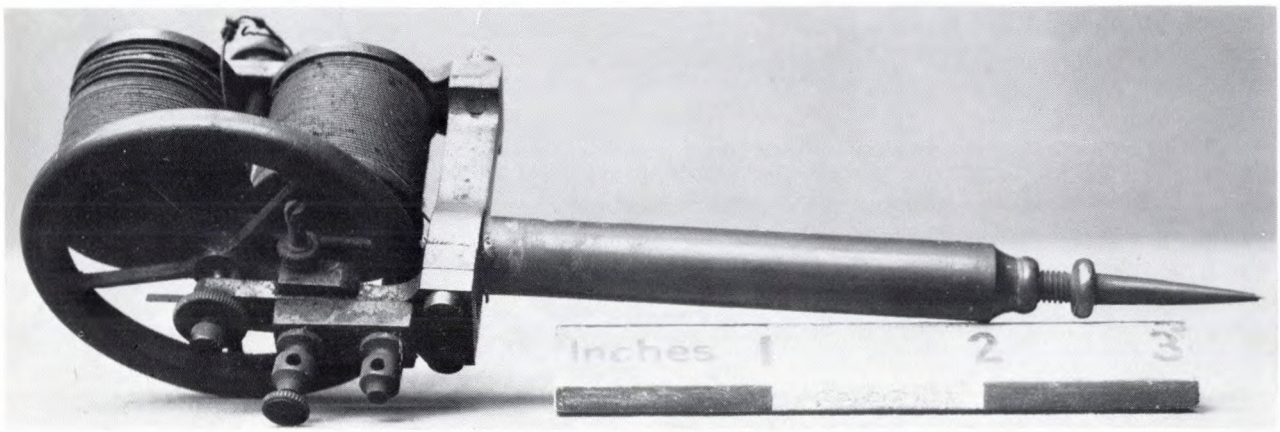


Fig. 19. Edison's 'electric pen', the first electric motor to be mass-produced. The oscillating needle point was used as a stencil cutter; the motor ran at about 65 revolutions per second. (Photo Science Museum, London.)

The most important early application of electric power was for railways and tramways. Siemens and Halske gave the first successful demonstration at the Berlin Industrial Exhibition in 1879. They had a railway with 325 yards (300 m) of track formed into a closed oval. The locomotive was a four-wheeled truck carrying an ordinary Siemens dynamo/electric machine which was geared to the wheels (*fig. 20*), and it pulled an open carriage with passengers. The electricity supply was through a central live rail with the running rails providing the return path. In the following year Egger showed a railway at Vienna in which the two running rails were insulated from one another and formed the positive and negative conductors [32].

The first permanent railway open to the public was constructed by Siemens and Halske at Lichterfelde and opened in 1881. It ran for $1\frac{1}{2}$ miles ($2\frac{1}{2}$ km) between Lichterfelde station and an Army Training School. Each carriage was rather like a tramcar in appearance, and had a single motor fixed centrally under the floor [32].

The first electric railway in the United Kingdom was the line from Portrush to Bushmills, opened in 1883. The total length was six miles ($9\frac{1}{2}$ km), and the railway had several features of interest. Electrical operation was not envisaged when the line was first proposed, but it was adopted because abundant water power was available at Portrush, where a hydroelectric generating station was built. The motors were placed centrally under the floor of the carriages, and the driver had a lever which enabled him to reverse the motor by moving the brushes. There was a driving position at each end of the carriages [32].

Direct-current motors for both traction and other applications were developed by many manufacturers during the later 1880s and 1890s. An important difference between early industrial motors and generators of the same period was that provision had to be made in the motors for reducing sparking at the commutator by keeping the brushes on the magnetic neutral axis as the load varied. Initially the engine-driver or other person in charge of the machine was given some means of adjusting the brush position as the load varied, but that was only practical when dealing with a single motor in an accessible location. The American engineer Frank J. Sprague, who established the Sprague Electric Railway and Motor Company in 1884, found that by making a compound-wound motor with the series field coil arranged at an angle to the main, shunt coil, he could maintain spark-free commutation under varying load with a fixed brush position [21].

The overall height of a traction motor is a critical dimension, especially when the motor is mounted on

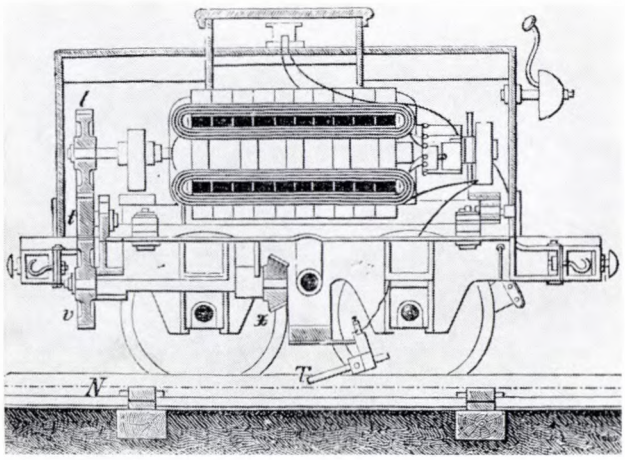


Fig. 20. Cross-sectional drawing of Siemens and Halske's electric locomotive (1879). A Siemens generator was used as the motor, and was coupled to the wheels by the gears *l*, *t*, *v*, *x*. The current was taken from a central live rail *N*.

the bogie of an eight-wheeled coach rather than on the floor of a simple four-wheeled vehicle. This requirement led to a design of motor which appears at first sight to have only two poles but does in fact have four, although only two poles are wound.

Alternating-current motors

In the last years of the nineteenth centrury there was much controversy between the supporters of direct-current electricity supply and those who favoured alternating current. The advantage of a.c. was that transformers could be used; electric power could then be converted to high voltage at the power station, transmitted more efficiently than at low voltage, and reduced to low voltage at the load. The supporters of d.c. pointed out that it was difficult to operate alternating-current generators in parallel, and d.c. motors were generally far better then a.c. motors.

In a lecture to the Institution of Civil Engineers in 1883, entitled 'Some Points on Electric Lighting', Dr John Hopkinson discussed the operation of alternating-current generators in parallel [34]. Even the possibility of parallel running was doubted by some engineers, but Hopkinson showed that it could be done if the machines had similar output waveforms. He then said that if two such generators were being run in parallel and the power driving one was cut off, then that machine would continue running as a motor. Furthermore, the motoring machine would maintain synchronism with the other machine which continued to generate.

Since the synchronous motor is essentially a constant-speed machine, it can only be used in a few special applications. With the introduction of alternating-current electricity-supply systems there was a need for a general-purpose a.c. motor. It was found that the ordinary direct-current motor could be modified to run on alternating current. The problem was that the inductance of the field winding led to excessive sparking, and eddy currents in various parts of the machine caused power loss and overheating. The use of laminated construction for the stator as well as the rotor alleviated the situation, but a.c. versions of the ordinary d.c. motor have never proved satisfactory except in small sizes. The Siemens firm were developing a.c. commutator motors as early as 1884 and Alexander Siemens spoke on the subject at the Institution of Electrical Engineers in November 1884 [35].

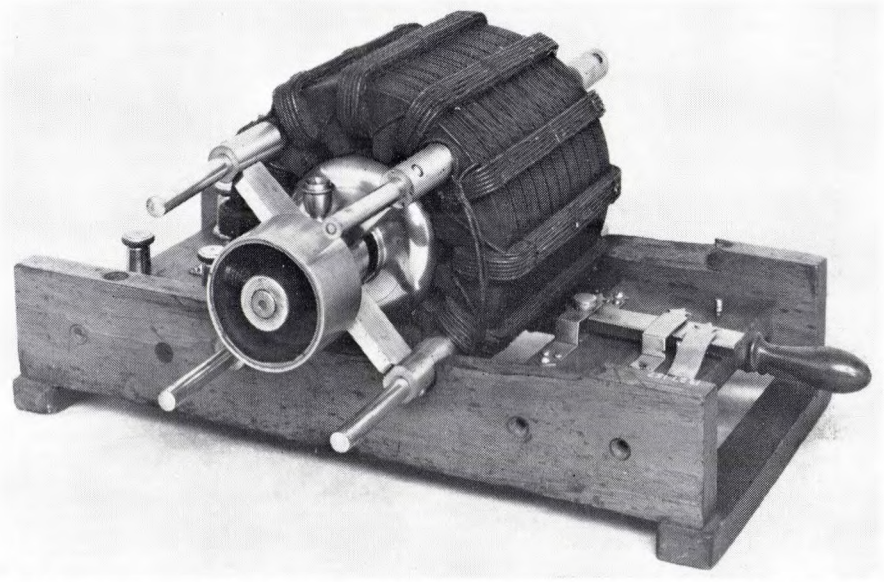


Fig. 21. Tesla's two-phase induction motor (1888). Neighbouring groups of three of the twelve stator coils are successively connected in series. The stator iron is laminated. (Crown Copyright. Science Museum, London.)

The most important type of motor today, in terms of total power installed, is the induction motor. The origin of this is to be found in Arago's disc, for both Arago's disc and the induction motor utilize the force between a moving magnetic field and a conductor carrying a current induced by the magnetic field. Arago produced a rotating field by spinning a bar permanent magnet about an axis perpendicular to the bar. In induction motors a similar magnetic field is produced by currents in the field coils, without the need for any moving parts.

In 1884 two men, Galileo Ferrari in Italy and Nicola Tesla in America, showed, independently of one another, how fixed coils carrying suitable phase-displaced

currents could produce a rotating magnetic field and cause an armature to rotate. In Turin, Ferrari used two coils arranged at right angles and fed from the same a.c. supply, one directly and the other through an inductance. Although he produced rotation he did not develop his device into a practical motor [9]. At about the same time Tesla read a paper to the American Institute of Electrical Engineers in which he described a two-phase generator and motor [36]. The generator was made by fitting two pairs of slip-rings to a d.c. generator; the motor consisted of an iron ring with four coils wound around it and each pair of opposite coils connected together and to one pair of slip-rings. He showed that the arrangement gave a rotating field and that pivoted metal discs placed within the iron ring would spin. The success of this experiment led Tesla to build the first practical two-phase induction motor (*fig. 21*). Within a few months he had made three-phase machines also, and the Westinghouse Company began to manufacture them.

In 1891 W. Langdon Davies showed that it was possible to run an induction motor from a single-phase supply. He made an experimental machine whose stator was an iron ring carrying six coils with six pole pieces between them (*fig. 22*). Opposite coils were connected together, so there were three groups of two coils. Two groups were connected directly to the supply, the third was connected through a resistance so as to introduce a phase shift. The rotor was a ring similar to the stator, but smaller, and wound with six short-circuited coils. By 1897 Langdon Davies was producing single-phase induction motors commercially [9].

By the end of the nineteenth century, less than one hundred years after Oersted's demonstration that an electric current could produce motion, the main types of electric motor had been invented and were in production.

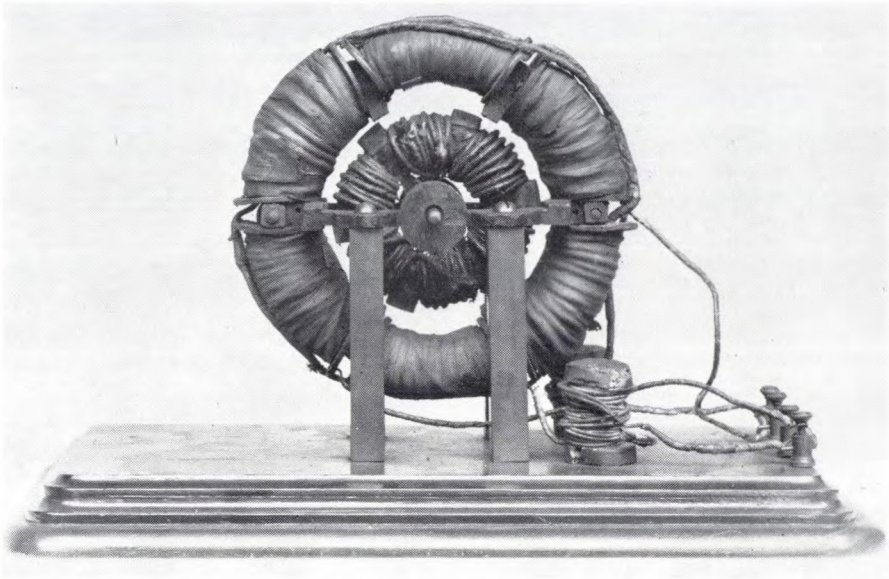


Fig. 22. Langdon Davies's induction machine for a single-phase supply (1891). Four of the six stator coils are connected directly to the supply, two via a resistance. (Lent to the Science Museum, London, by the Langdon Davies Motor Co.)

Bibliography

[1] B. Dibner, Oersted and the discovery of electromagnetism, New York 1961.
[2] H. C. Oersted, Experimenta circa effectum conflictus electrici in acum magneticam, Copenhagen 1820.
[3] Phil. Trans. Roy. Soc. **115**, 467, 1825.
[4] See a biography of Faraday, such as: L. Pearce Williams, Michael Faraday, London 1965, or: Brian Bowers, Michael Faraday and electricity, London 1974.
[5] Phil. Mag. **59**, 241, 1821.
[6] Mechanical effect of electricity, Mechanics' Magazine **3**, 249-250, 1824.
[7] G. Moll, Electro-magnetic experiments, Edinburgh J. Sci. **3**, 209-218, 1830; also in Amer. J. Sci. **19**, 329-337, 1830.
[8] Phil. Mag. **8**, 45, 1832.
[9] P. Dunsheath, History of electrical engineering, London 1962.
[10] J. Henry, On a reciprocating motion produced by magnetic attraction and repulsion, Amer. J. Sci. **20**, 340-343, 1831.
[11] Phil. Trans. **123**, 319, 1833.
[12] Phil. Mag. **7**, 107, 1835, and **12**, 190, 1838.
[13] Personal communication from Dr P. van der Star, Keeper of the Physics Department, Rijksmuseum voor de Geschiedenis der Natuurwetenschappen, Leiden, The Netherlands.
S. Stratingh and C. Becker, Electromagnetische bewegingskracht en aanwending daarvan tot een electromagnetischen wagen, Algemeene Konst- en Letterbode, 1835, 402-422, and plate.
[14] B. Woodcroft, Abridgements of specifications relating to electricity and magnetism 1766-1857, London 1859; in the introduction on p. lxx he quotes the article 'Voltaic electricity' in the Encyclopaedia Britannica.
[15] English Patent No. 7386/6 June 1837.

[16] H. C. Passer, The electrical manufacturers 1875-1900, Harvard 1953.
[17] M. MacLaren, The rise of the electrical industry in the nineteenth century, Princeton 1943.
[18] Th. Davenport, Amer. J. Sci. 32, app. 4, and 33, app. 2, 1837.
[19] Phil. Mag. 15, 164, 1839.
[20] P. Forbes, On the application of electro-magnetism as a motive power, Annals of Electricity 5, 239-240, 1840.
[21] F. J. Sprague, Digging in the mines of the motors, Electr. Engng. 53, 695-707, 1934.
[22] Taylor's electro-magnetic engine, Mechanics' Magazine 32, 693-696, 1840.
[23] D. Mackie, Prospects of electro-magnetism as a prime mover, with a notice of Mr. Robert Davidson's electro-magnetic locomotive, lately tried on the Edinburgh and Glasgow railway, Pract. Mechanic 2, 1843.
[24] Personal communication from Dr J. Coiley, Keeper of the National Railway Museum, York, England.
[25] Personal communication from Dr D. Vaughan, Deputy Keeper of the Physics Department, the Science Museum, London.
[26] C. Wheatstone, An account of several new instruments and processes for determining the constants of a Voltaic circuit, Phil. Trans. 133, 303-327, 1843.
[27] Phil. Mag. 35, 550, 1850.
[28] U. Clarke, Annals of Electricity 5, 33-34 and plate 1, and 304-305, 1840.
[29] Th. Wright, Annals of Electricity 5, 108-110 and plate 3, 1840.
[30] Mechanics' Magazine 60, 265-268, 1854; 66, 389, 1857; 67, 4-5, 1858.
[31] Th. du Moncel, Exposé des applications de l'électricité, 1878, quoted in translation in: D. P. Heap, Report on the international exhibition of electricity held at Paris, 1881, Washington 1884.
[32] R. Wormell, Electricity in the service of man, London 1890.
[33] Edison's electrical generator (not really an adequate title for an article that also deals with motors), Sci. Amer., 18 October 1879, 242-243.
[34] J. Hopkinson, Some points on electric lighting, Minutes of Proceedings of the Institution of Civil Engineers, 1883; reprinted in: J. Hopkinson, Original papers, Cambridge 1903.
[35] A. Siemens, contribution to the discussion on the lecture by W. Grylls Adams, The alternate-current machine as a motor, J. Soc. Telegraph Engrs. 13, 515-540, 1884.
[36] N. Tesla, J. Amer. Inst. Electr. Engrs. 5, 308, 1888.

Summary. The electric motor has a history of a century and a half. The article covers the period until about 1900. After Oersted's discovery in 1820 machines were made in many places to demonstrate that mechanical movement could be produced by electricity. From this time 'electromagnetic engines' began to appear; these were machines in which a continuous movement was produced by the attractive and repulsive forces from periodically switched electromagnets. Such machines were used as practical sources of mechanical power, and also in experiments to demonstrate the electric propulsion of boats (Jacobi) and railway carriages (Davidson). The electricity supply was provided by batteries, so that these machines were very expensive to operate. With the development of the d.c. dynamo (Gramme, Siemens), cheaper electricity became available; at the same time it was found that these dynamos could also be put to good use as motors. Many practical applications now began to appear: as a drive for tools (Edison) and the very important one of motive power on railways and in trams (Siemens and Halske, Sprague). Alternating current had advantages for the transmission of current over long distances; and the induction motor was developed (Tesla).

Designing a small d.c. motor

R. H. Dijken

Small d.c. motors

Every year, more than a hundred million small d.c. commutator motors are made throughout the world — far more than any other type of electric motor. These motors are generally powered by dry batteries (toys, electric shavers, portable recorders and players), but sometimes by rechargeable nickel-cadmium batteries (shavers, film cameras) or from the mains via a transformer and rectifier (model trains).

In designing a small d.c. commutator motor it is still quite common to start by modifying an existing design in the light of practical experience. In this article an attempt will be made to contribute towards a more theoretical approach mainly concerned with the efficiency and the dimensions. First of all it will be shown that the efficiency and the speed can both be expressed in terms of two dimensionless parameters,

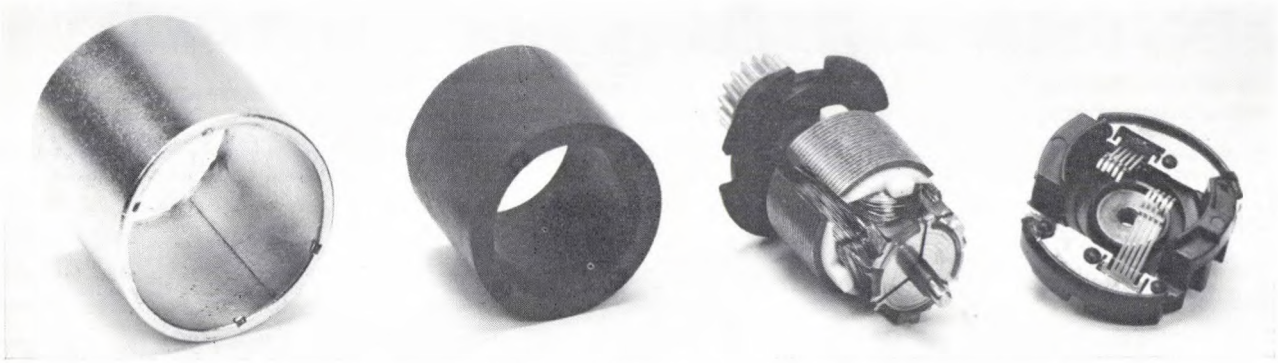


Fig. 1. A small d.c. commutator motor, taken apart. From left to right: steel outer housing, plastoferrite ring magnet, rotor with three teeth and three flat commutator segments, bearing cap with brushes and electrical connections.

The supply voltage depends on the application, but is rarely lower than 1 volt (one dry cell or one nickel-cadmium cell) or higher than 12 volts (one car battery). The power delivered generally lies between a few tenths of a watt (record players, small toys) and a few watts (shavers). The life required of the motor may be as much as several thousand hours (recorders and players).

Two important aspects of the design of a small d.c. commutator motor are the efficiency and the life; the life is mainly determined by the commutator and the brushes. A high efficiency is especially important for battery-operated equipment, because of the high cost per kilowatt hour of the electrical energy used; when the motor is powered by nickel-cadmium cells, sufficiently long operation per charging cycle is desirable. Other factors apart from the efficiency and the life are also important, of course; these include the motor torque at various speeds and also the dimensions, weight and prime cost (the cost of manufacture).

Dr Ir R. H. Dijken is with Philips Small Domestic Appliances Product Division, Hoogeveen, the Netherlands.

one containing motor characteristics only and the other also including the torque required by the load. Two other important points are then considered: first, the choice of the magnetic material, and then the method of choosing the relative dimensions of the rotor to give the best compromise between the power delivered, the dimensions and the efficiency. It is found that the optimum length-to-diameter ratio of the rotor is much smaller than that normally encountered, and the optimum ratio of tooth width to rotor diameter much larger.

Construction of a small d.c. commutator motor

To give an impression of the construction of the d.c. motors discussed in this article, the various parts of a motor are shown in *fig. 1*. The rotor consists of a laminated core with three slots. There are three rotor coils, connected to the three segments of a planar commutator. The coils are insulated from the core by a layer of injection-moulded plastic, shaped to give the correct support to the coils.

The stator consists of a permanent ring magnet of plastoferrite (a plastic-bonded ferrite, type SP 130 here). The magnet is surrounded by a steel ring that completes the magnetic circuit externally (*fig. 2*).

Bearing caps are fitted at each end of the motor. The cap at the commutator end contains the brush assembly, the other is used to attach the motor to the system it drives.

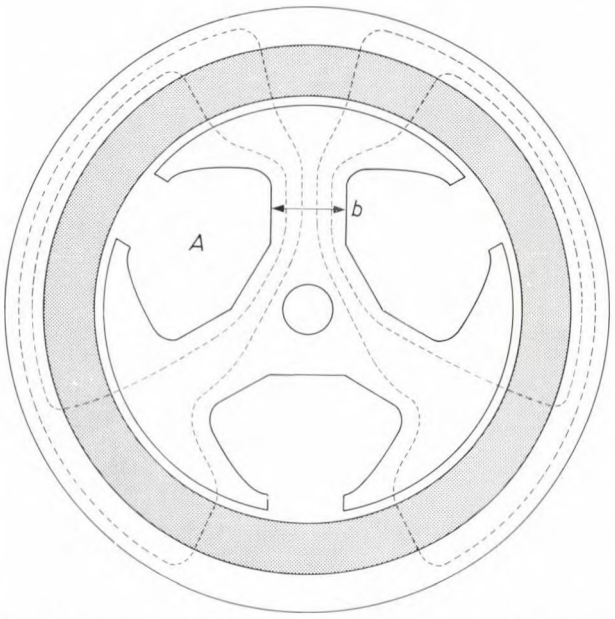


Fig. 2. Cross-section through a small d.c. motor, showing the magnetic lines of force. The shaded ring is the plastoferrite magnet. *A* space for windings. *b* width of rotor tooth.

Efficiency

The losses occurring in the motor can be classified as either electrical or mechanical losses. The electrical losses are those of the rotor and commutator resistances, and those caused by eddy currents and magnetic hysteresis. The mechanical losses are caused by friction in the brushes, in the bearings and due to air resistance.

There is however another way of classifying the losses: they can be split up into losses that appear when the motor is switched on but is not rotating (e.g. because the shaft is kept fixed), and those that occur when the rotor carries no current but is rotating (e.g. because it is driven externally). The first category only includes the conduction losses, and the second only the iron losses and the various frictional losses. This second method of classification is more suitable for efficiency calculations, since all the losses in the second category are approximately proportional to the motor speed.

The conduction losses in the rotor winding depend on the rotor resistance R between the two commutator segments with which the brushes make contact. This resistance is called the diametral rotor resistance, and is the resistance of the two rotor paths in parallel [1], one path consisting of one rotor coil and the other of two rotor coils in series. There is also a contact resistance between the brushes and the commutator. This contact resistance is almost independent of the current with the noble-metal brushes and segments used here (this is not the case with carbon brushes). It can therefore be combined with the resistances of the actual brushes, the lead wires and possibly the switch, to give the constant resistance R_b.

The equivalent circuit of the d.c. motor then takes the simple form shown in *fig. 3*. The applied voltage V, which is the sum of the voltage across the resistance $R + R_b$, through which the current I flows, and the speed voltage (or back e.m.f.) E induced in the rotor is therefore given by:

$$V = E + I(R + R_b). \tag{1}$$

The power supplied P_i is equal to VI, so that:

$$P_i = EI + I^2(R + R_b). \tag{2}$$

The conduction losses are $I^2(R + R_b)$; the remainder EI of the power supplied is converted into mechanical power and is called the electromechanical power P_e. An efficiency η_e, which is often called the electrical efficiency, can be defined as follows:

$$\eta_e = \frac{P_e}{P_i} = \frac{V - I(R + R_b)}{V} = \frac{E}{V}. \tag{3}$$

The speed voltage is given by:

$$E = 2nN\hat{\Phi}, \tag{4}$$

where n is the number of revolutions per second, N is the total number of turns of all three rotor coils and $\hat{\Phi}$ is the maximum flux through one rotor tooth (this flux is at a maximum when the tooth is exactly opposite the pole of the magnet).

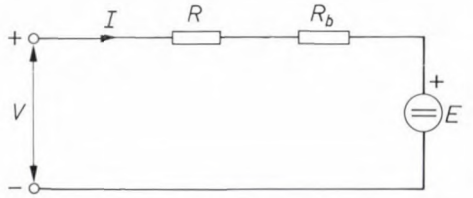


Fig. 3. Equivalent circuit of the d.c. motor. V supply voltage. I supply current. R diametral rotor resistance. R_b commutator resistance. E speed voltage (back e.m.f.) induced in the rotor coils.

Equation (4) gives the *average* speed voltage. In reality, the speed voltage has a ripple, since it is the resultant of the voltages induced in the three separate rotor coils. If we assume that the

[1] An introductory treatment of the commutator motor is given in E. M. H. Kamerbeek, Electric motors, Philips tech. Rev. **33**, 215-234, 1973.

flux through the rotor teeth varies sinusoidally and that the commutation occurs instantaneously and at the zero crossing of the induced voltage, then the voltages induced in rotor coils *1*, *2* and *3* have the waveform shown in *fig. 4*. The resultant voltage is given by the thick line, and is found to be the same for both rotor paths at any given moment. It is clear from the figure that the average value is given by

$$E = \frac{3}{\pi}\int_{\frac{\pi}{3}}^{\frac{2\pi}{3}} \hat{e}\sin\theta\,\mathrm{d}\theta = \frac{3}{\pi}\hat{e}.$$

The angle θ gives the position of the rotor in relation to the stator. The voltage induced in one rotor coil is

$$e = -\frac{N}{3}\frac{\mathrm{d}}{\mathrm{d}t}(\hat{\Phi}\cos\theta) = -\frac{N}{3}\frac{\mathrm{d}}{\mathrm{d}t}(\hat{\Phi}\cos 2\pi nt).$$

The peak value of this voltage is

$$\hat{e} = 2\pi n\frac{N}{3}\hat{\Phi}.$$

Equation (4) follows directly from this.

The electromechanical power P_e, which is equal to EI in terms of electrical quantities, is also equal to the product of the angular velocity and the torque T_e:

$$P_e = 2\pi nT_e. \tag{5}$$

It follows from the above that the electromechanical torque may be written:

$$T_e = N\hat{\Phi}I/\pi. \tag{6}$$

This torque is not all available at the shaft, since the rotating rotor is subject to a braking or loss torque T_d due to the iron losses and the frictional losses. The available torque T is therefore smaller:

$$T = T_e - T_d. \tag{7}$$

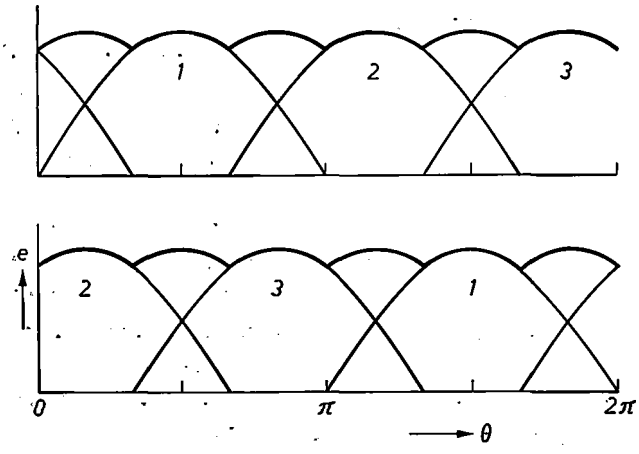


Fig. 4. The induced voltage e in the three rotor coils *1*, *2* and *3*; the top and bottom curves are for the two rotor paths; θ indicates the angular position of the rotor in the stator. Each rotor coil forms part of both rotor paths in turn; commutation is assumed to be ideal. The thick line gives the resultant speed voltage.

The braking torque, which only increases slightly with the speed, may be taken as constant when designing a motor for a limited speed range.

A mechanical power

$$P = 2\pi nT$$

is thus delivered at the shaft. The efficiency P/P_e, which expresses the iron and mechanical losses, is often called the mechanical efficiency. It can be written:

$$\eta_{mec} = \frac{T}{T + T_d}. \tag{8}$$

The overall efficiency η of the motor, the product of the mechanical and electrical efficiencies, is therefore:

$$\eta = \frac{T}{T + T_d}\,\frac{V - I(R + R_b)}{V}. \tag{9}$$

The efficiency varies with the torque T delivered by the motor and with the supply voltage V. The current I can be eliminated with the aid of the expressions given above, so that the efficiency may finally be written in the form:

$$\eta = \frac{T}{T + T_d} - \frac{\pi T(R + R_b)}{N\hat{\Phi}V}. \tag{10}$$

A number of parameters will now be introduced which can simplify the form of this expression. The motor can be driven externally at a speed n_0 such that the induced speed voltage E is equal to the applied voltage V:

$$V = 2n_0N\hat{\Phi}. \tag{11}$$

No current will then flow, and the mechanical power P_0 will all be consumed in the iron and frictional losses:

$$P_0 = 2\pi n_0T_d. \tag{12}$$

If on the other hand the shaft is held stationary and a voltage V applied, all the electrical power P_s supplied will be dissipated in the resistances:

$$P_s = \frac{V^2}{R + R_b}. \tag{13}$$

The ratio G of these two powers:

$$G = \frac{P_0}{P_s} = \frac{2\pi n_0T_d(R + R_b)}{V^2}, \tag{14}$$

is an important parameter, which characterizes a motor developed for a particular supply voltage and speed. Introducing G into the expression for the efficiency, we obtain:

$$\eta = \frac{T}{T + T_d} - G\frac{T}{T_d}. \tag{15}$$

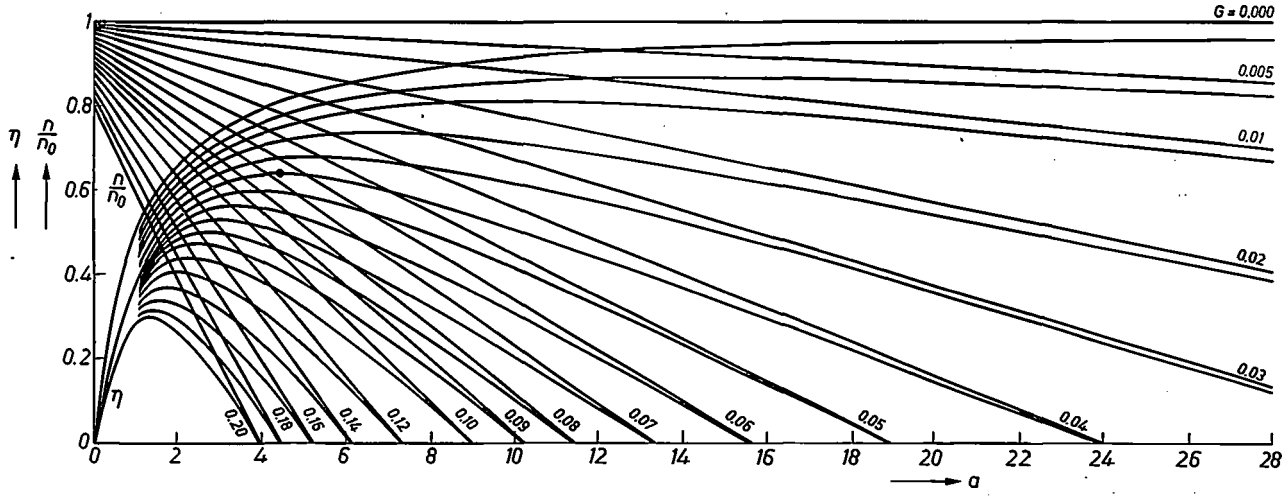


Fig. 5. *Curved lines:* The efficiency η as a function of $a = T/T_d$ (where T is the delivered torque and T_d the loss torque), with G as parameter. For the motor shown in fig. 1, $G = 0.04$ and the maximum efficiency is 0.64. *Straight lines:* speed-torque characteristics. n speed. n_0 reference speed.

It is now useful to introduce a second dimensionless parameter

$$a = \frac{T}{T_d}. \tag{16}$$

This parameter simplifies the expression for the efficiency to

$$\eta = \frac{a}{a+1} - Ga. \tag{17}$$

In *fig. 5* the curved lines give the efficiency as a function of a, with G as parameter. The smaller G is, the higher the efficiency that can be achieved. If (17) is differentiated with respect to a, it is found that for any given value of G the maximum efficiency is obtained at a load such that

$$a = 1/\sqrt{G} - 1. \tag{18}$$

This maximum efficiency is given by

$$\eta_{max} = (1 - \sqrt{G})^2. \tag{19}$$

With this optimum load the electrical and mechanical efficiencies are equal $(1 - \sqrt{G})$, as appears on substituting $T/T_d = a = 1/\sqrt{G} - 1$ in (8).

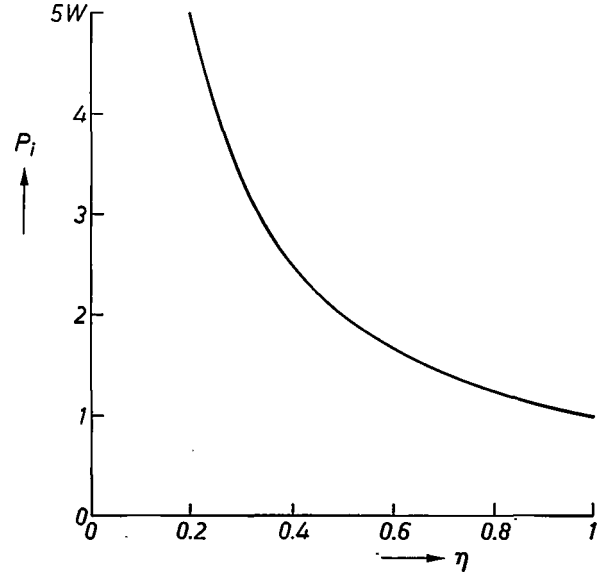


Fig. 6. The mechanical power P_i to be supplied for an output power of 1 watt, plotted against the efficiency η. The higher the efficiency, the less power can be gained by increasing the efficiency further.

What should be the target efficiency?

As mentioned in the introduction, in designing a motor a compromise has to be made between efficiency, life and prime cost. If a motor is intended for cheap toys its prime cost must be kept as low as possible, even if this has an adverse effect on the efficiency and life. In small battery-powered consumer articles such as shavers, cameras and toothbrushes, on the other hand, the efficiency should be high, even if this makes the motor more expensive; in this case, the expense and inconvenience of changing dry cells or recharging accumulators are important. But just *how* high an efficiency should be aimed at?

To find an answer to this question, let us start by looking at *fig. 6*. This figure shows a graph of the power P_i supplied as a function of the efficiency η, for a motor delivering a power P of 1 watt. It will be seen that P_i decreases as η increases, but flattens out as η becomes higher. Making the efficiency very high does not therefore give much extra saving of power.

In the second place, experience shows that at the present state of the art it is practically impossible to develop a small d.c. motor with an efficiency of more than about 80% for a delivered power of about 1 W,

even if a very high prime cost is acceptable. For these two reasons, an efficiency of about 80% should be considered as the practical maximum against which the efficiency obtained in a particular case can be compared.

A general argument for aiming at a high efficiency (energy considerations apart) can be derived from fig. 5. If $G = 0.1$, it can be seen from the figure that the maximum efficiency is 47% at $a = 2.16$. If a is varied from this value by a factor of 3, the efficiency falls to 32% on increasing a and to 22% on decreasing a. If on the other hand $G = 0.01$, the maximum efficiency (81%) is found at $a = 9$; if a is again varied by a factor of 3, the efficiency remains high (72% and 69%). This means that by making motors with a low value of G, i.e. a high maximum efficiency, the manufacturer can cover a wide range of delivered powers with a limited number of motor types.

Speed

Like the efficiency, the motor speed (or more accurately the ratio of the speed to the reference speed n_0 defined above) can be given by a simple expression containing only a and G, since the ratio n/n_0 is equal to η_e, one of the factors of η:

$$\frac{n}{n_0} = \frac{E}{V} = \frac{V - I(R + R_b)}{V} = \eta_e. \qquad (20)$$

The quantity I can be eliminated as before for the efficiency; substitution of a and G then gives:

$$\frac{n}{n_0} = 1 - G(a + 1). \qquad (21)$$

This function is plotted with G as a parameter in fig. 5 (the straight lines). These are therefore speed-torque characteristics, the converse of the more usual torque-speed characteristics. It can be seen that the speed falls off faster under the influence of the load as G increases.

Two important design aspects

Choice of magnetic material

In designing the motor it is necessary to choose the type of material for the permanent-magnet stator. At present there are four materials that might be considered when prime cost is taken into account:

1. Magnet steel (of various compositions), with a remanence of from 0.6 to 1.3 teslas; this group will be referred to here as 'Ticonal';
2. Sintered ferrite (not magnetically aligned), with a remanence of about 0.18 T (unaligned ferroxdure);
3. Sintered ferrite (magnetically aligned), with a remanence of 0.28 to 0.36 T (aligned ferroxdure);
4. Plastic-bonded aligned ferrite with a remanence of about 0.25 T (plastoferrite).

'Ticonal' is not so suitable for our purposes. The gaps between the rotor teeth make the reluctance of the magnetic circuit dependent on the position of the rotor. The flux from the magnet therefore varies periodically as the motor turns. Since a 'Ticonal' magnet is a solid piece of electrically conducting material, the flux variations will set up eddy currents in the magnet, at a frequency of six times the number of revolutions per second of the motor. The Joule heating due to the eddy currents introduces a relatively large extra loss torque.

Unaligned ferroxdure can be made in various forms with quite small dimensional tolerances. For example, it can be made in the form of a ring, which is very suitable for these motors; however, unaligned ferroxdure is not so useful because of its low remanence, which means that the rotor and magnet have to be made relatively large, thus reducing the power-to-weight ratio of the motor.

Aligned ferroxdure cannot easily be made in the form of a ring. It therefore has to be used in the motor in the form of segment magnets (*fig. 7a*) or blocks (*fig. 7b*); in the latter case, the flux is conducted to the rotor via

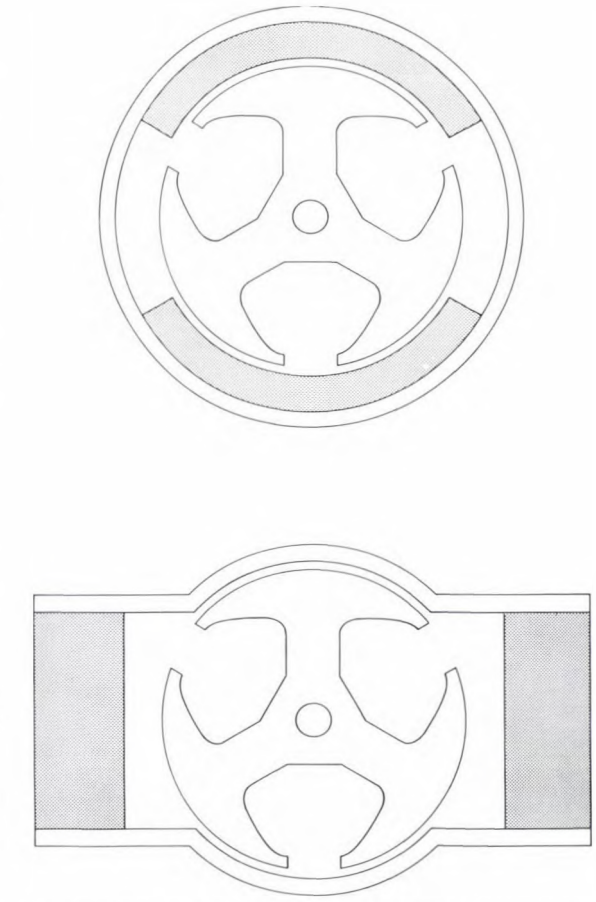


Fig. 7. Cross-section of a motor: *a*) with ferroxdure segments in a steel ring, *b*) with ferroxdure blocks and steel pole pieces.

the steel pole pieces of the magnet. The version with segment magnets gives an adequate flux in the rotor, but it is not easy to attach the magnets. Moreover, since there are air gaps between the magnets, the rotor has certain preferential positions, and this effect leads to vibration and noisy operation. The version with blocks and steel pole pieces is still used in the motors of some shavers powered by dry batteries or rechargeable cells. This motor also has preferential positions so that vibration and noise cannot be completely avoided.

At the moment, a ring of plastoferrite would seem to be the best solution. The dimensional accuracy of this material is very high; the magnet can therefore be used as part of the motor structure, and the air gap can be made very small. The magnets are injection-moulded. During injection, the magnetic particles in the binder are aligned in a magnetic field. The magnet is thus already magnetized when it leaves the mould; this process gives a low spread in the available flux. The magnetic field in the mould can be shaped in such a way that the distribution of the magnetization creates only weakly preferential positions. The flux obtained in the rotor is in fact nearly as high as that with segments of aligned ferroxdure, even though plastoferrite has a lower remanence.

Plastoferrite magnets are relatively cheap. Their only disadvantage is that they are relatively fragile; the construction of the motor must be compatible with this fragility.

Rotor design

The rotor design is based on the specification for the motor. The supply voltage and the output power will always be stated here, so that the current I will be more or less fixed. The slope of the torque-speed curve is also often specified. This is connected with the voltage IR across the rotor resistance: the slope increases with decreasing R. The copper losses I^2R in the rotor are of importance for the efficiency and the heat dissipation in the motor, which again is generally laid down in the specification. It follows that most of the requirements placed on the motor amount to a specification of the maximum permissible value of the diametral rotor resistance R.

A simplified description will now be given [2] of a procedure that starts from this value of the diametral rotor resistance, and determines the rotor dimensions for minimum motor volume; these correspond approximately to the minimum iron losses. The first step is to express the diametral rotor resistance in terms of the rotor dimensions.

The resistance R is the resultant of the parallel combination of two rotor paths, one consisting of one coil of resistance R_1 and the other of two coils in series with a total resistance of $2R_1$; it follows that $R = \frac{2}{3}R_1$. Each rotor coil has $N/3$ turns of average length s. If the resistivity of the copper is ϱ and the cross-sectional area of the wire q, it follows that

$$R = \frac{2\varrho sN}{9q}. \tag{22}$$

Of the total cross-sectional area of the rotor (see fig. 2), only the part A is available for the coil windings. Of this, only a fraction fA is available for the copper core of the wires ($2Nq$ for the rotor as a whole), since space is also required for the insulation, etc.; f is the space factor. With the aid of these data, q can be expressed in terms of the rotor dimensions if N is known. The number of turns N also depends on the specified supply voltage V, the reference speed n_0 mentioned earlier and the magnetic flux $\hat{\Phi}$ through one tooth of the rotor:

$$V = 2n_0N\hat{\Phi}. \tag{11}$$

R can therefore be written:

$$R = \frac{\varrho sV^2}{9fAn_0^2\hat{\Phi}^2}. \tag{23}$$

It now remains to express s, A and $\hat{\Phi}$ in terms of the rotor dimensions, which requires a closer look at the shape of the rotor teeth. Each tooth has straight parallel edges, and if we call the width of a tooth b and the length of the rotor l, then the iron cross-section of a tooth is bl; we may thus write for the flux:

$$\hat{\Phi} = bl\hat{B}, \tag{24}$$

where $\hat{B}$ is the maximum flux density in the rotor tooth. The average length s of a turn is $2b + 2l + 4c$; b and l are the lengths of the straight sections, and c the length of a curved section. The area A available for the windings in the total cross-sectional area of the rotor decreases as the tooth width b increases. The total cross-sectional area of the rotor is $\frac{1}{4}\pi d^2$, where d is the rotor diameter; A is a fraction of this area which decreases with increasing b as a function α of b/d:

$$A = \frac{\pi}{4}d^2\alpha(b/d). \tag{25}$$

Making use of the expressions for s, A and $\hat{\Phi}$, we can write for the resistance:

$$R = \frac{8}{9\pi}\frac{\varrho}{f\hat{B}^2}\frac{b+l+2c}{d^2b^2l^2\alpha(b/d)}\frac{V^2}{n_0^2}. \tag{26}$$

[2] A detailed description of the dimensioning of the rotor of a commutator motor is given in R. H. Dijken, Optimization of small AC series commutator motors, thesis, Eindhoven 1971 (also published as Philips Res. Repts. Suppl. 1971, No. 6).

The third factor in this expression contains the rotor dimensions. It is these that we are interested in; we bring the most important dimension, the rotor diameter d, to the left-hand side of the equation, and normalize the other rotor dimensions in terms of d:

$$b = \beta d, \qquad l = \lambda d, \qquad c = \gamma d. \tag{27}$$

We then obtain:

$$d^5 = \frac{8}{9\pi} \frac{\varrho}{f\hat{B}^2} \frac{\beta + \lambda + 2\gamma}{\lambda^2} \frac{1}{\beta^2 \alpha(\beta)} \frac{V^2}{n_0^2 R}. \tag{28}$$

This expression is grouped in such a way as to permit the effect of the various parameters to be easily distinguished. The first factor is a constant, the second contains material properties (ϱ and $\hat{B}$) and reflects the manufacturing method used (f). The third factor includes $\beta + 2\gamma$. It is found that, owing to changes in the shape of the rotor, γ decreases as β increases, so that $\beta + 2\gamma$ remains fairly constant. The third factor thus mainly reflects the effect of λ. The fourth factor reflects the effect of β, and the fifth factor indicates the performance required of the motor.

The variation of the required rotor volume with β and λ can easily be derived from (28). If λ is kept constant the rotor volume is proportional to d^3. The third factor in expression (28) is then practically constant; the variation of the fourth factor with β is such that d^5, and hence also d^3, goes to a minimum at a value of β of about 0.45 (see *fig. 8*). This is the optimum value for small rotor volume; if the thickness of the insulation in the rotor slot is also taken into consideration, the optimum value of β is found to be a little smaller (about 0.40). In most existing small d.c. motors with three teeth, the value of β is less than 0.15, well away from the optimum.

If on the other hand β is kept constant and λ varied, the required rotor volume is proportional to $d^3\lambda$. It follows from (28) that this quantity has a minimum at $\lambda = (\beta + 2\gamma)/2$ (see *fig. 9*). The minimum becomes more pronounced as $\beta + 2\gamma$ decreases.

For a rotor with three teeth, $\beta + 2\gamma$ has a value of 0.4 to 0.6. The corresponding optimum value of λ is thus 0.2 to 0.3. If the rotor has more than three teeth, the length of the coil ends must be given by another quantity, since the coils are wound in a different way; the value of this quantity is about 1, and the optimum value of λ is then about 0.5.

With a three-toothed rotor, for which $\beta + 2\gamma$ therefore has the lowest value, it is particularly important to choose the optimum value of λ. In existing small d.c. motors, λ is generally between 0.5 and 2 — i.e. far above the optimum range of 0.2 to 0.3. Our optimum dimensioning method gives rotors that are shorter and have a larger diameter than is usual.

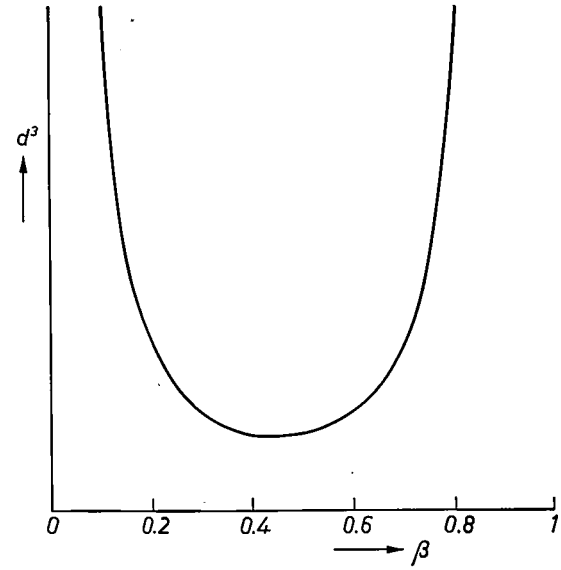


Fig. 8. The required rotor volume (proportional to the third power of the diameter d) has a minimum at a certain ratio β of tooth width to rotor diameter. The ratio of rotor length to rotor diameter is assumed to be constant.

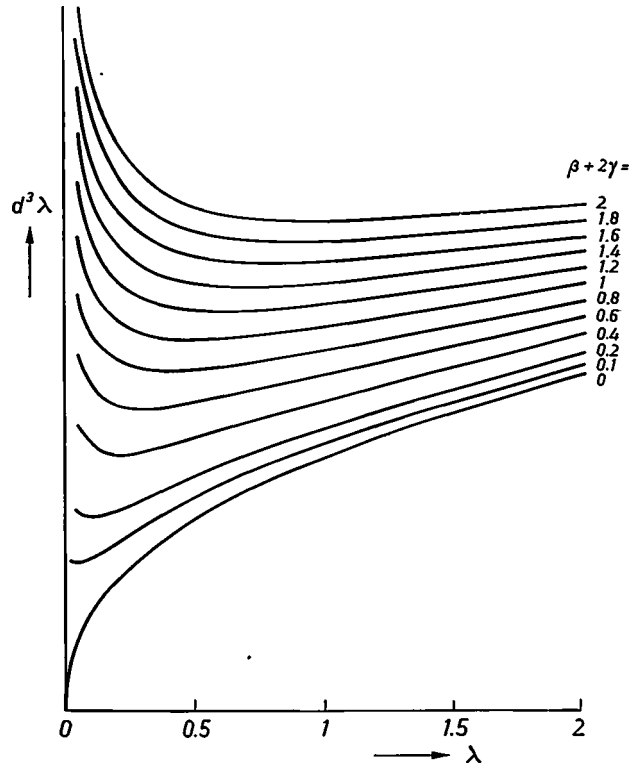


Fig. 9. The required rotor volume (proportional to $d^3\lambda$) as a function of the ratio λ of length to diameter. There is a minimum at $\lambda = (\beta + 2\gamma)/2$; γ is a measure of the length of the curved portions of a winding.

Design limitations

In practice, there may be limitations on the design that make it impossible to apply the optimum dimensioning described here. Certain dimensions are often limited by the nature of the application. For example, the external diameter of the motor shown in fig. 1 could not exceed 26 mm in connection with its application

in toothbrushes and certain shavers. This diameter had to include a 1.5-mm steel outer sheath to complete the magnetic circuit, a plastoferrite ring which had to be 2.8 mm thick (partly for mechanical strength) and an air gap of 0.3 mm; this left no more than 16.8 mm for the rotor diameter.

Besides the dimensional limitations of this kind, there is also a general limitation connected with the available flux of the magnet. In general, it is desirable to magnetize the rotor teeth to saturation, i.e. to about 1.5 teslas. To produce this magnetic flux density in rotor teeth with optimum dimensions ($\beta = 0.4$ and $\lambda = 0.25$), the remanence of the magnet material must be at least 0.4 T, which is appreciably more than can now be obtained with plastoferrite (about 0.25 T).

As a result of the two limitations mentioned above, it is impossible to give the rotor the theoretical minimum volume. This is also the case with the motor of fig. 1, but a step in the right direction has been taken here by giving β the maximum value permitted by the magnet, which was 0.22 as compared with less than 0.15 for most comparable motors.

The rotor volume required could only be obtained by making the length-to-diameter ratio of the rotor greater than the optimum value, since the diameter was fixed. The length of the rotor was set at 10 mm. The stator magnet was also made longer than the rotor; it projected 6 mm at each end, so that an appreciable flux was also provided by the projections. In this motor, $\lambda = 0.6$ (twice the optimum value from fig. 9 for $\beta + 2\gamma = 0.6$).

In a version of this motor designed for a 4.8-V supply, $G = 0.04$; this means a maximum efficiency of 64% for $a = 4$ (see fig. 5); a power of more than 1 watt is delivered at 6500 rev/min. If the maximum permissible diameter and the available magnetic material had not set certain limitations, it would have been possible to develop a motor optimized to the theory described in this article. A number of characteristics of the actual and optimum motors are compared in *Table I.*

Table I. Comparison of the motor of fig. 1, whose design was subject to certain restrictions, and an optimum design.

Quantity	Actual motor	Optimum motor
Rotor diameter	16.8 mm	18 mm
Stack height of rotor laminations	10 mm	5.4 mm
λ	0.6	0.3
β	0.22	0.4
Height of coil end	3 mm	1.5 mm
Rotor length, including coil ends	18 mm	10.4 mm
Remanence of magnet material	0.25 T	0.4 T
Projecting length of magnet ring	6 mm	6 mm
Magnet length	22 mm	16.4 mm
Total motor length	31 mm	23.4 mm
Total motor diameter	26 mm	27.2 mm
Rotor volume	16.5 cm^3	13.6 cm^3

Summary. More than a hundred million small d.c. commutator motors are produced every year throughout the world. The most important point in the design of such motors is often the efficiency, since they are generally battery-powered. The efficiency can be expressed by means of two dimensionless parameters, whose values depend on the motor characteristics and on the torque required by the load. The speed as a function of the torque is also given by these two parameters. Two important aspects of the design are the choice of the magnetic material for the stator and of the rotor dimensions. Optimum dimensioning can be obtained by making the motor as small as possible without adversely affecting the efficiency; the rotor is then shorter than usual for a given diameter, and has wider teeth.

Applications of stepping motors

A stepping motor responds to a current pulse with a small, well defined angular rotation [1]. In this way digital information — the number of pulses in a pulse train — is converted into a proportional mechanical displacement. A stepping motor is therefore a useful device for coupling digital data processing with machine tools. Since it accurately reproduces angular rotations, it can be used in open control systems, i.e. systems without feedback. The speed can be adjusted over a wide range.

Stepping motors have many and various applications. Three typical examples are given here, in which the motor speed is required to be accurately proportional to the number of pulses supplied per second. The first two examples relate to the controlled supply of liquids. The third relates to the slide movement for a numerically controlled machine tool; accurate positioning of the workpiece is also necessary in this application.

The 'Precidor' infusion pump made by Infors AG of Binningen, Switzerland (*fig. 1*) is used for the accurate dosing of prolonged infusions in medical treatment. The stepping motor (Philips ID 04) drives a lead-screw, which converts the rotary movement into a slow linear movement of the piston in a syringe. Syringes of different sizes can be fitted to the machine, including the sterile-packed disposable syringes now in common use.

The pulses for the stepping motor are produced by a stabilized oscillator with a unijunction transistor. The frequency of this oscillator, and hence the piston speed, is continuously variable over a decade. The frequency can also be divided twice by 10 in a divider, giving a

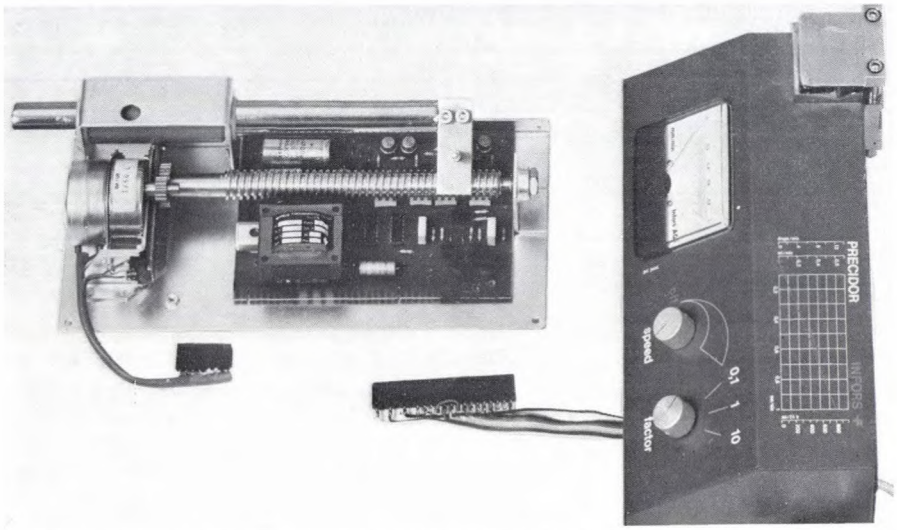


Fig. 1. 'Precidor' infusion pump, type 5003, made by Infors. *Left:* mechanism with Philips ID 04 stepping motor; *right:* the case. The motor drives a lead-screw, which transmits a linear motion to the piston of a syringe (not shown in the photograph). (Photo Infors AG, Binningen, Switzerland.)

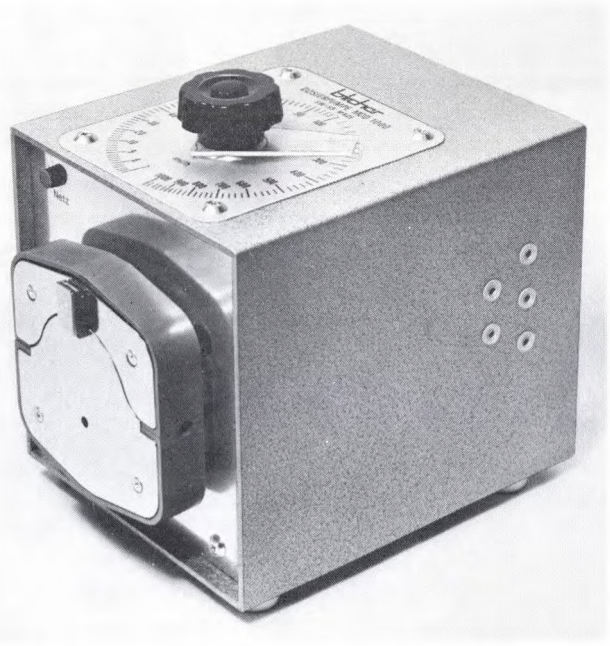


Fig. 2. Dose pump made by Bircher. A star-section rotor with pressure rollers squeezes a silicone-rubber tube as it rotates, forcing the fluid along the tube. The drive is provided by a Philips ID 06 stepping motor. (Photo Max Bircher Ingenieur, Schaffhausen, Switzerland.)

[1] See E. M. H. Kamerbeek, Electric motors, Philips tech. Rev. **33**, 215-234, 1973.

range of piston speeds from 0.01 mm to 10 mm per minute. The piston speed can be read off from a moving-coil meter controlled by a pulse counter.

The dose pump made by Bircher (Schaffhausen, Switzerland) has a similar function (*fig. 2*). A stepping motor (Philips ID 06) drives a star-section rotor with pressure rollers. When the rotor rotates, the rollers squeeze a silicone-rubber tube, forcing the liquid along it. As in the Infors infusion pump, there is no control feedback. The pump has a dosing accuracy of about 2% and a wide flow range (7 cm^3 to 1000 cm^3 an hour). The dose pump is used in laboratories, in the chemical and pharmaceutical industries and in hospital intensive-care units.

Fig. 3 shows a numerically controlled vertical milling machine made by Terco AB of Skärholmen (Sweden). Three stepping motors (Philips PD 20) displace the slide in three perpendicular directions. The drive unit, on the right, provides the stator coils with currents that correspond to the control input, and ensures that the motors deliver a constant torque. The control signals come from a punched-tape reader and a recording unit (not shown). The position of the stepping motors is monitored by optical sensors; their output signal is fed back to the input — in a closed-loop control system — and also supplies the data for a digital read-out of the position of the workpiece. Each stepping motor is coupled to a 50 : 1 reduction gear that drives a lead-screw for the slide displacement. The workpiece speed (the 'feed') can be adjusted in all directions in 10 steps between 0.5 mm/s and 0.05 mm/s. For a workpiece displacement of 50 mm the maximum position error is $\pm$ 0.02 mm. The whole system was specially designed for industrial training courses.

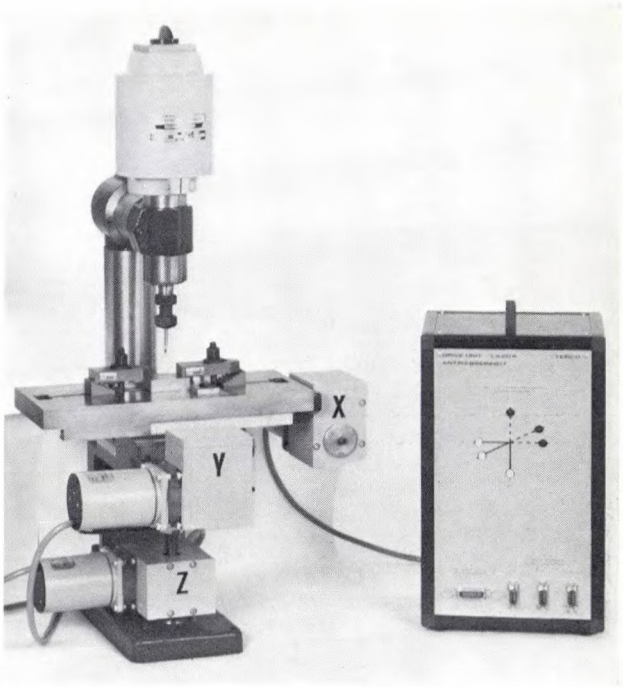


Fig. 3. Numerically controlled vertical milling machine, made by Terco, with drive unit. Three stepping motors (Philips PD 20) displace the workpiece along the *x*-, *y*- and *z*-axes. (Photo Terco AB, Skärholmen, Sweden.)

Steady-state performance of a class of electronically commutated d.c. machines

W. Radziwill

Introduction

In conventional d.c. machines the time during which the commutated armature windings are short-circuited is determined by the relative dimensions of the brushes and the commutator segments; it varies only with the rotor speed. With electronically commutated machines [1], however, the short-circuit time can be varied by control of the semiconductor switches. In addition, the current can also be interrupted periodically in this way, so that new and interesting possibilities become available for controlling the speed and torque [2] [3]. Such control methods are of course more easily applied if a theory is available that gives a satisfactory description of the behaviour of the machine in this kind of operation.

In the case of periodic interruption of the current the behaviour of the machine can only be treated numerically. If the current is not interrupted but only the short-circuit time is controlled, it is still possible to obtain expressions in closed form for the torque, the input power and the output power. Moreover, the efficiency can be determined by a simple graphical procedure.

These expressions and quantities will be derived for d.c. machines satisfying the following three conditions:
— The excitation flux is constant with time (e.g. permanent-magnet rotor) and the flux linkage with the armature winding varies sinusoidally with the angular position of the rotor.
— The stator winding is divided into m phases with an identical number of turns and each phase is connected to the supply voltage via four controlled switches, as shown in *fig. 1*.
— The switches are only allowed to take up the combinations shown in *Table I*. The time interval I_1 is identical to I_3 and I_2 is identical to I_4.

The theory derived for these machines can also be used to give a first approximation in the more general case with periodic interruption of the current [2] [3].

Considerable simplification is obtained by confining the investigation to steady-state operation — i.e. with only gradual changes of load and disregarding acceleration transients: the angular velocity can then be taken as constant over a switching period I_1-I_4 and the machine always rotates at such a speed that the period of the voltage induced in the windings by the permanent magnet coincides exactly with the switching period mentioned above — a situation necessarily imposed by the conventional commutator but which need not be so with the more general control available with electronic commutation.

Dipl.-Ing. W. Radziwill is with Philips Forschungslaboratorium Aachen GmbH, Aachen, West Germany.

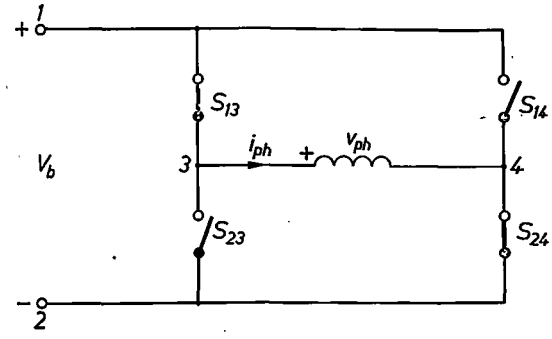


Fig. 1. Schematic circuit diagram for one phase winding of an electronically commutated d.c. machine. V_b supply voltage. v_{ph} voltage across phase winding. i_{ph} armature current in phase winding. S electronically controlled switches.

First of all, the differential equation describing the voltage equilibrium for one phase will be solved for the armature current. The resulting expression will then be used to describe the power balance of the machine and to derive its speed-torque characteristic.

It will be shown that higher harmonics of armature voltage and current do not contribute to the mechanical output power. The steady-state behaviour of this class of machines can therefore be represented by locus diagrams of the vector symbolizing the fundamental of the armature current in the complex plane. It will also be shown that losses due to higher harmonics are independent of special machine parameters, so that a general plot of these loss functions can be used in combination with the locus diagrams to provide a quick and simple evaluation of the motor efficiency.

Table I. Positions of the four switches S_{13}-S_{24} in fig. 1 during the four successive adjacent time intervals I_1-I_4 of each rotation of the rotor over two pole-pitches.

	I_1	I_2	I_3	I_4
S_{13}	on	on	off	off
S_{14}	on	off	off	on
S_{23}	off	off	on	on
S_{24}	off	on	on	off

The armature current

To calculate the torque and the power as a function of the speed of rotation it is first necessary to know the armature current as a function of angular position over one pole pitch. In the derivation of this current we shall assume that the magnetic saturation can always be neglected. It is further assumed that the mutual inductance between the stator phases is zero (in two-phase machines) or negligibly small (the reluctance of the magnetic circuit is high because of the low permeability of the permanent-magnet rotor).

Under the conditions imposed earlier we then have the simple equivalent circuit shown in *fig. 2*. R and L are the resistance and the inductance per phase and e, v_{ph} and i_{ph} are the instantaneous values of the induced voltage (or back e.m.f.), the applied voltage (via the switches) and the armature current, respectively (see fig. 1).

If $N\hat{\Phi}$ is the peak value of the flux of the permanent magnet linked by a phase winding, p the number of pole pairs and x the instantaneous electrical angular displacement of the rotor (x/p is the actual physical angular displacement), then we can write the sinusoidal change of the linked flux as

$$N\Phi = -N\hat{\Phi}\cos(x-b), \tag{1}$$

where b is a phase displacement that will be defined later. Using the relation between the speed of rotation n and the electrical angular frequency

$$\omega = \dot{x} = 2\pi pn, \tag{2}$$

and applying the induction law, we can obtain the induced voltage

$$e = \frac{dN\Phi}{dx}\dot{x} = \omega N\hat{\Phi}\sin(x-b). \tag{3}$$

The waveform of the armature voltage v_{ph} follows from Table I and fig. 1. The relationship between the induced voltage e and the armature voltage v_{ph} is shown in *fig. 3*. The phase displacement b is fixed by the adjustment of the electronic commutator and can in principle have any value. The angle γ during which the armature voltage remains 'on' must lie between 0 and π; by varying γ within these limits the machine can be controlled.

To obtain an expression for the armature current i_{ph} we now solve the linear differential equation describing the circuit of fig. 2. Since $L\,di_{ph}/dt = L\dot{x}\,di_{ph}/dx$, the time t can be eliminated and the equation written in a form dependent only on the angular displacement x:

$$\omega L\frac{di_{ph}}{dx} + Ri_{ph} = v_{ph} - \omega N\hat{\Phi}\sin(x-b). \tag{4}$$

To bring this equation to the most general form we introduce the dimensionless quantities

$$v = v_{ph}/V_b, \qquad y = i_{ph}/(V_b/R) \tag{5}$$

and the constant coefficients

$$a = \omega L/R, \qquad c = N\hat{\Phi}/(LV_b/R). \tag{6}$$

The equation then becomes:

$$ay' + y = v(x) - ac\sin(x-b). \tag{7}$$

In accordance with fig. 3, $v(x)$ is now defined as:

$$v(x) = \begin{cases} 0 & \text{for } 0 < x \leqslant \pi - \gamma \\ 1 & \text{for } \pi - \gamma < x \leqslant \pi \end{cases} \tag{8}$$

with the symmetry condition:

$$v(x) = -v(x+\pi). \tag{9}$$

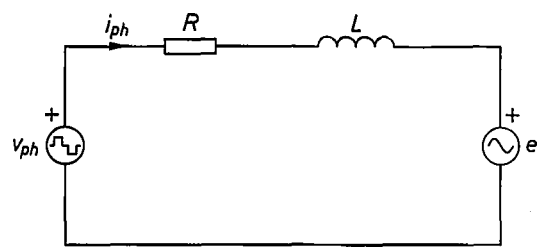


Fig. 2. Equivalent circuit of a stator phase. e induced voltage (or back e.m.f.). R resistance. L inductance.

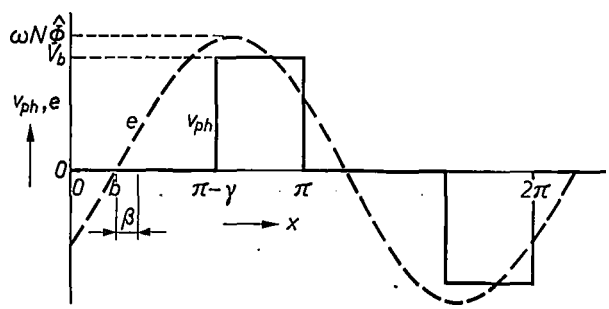


Fig. 3. The voltage v_{ph} across a phase winding and the induced voltage e as functions of the electrical phase angle x. b phase angle. β phase difference between the induced voltage and the fundamental of the supply voltage. ω angular frequency. $N\hat{\Phi}$ maximum linked flux.

[1] The principle of these motors has been described in E. M. H. Kamerbeek, Electric motors, Philips tech. Rev. 33, 215-234, 1973. However, in the motors discussed in that article the current can only be switched on and off, not reversed.

[2] W. Radziwill, A highly efficient small brushless d.c. motor, Philips tech. Rev. 30, 7-12, 1969.
W. Radziwill, Theorie des Betriebsverhaltens kleiner elektronisch kommutierter Gleichstrommaschinen mit Drehzahlregelung durch Variation der Pulsbreite und Pulsphase der Ankerspannung, Ingenieur 84, E 41-49, 1972.

[3] W. Radziwill, Elektronische Kommutierung bei kleinen permanentmagnetisch erregten Gleichstrommaschinen, in: Philips — Unsere Forschung in Deutschland, Part II, Aachen/Hamburg 1972, pp. 238-244.

Solving (7), e.g. by Laplace transforms, we obtain the normalized armature current per phase:

$$y = -\frac{ac}{1+a^2}\{\sin(x-b) - a\cos(x-b)\} + \begin{cases} \dfrac{\exp(-\gamma/a)-1}{1+\exp(-\pi/a)}\exp(-x/a) & \text{for } 0 < x \leqslant \pi-\gamma. \\ 1 - \dfrac{1+\exp\{-(\pi-\gamma)/a\}}{1+\exp(-\pi/a)}\exp\{-(x-\pi+\gamma)/a\} & \text{for } \pi-\gamma < x \leqslant \pi. \end{cases} \quad (10)$$

We are now in a position to derive expressions for the input power, the output power and the torque.

Power balance

If (4) is multiplied by i_{ph} the power balance per phase is obtained. The first term is evidently the reactive power, the second the losses, the third the input power and the fourth the output power. The total mean power is obtained by taking the mean over a half-period and multiplying this by the number of phases m. The first term then vanishes and for the input power we find:

$$P_i = \frac{m}{\pi}\int_0^{\pi} v_{ph} i_{ph} dx = \frac{mV_b}{\pi}\int_{\pi-\gamma}^{\pi} i_{ph} dx = \frac{mV_b^2}{R\pi}\int_{\pi-\gamma}^{\pi} y dx,$$

and for the output power:

$$P_o = \frac{m}{\pi}\int_0^{\pi} i_{ph}\omega N\bar{\Phi}\sin(x-b)dx = \frac{mV_b^2}{R}\frac{ac}{\pi}\int_0^{\pi} y\sin(x-b)dx.$$

Substituting (10) for the normalized armature current y and putting

$$\beta = \frac{\pi-\gamma}{2} - b \quad (11)$$

to simplify the expressions, we have finally:

$$P_i = \frac{mV_b^2}{R}\left\{\frac{\gamma}{\pi} - \frac{a}{\pi}\frac{1-\exp(-\gamma/a)+\exp\{-(\pi-\gamma)/a\}-\exp(-\pi/a)}{1+\exp(-\pi/a)} - \frac{2}{\pi}\sin(\gamma/2)\frac{ac}{1+a^2}(\cos\beta + a\sin\beta)\right\}, \quad (12)$$

$$P_o = \frac{2}{\pi}m\frac{V_b^2}{R}\sin(\gamma/2)\frac{ac}{1+a^2}\left\{\cos\beta - a(\sin\beta + \frac{\pi}{4}\frac{c}{\sin(\gamma/2)})\right\}. \quad (13)$$

Differentiating (13) with respect to the independent quantities c and β gives the maximum output:

$$P_{om} = \frac{2m}{\pi^2}\frac{V_b^2}{R}\sin^2(\gamma/2), \quad (14)$$

which can be obtained for $a = -\tan\beta$ provided that the condition $c = -(2/\pi)\sin(\gamma/2)/\sin\beta$ is satisfied by suitable adjustment of the machine parameters. This is in general not to be recommended since — as will appear later — the efficiency obtained in this case is not very high. Nevertheless it is advisable, to show the essential relations more clearly, to normalize all the powers with respect to P_{om}. The normalized output power is then

$$P_{on} = \frac{P_o}{P_{om}} = \frac{\pi c\cos\beta}{\sin(\gamma/2)}\frac{a}{1+a^2}\left(1-\frac{a}{a_0}\right), \quad (15)$$

with

$$a_0 = \frac{\cos\beta}{\sin\beta + \frac{\pi}{4}\frac{c}{\sin(\gamma/2)}}. \quad (16)$$

It will be useful before going further to comment on the significance of the various parameters. As can be seen from (11) and fig. 3, β is the phase difference between the induced voltage e and the fundamental of the pulse voltage v_{ph}. Provided γ is not varied, β has

a constant value given by the setting of the electronic commutator, equivalent to the brush displacement in a conventional commutator machine. (If the machine were to be used without a commutator as a synchronous machine, supplied by an a.c. voltage of fixed frequency, so that it ran at a fixed speed, then β would be variable, denoting the load angle between the a.c. supply voltage and the induced voltage.) When the motor is run as a d.c. machine the quantity a is the load-dependent variable [3].

As can be seen immediately from (2) and (6), a can be regarded as a normalized speed of rotation:

$$a = \frac{n}{R/2\pi pL}.$$

This may be positive or negative, depending on the direction of rotation. The important machine constant c, as equation (6) shows, gives the ratio of the peak rotor flux linked by a phase winding to the flux produced in the phase winding by the current V_b/R. It is always positive, since any reversal of the voltage V_b can be dealt with mathematically not by a negative sign for V but by substituting $\beta \pm \pi$ for β instead.

A further study of the expression (15) for the normalized output power enables the speeds to be found at which the machine operates as a motor ($P_{on} > 0$) and at which it operates as a generator ($P_{on} < 0$).

If $a_0 \neq 0$, there are two values of a for which (15) is zero: $a = 0$ and $a = a_0$. The condition $dP_{on}/da = 0$ then yields two speeds at which the normalized power is at a turning point. These speeds are

$$a_{Pe1} = \tan(\tfrac{1}{2}\arctan a_0) \quad \text{and} \quad a_{Pe2} = \tan(\tfrac{1}{2}\arctan a_0 - \pi/2).$$

The turning point is

$$P_{on,e} = \frac{\pi}{2}\frac{c\cos\beta}{\sin(\gamma/2)}\, a_{Pe}.$$

Evidently a_{Pe1} lies in the interval $[0,a_0]$; a_{Pe2} has the opposite sign and is only reached when the direction of rotation is reversed. At the value of a_{Pe} for which $(\cos\beta)/a_{Pe} > 0$, $P_{on,e}$ is a maximum. This follows from the negative sign of d^2P_{on}/da^2 at $a = a_{Pe}$. The behaviour of P_{on} as a function of speed is also characterized by three points of inflexion at

$$a = -\tan(\tfrac{1}{3}\operatorname{arccot} a_0) \quad \text{and} \quad a = -\tan(\tfrac{1}{3}\operatorname{arccot} a_0 \pm \pi/3),$$

where $d^2P_{on}/da^2 = 0$. Outside the region in which these speeds lie, the output power converges monotonically, for positive and negative a, from different sides to the limiting value

$$\lim_{|a|\to\infty} P_{on} = -\frac{\pi c}{a_0}\frac{\cos\beta}{\sin(\gamma/2)} = -\frac{\pi c}{\sin(\gamma/2)}\left(\sin\beta + \frac{\pi}{4}\frac{c}{\sin(\gamma/2)}\right).$$

To illustrate this P_{on} is plotted in *fig. 4* as a function of a for three cases.

The following conclusions can be drawn from the information already given. If the power has a maximum at a_{Pe1}, the machine operates as a motor in the interval $[0,a_0]$; outside this interval it operates as a generator. In this case a_0 is the no-load speed (fig. 4*a*). If the power has a minimum at a_{Pe1} the situation is reversed: in the interval $[0,a_0]$ the machine acts as a generator and outside it as a motor. Of the two motor regions in this case, only the one adjacent to $a = 0$, where the direction of rotation is opposite to that in the range $(0,a_0)$, is of any technical interest. Since the power has a maximum here at a_{Pe2} but there is no speed at which the power is zero, there is no no-load speed and the motor 'runs away' if the load is removed (fig. 4*b*).

Finally, let us consider the special case $a_0 = 0$, for which equation (16) shows that $\beta = \pm\pi/2$. The normalized output power here is

$$P_{on} = -\frac{a^2}{1+a^2}\frac{\pi c}{\sin(\gamma/2)}\left(\sin\beta + \frac{\pi}{4}\frac{c}{\sin(\gamma/2)}\right).$$

At $\beta = +\pi/2$, the machine evidently cannot operate as a motor since the two zero points and the maximum coincide at the origin and P_{on} is negative for all other values of a. At $\beta = -\pi/2$, the machine again functions only as a generator unless also

$$\frac{\pi}{4}\frac{c}{\sin(\gamma/2)} < 1.$$

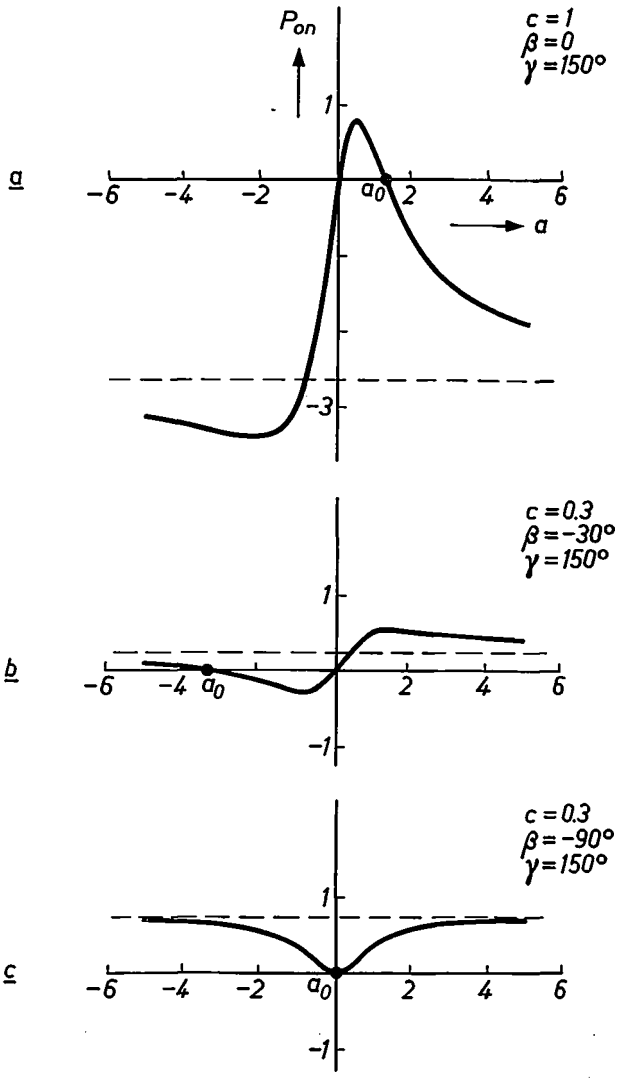


Fig. 4. Curves showing the normalized output power P_{on} as a function of the normalized speed a, for three different combinations of β and c; in all cases $\gamma = 150°$. The dashed line is the asymptote. *a*) The machine can function as a motor at speeds between zero and the no-load speed a_0. *b*) The machine functions as a motor at speeds $a > 0$ and can 'run away'. *c*) The special case $\beta = -\pi/2$. For the given values of c and γ the machine functions as a motor at all speeds and in both directions; the starting torque, however, is zero.

In this case the two zero points coincide with the minimum (fig. 4c). The machine does indeed operate then as a motor at all values of a, but is of no real interest because the starting torque is zero, as will appear below.

Torque-speed characteristics

The torque T is found by taking the derivative of the mechanical work $\int P_0 \, dt$ with respect to the angle of rotation x/p:

$$T = p\frac{P_0}{\dot{x}} = \frac{P_0}{2\pi n}.$$

Substituting from (13) and normalizing with respect to the maximum value that the expression could assume at the speed $a = 0$, if β and γ were variable,

$$T_{0m} = \frac{2}{\pi} pm \frac{V_b}{R} N\hat{\Phi}, \qquad (17)$$

which is the maximum starting torque, we obtain the normalized torque

$$T_n = \frac{T}{T_{0m}} = \frac{\sin(\gamma/2)\cos\beta}{1+a^2}\left(1-\frac{a}{a_0}\right). \qquad (18)$$

From (15) and (18) a simple relation between T_n and the normalized power can be obtained:

$$P_{on} = \frac{\pi c}{\sin^2(\gamma/2)} T_n a.$$

The machine therefore operates as a motor if $T_n a > 0$. This means that the characteristic curves of T_n as a function of a lie in the first and the third quadrants for operation as a motor, and in the second and fourth quadrants for operation as a generator.

Fig. 5 shows some typical examples of such curves; these examples all refer to motors developed and tested at PFA. It can be seen immediately that the characteristic of a shunt-wound machine ought not to be expected automatically, even though it would seem obvious in the case of a machine with a constant magnetic excitation [6]. Indeed, curve *4* would be typical for a series-wound motor; it represents the case mentioned earlier of a motor that can 'run away' and thus where a_0 is not the no-load speed.

Expression (18) for the normalized torque is zero, in the case just referred to, only for the single value $a = a_0$. It has two turning points, at

$$a_{Te1} = -\tan(\tfrac{1}{2}\operatorname{arccot} a_0)$$

and

$$a_{Te2} = -\tan(\tfrac{1}{2}\operatorname{arccot} a_0 - \pi/2).$$

The turning point is a maximum if the following conditions are satisfied:

	$\cos\beta > 0$	and	$\lvert a_{Te}\rvert < 1$,
or	$\cos\beta < 0$	and	$\lvert a_{Te}\rvert > 1$,
or	$\beta = -\pi/2$	and	$a_{Te} = 1$,
or	$\beta = \pi/2$	and	$a_{Te} = -1$.

In addition there are generally three points of inflexion at $a = \tan(\frac{1}{3}\arctan a_0)$ and $a = \tan(\frac{1}{3}\arctan a_0 \pm \frac{1}{3}\pi)$ and a limiting value $\lim_{|a|\to\infty} T_n = 0$.

In practice it is chiefly the decreasing portions of the curves that are of importance, since stable steady-state operation is possible only at working points where dT_n/da is negative. This is because the demand for more power when the load suddenly increases slightly causes a drop in speed, since it has to be temporarily supplied from the kinetic energy of rotation, which implies that a new equilibrium can be set up only when the lower speed corresponds to an increased torque.

The stability condition $dT_n/da < 0$ must also be satisfied at the no-load speed a_0. From this it follows that a_0 is the no-load speed only when

$$\sin\beta + \frac{\pi}{4}\frac{c}{\sin(\gamma/2)} > 0.$$

In this case the maximum of the torque must lie in the second quadrant and the minimum in the fourth quadrant. If the condition is not satisfied, we have a machine that can 'run away', with the maximum lying in the first quadrant and the minimum in the third quadrant. Between the starting torque $T_0 = \sin(\gamma/2)\cos\beta$ and the maximum there is a region of instability; this is however of little importance if the motor can be started with a load that is less than the starting torque.

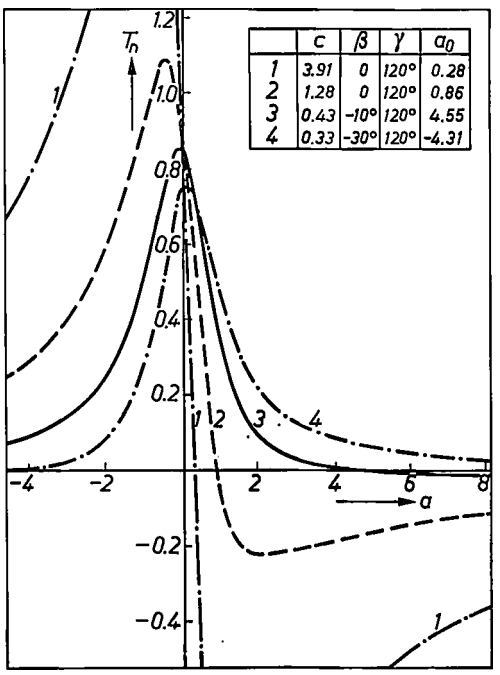


Fig. 5. Normalized torque-speed characteristics. Curve *1* refers to the motor in the horizon sensor of the Netherlands astronomical satellite ANS [4]. Curve *2* refers to the attitude-control motors in the ANS [5]. Curve *3* relates to an experimental motor for a tape recorder ($V_b = 9$ V, $N\hat{\Phi} = 5.7$ mWb, $R = 10.2\ \Omega$, $L = 15$ mH), and curve *4* to a motor that can 'run away', for which the starting torque is less than the maximum torque.

It can now be seen why the possibility mentioned above of motor operation at all speeds when $\beta = -\pi/2$ and $\frac{1}{4}\pi c/\sin(\gamma/2) < 1$ is of no interest. In this case, from (18) and (16): $T_n = \{\sin(\gamma/2) - \frac{1}{4}\pi c\}a/(1 + a^2)$. This expression is zero for $a = 0$ and also has a positive slope in the neighbourhood of this point. Such a machine is thus initially unstable when starting from rest in either direction.

In addition, it has no finite no-load speed. On the other hand it is possible to think of practical applications for a machine working only as a generator at $|\beta| = \pi/2$, and producing a braking torque proportional to the speed in the neighbourhood of the zero.

Losses due to higher harmonics

If the losses of the machine are simply taken as the difference $P_i - P_o$ of (12) and (13) and the efficiency as the ratio P_o/P_i, the expressions obtained are very complicated and it is difficult to see how these important quantities depend on the parameters of interest. It is possible, however, to split the losses into those due entirely to the higher harmonics and independent of the machine parameters β and c, and those due to the fundamental and dependent on the machine parameters. The r.m.s. value of the fundamental of v_{ph} is given by:

$$V = \frac{2\sqrt{2}}{\pi} V_b \sin(\gamma/2). \tag{19}$$

In the differential equation (4), we can now introduce the fundamental

$$v_1 = \sqrt{2}\, V \sin(x - b - \beta) \tag{20}$$

instead of v_{ph}, and i_1 instead of i_{ph}. We then find the solution for the fundamental component of the current:

$$i_1 = \sqrt{2}\,\frac{V}{R}\left\{\cos\alpha \sin(x-b-\beta-\alpha) - \frac{\pi}{4}\frac{c}{\sin(\gamma/2)}\sin\alpha\sin(x-b-\alpha)\right\}, \tag{21}$$

where

$$\alpha = \arctan a. \tag{22}$$

The contribution to the input power from the fundamental components of voltage and current is then:

$$P_{i1} = \frac{m}{\pi}\int_0^{\pi} v_1 i_1 dx = \frac{4P_{om}}{1+a^2}\left\{1 - \frac{\pi}{4}\frac{ac}{\sin(\gamma/2)}(\cos\beta + a\sin\beta)\right\}. \tag{23}$$

If this is subtracted from the total input power (12) and the result is divided by the maximum output power (14), we obtain the normalized losses due to the higher harmonics:

$$P_{hn} = \frac{P_i - P_{i1}}{P_{om}} = \frac{\pi}{2\sin^2(\gamma/2)}\left(\gamma - a\,\frac{1 - \exp(-\gamma/a) + \exp\{-(\pi-\gamma)/a\} - \exp(-\pi/a)}{1 + \exp(-\pi/a)}\right) - \frac{4}{1+a^2}. \tag{24}$$

This can also be expressed as a series:

$$P_{hn} = \frac{4}{\sin^2(\gamma/2)}\sum_{n=1}^{\infty}\frac{1}{(2n+1)^2}\,\frac{\sin^2\{(2n+1)\gamma/2\}}{1 + (2n+1)^2a^2}. \tag{24a}$$

It is not immediately obvious that these two expressions are identical. The expression (24a) is obtained by expanding $i_{ph} = (V_b/R)y$ as a Fourier series, where y is given by (10), and then solving

$$\frac{1}{P_{om}}\frac{1}{\pi}\int_0^{\pi}(i_{ph} - i_1)^2 R dx.$$

Fig. 6 shows P_{hn} plotted against the normalized speed a for various values of γ. The losses are a minimum at γ_{opt}, a quantity shown in *fig. 7* as a function of a. The curve of γ_{opt} is obtained by computing the minimum of P_{hn} numerically from the relation $dP_{hn}/d\gamma = 0$. Since γ_{opt} only varies over a small range, a fixed value of γ in this range gives an adequate approximation to the minimum losses from the higher harmonics for all speeds.

[4] P. van Dijk, The optical sensors of the Netherlands astronomical satellite (ANS), II. The horizon sensor, Philips tech. Rev. 34, 213-217, 1974 (No. 8).
[5] J. Crucq, The reaction wheels of the Netherlands satellite ANS, Philips tech. Rev. 34, 106-111, 1974 (No. 4).
[6] See E. M. H. Kamerbeek, Electric motors, Philips tech. Rev. 33, 215-234, 1973.

Locus diagram of the fundamental component of the current in the complex plane

If the contribution to the output power P_0 from the fundamental of the armature current is determined, the expression (13) is again obtained. This result indicates that the higher harmonics of the current i_{ph} and of the voltage v_{ph} make no contribution to the output power; it is not surprising, since the induced voltage is purely sinusoidal and only those harmonics of i_{ph} and e that are of the same frequency can make a contribution to the mean value

$$\frac{1}{\pi}\int_0^{\pi} e i_{ph} dx.$$

Apart from the losses, therefore, the machine behaves as if it were driven by a purely sinusoidal a.c. voltage. For the description of its speed-dependent behaviour

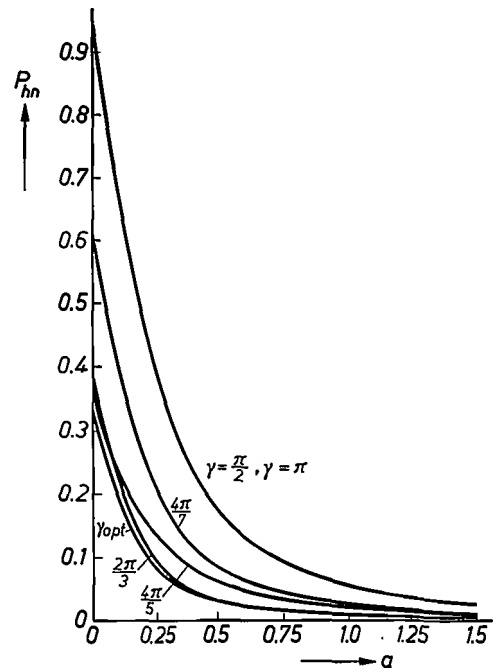


Fig. 6. The losses P_{hn} (normalized) that arise at higher harmonics as functions of the (normalized) speed a, for various values of the switching angle γ. The curve for γ_{opt} represents the theoretical minimum losses; γ_{opt} is itself a function of a (see fig. 7).

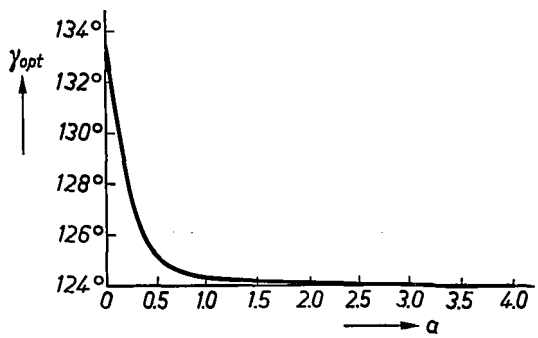


Fig. 7. The switching angle γ_{opt} for which the losses due to higher harmonics are a minimum. The angle γ_{opt} does not vary much with the (normalized) speed a.

it is therefore convenient to use locus diagrams of the position vector representing the fundamental of the current in the complex plane [7]. In this approach v_1 and i_1 are taken to be the imaginary parts of the complex quantities

$$v_1 = \mathrm{Im}\{\sqrt{2}\,\underline{V}\exp \mathrm{j}(x-b-\beta)\}, \quad (25)$$

$$i_1 = \mathrm{Im}\{\sqrt{2}\,\underline{I}\exp \mathrm{j}(x-b-\beta)\}. \quad (26)$$

Here $\underline{V}$ and $\underline{I}$ are the complex r.m.s. values of the fundamentals of voltage and current:

$$\underline{V} = V\exp \mathrm{j}0, \quad (27)$$

which is real by definition, and

$$\underline{I} = \frac{V}{R}\left\{\cos\alpha\exp(-\mathrm{j}\alpha) - \frac{\pi}{4}\frac{c}{\sin(\gamma/2)}\sin\alpha\exp \mathrm{j}(\beta-\alpha)\right\}. \quad (28)$$

If $\cos\alpha$ and $\sin\alpha$ in (28) are expressed by means of complex conjugates and the expression is normalized with respect to V/R, we have the desired locus diagram of the complex current for varying α:

$$\underline{Y} = \frac{\underline{I}}{V/R} = \frac{1}{2} + \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\exp \mathrm{j}\left(\beta+\frac{\pi}{2}\right) + \left\{\frac{1}{2} + \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\exp \mathrm{j}\left(\beta-\frac{\pi}{2}\right)\right\}\exp(-\mathrm{j}2\alpha). \quad (29)$$

This is a circle with centre at

$$\left(\frac{1}{2} - \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\sin\beta, \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\cos\beta\right)$$

and of radius

$$\left\{\left(\frac{1}{2} + \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\sin\beta\right)^2 + \left(\frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\cos\beta\right)^2\right\}^{1/2}.$$

Since, from (27), $\underline{V}$ is real, the real part Y_a of the normalized current ($\underline{Y} = Y_a + \mathrm{j}Y_r$) is the active component, and the imaginary part the reactive component of the current. The curves of constant power in the complex plane can now be determined very simply from the well known relations:

$$P_{i1} = mVI_a \quad \text{and} \quad P_0 = P_{i1} - m(I_a^2 + I_r^2)R. \quad (30)$$

If these expressions are normalized to $P_{om} = mV^2/4R$ (see (14) and (19)), so that

$$P_{i1n} = P_{i1}/P_{om} \quad \text{and} \quad P_{on} = P_0/P_{om}, \quad (31)$$

then for the curves of constant normalized power:

$$Y_a = \tfrac{1}{4}P_{i1n} \quad \text{and} \quad (Y_a - \tfrac{1}{2})^2 + Y_r^2 = \tfrac{1}{4}(1-P_{on}). \quad (32)$$

The input power from the fundamental component is thus constant on lines running parallel to the imaginary axis; the curves of constant output power are concentric

circles with the point $(\frac{1}{2}, 0)$ as centre. The circle representing the maximum power $P_{on} = 1$ coincides with this point since its radius is zero. The no-load power $P_{on} = 0$, and also the output power at rest ($a = 0$), both lie on the circle of radius $\frac{1}{2}$.

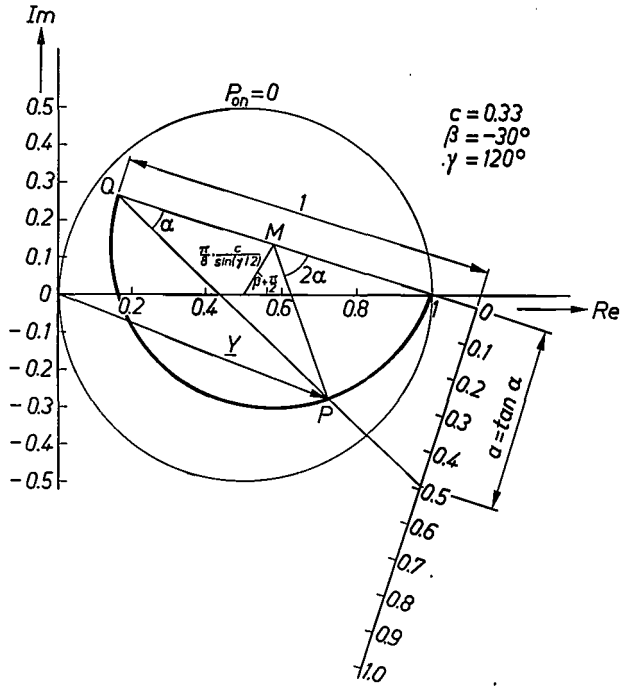


Fig. 8. Locus diagram of the quantity $\underline{Y}$ in the complex plane; Y is the normalized r.m.s. value of the fundamental component of the current. The working point P describes the semicircle from (1, 0) to Q as the normalized speed a runs from 0 to ∞. This speed $a = \tan\alpha$ can be read off the scale. This curve refers to the same motor as curve 4 in fig. 5.

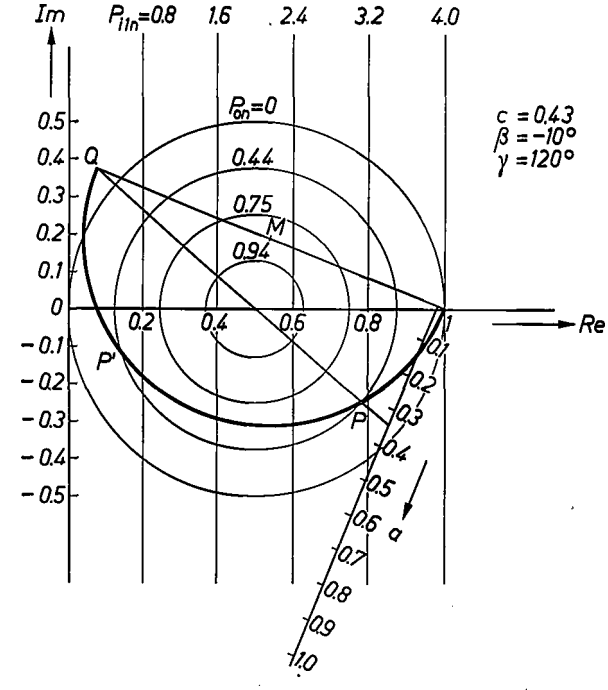


Fig. 9. Plot of $\underline{Y}$ for a motor designed for a tape recorder (see fig. 5, curve 3).

Since the relations (32) contain no machine parameters their representation by straight lines and circles is universally applicable. For a given machine it is only necessary to draw the locus diagram (29). This circle diagram can be constructed very simply as follows. The centre M is found by measuring off a distance $(\pi/8)c/\sin(\gamma/2)$ along a line starting at $(\frac{1}{2}, 0)$ and at an angle $\beta + \pi/2$ to the real axis. The radius of the circle is the distance from M to the point (1, 0), since in all cases this point represents locked-rotor operation ($a = \alpha = 0$), as can easily be seen from (29). Since on the other hand, $a = \tan\alpha$ becomes infinite for $2\alpha = \pi$, the locus diagram for $a \geqslant 0$ is a semicircle with the other end at Q (*fig. 8*).

To give a direct geometrical representation of the normalized speed a, a unit length can be measured off along the straight line starting at Q and passing through the locked-rotor point (1, 0); the point at unit distance from Q then forms the zero of a scale for a, constructed at right angles. Since the angle subtended at the centre by the chord between the locked-rotor point (1, 0) and the working point P is 2α, from (29), the angle subtended at the circumference by the same chord must be α, by elementary geometry. From this it follows that a straight line through P and Q intersects the a scale at a distance $\tan\alpha = a$ from the zero.

If Q lies inside the circle $P_{on} = 0$, i.e. on a circle $P_{on} > 0$, the machine can 'run away' (fig. 8). If Q lies outside this circle, i.e. at $P_{on} < 0$, then the intersection of the locus of $\underline{Y}$ with the circle $P_{on} = 0$ is the no-load point and the corresponding speed is the no-load speed a_0 (*fig. 9*).

In this case it is frequently sufficient to measure the unit distance not from the point Q but from a closer point on the circumference. This is done in *fig. 10*. The speeds can then be read off for all working points up to the chosen reference point A, but not for the points between A and Q.

Efficiency

The equations derived above can now be used to give a simple calculation of the efficiency of the machine at any working point of the motor region $P_{on} > 0$. From equations (32), the radius of a power circle is given by:

$$r = \tfrac{1}{2}\sqrt{1 - P_{on}}\,.$$

The normalized output power is therefore $P_{on} = 1 - 4r^2$.

[7] An essentially similar method has also been used by H. Schemmann for motors with the special characteristics $p = 1$, $m = 1$, $\gamma = \pi$. See H. Schemmann, Wirkungsgrad von kleinen zweipoligen Gleichstrommotoren mit dauermagnetischem Läufer, Elektrotechn. Z. A 88, 225-229, 1967.

For the efficiency we then have immediately:

$$\eta = \frac{P_{on}}{P_{i1n} + P_{hn}} = \frac{1 - 4r^2}{4Y_a + P_{hn}}. \qquad (33)$$

It is therefore only necessary to measure the distance r from the working point P to the point $(\frac{1}{2}, 0)$ and the perpendicular distance Y_a to the imaginary axis, and to read off in fig. 5 the losses P_{hn} due to the higher harmonics at the corresponding normalized speed a. For example, for the working point P in fig. 10, we find the values $a = 0.248$, $P_{hn} = 0.091$, $r = 0.387$, $Y_a = 0.114$. The efficiency is $\eta = 73.7\%$.

In fig. 9 the point P corresponds to a speed $a = 0.343$ and an efficiency of 13.6%. Although such a low efficiency is hardly liable to produce overheating in small electric motors with losses of a few watts, this is of course not a very good working point to choose. The same output power $P_{on} = 0.44$ can also be achieved at a higher speed $a = 1.64$ at the point P', with a much higher efficiency (72.7%). By choosing appropriate design dimensions or by introducing gearing it is always possible to save energy by operating the motor at P' instead of P.

A design criterion

From the shape of the curves P_{i1n} = constant and P_{on} = constant it can be seen that the minimum of P_{i1n} lies on the real axis between 0 and $\frac{1}{2}$ for all powers $P_{on} > 0$:

$$P_{i1n,min} = 2 - 2\sqrt{1 - P_{on}}. \qquad (34)$$

It can also be seen from fig. 6 that P_{hn} falls off exponentially with increasing a. These two facts are important in choosing a suitable working point. In the process of optimizing the design, however, they must always be considered in combination with other important conditions. These can differ considerably, depending on the application. Besides a high power-to-volume ratio it could for example be necessary to have a torque-speed characteristic with a particular slope at the working point. Such design problems can be solved with the theory presented here.

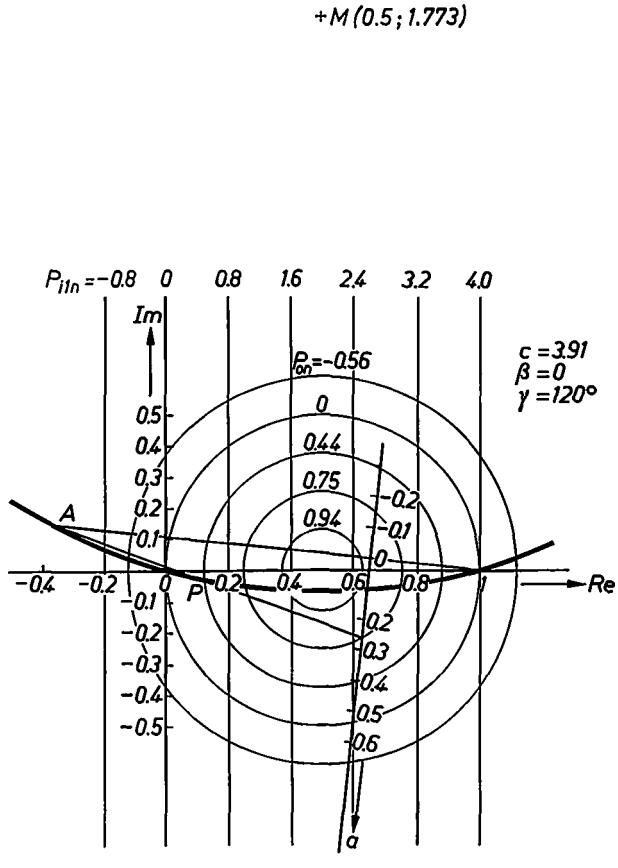


Fig. 10. Plot of $\underline{Y}$ for the ANS motor (see fig. 5, curve *1*).

Variation of the short-circuit interval

So far γ has been taken as a constant quantity. But if, as mentioned in the introduction, the short-circuit interval is to be varied to control the speed or the torque of such a motor then γ would have to be a variable. Nevertheless, the theory developed so far remains valid in this case since the treatment has been confined to steady-state performance: the dynamic transient effects can then be ignored and it can be assumed that γ does not change within a half-period. Curve (29) is of course no longer a circle but assumes a more complicated shape depending on the relation between γ and a. If this relation is introduced into (17) and (29), the torque and the locus diagram for $\underline{Y}$ can be computed point by point; it is also possible to derive the curves by interpolation between a number of curves for fixed values of γ.

It is necessary to bear in mind that β in this case is a function of γ (see 11). Since b remains constant, however, β can be eliminated from the equations of interest, (18) and (29). We then get:

$$T_n = \tfrac{1}{2}\frac{1}{1+a^2}\left[\cos b - \cos(\gamma + b) - a\left\{\frac{\pi}{2}c - \sin b + \sin(\gamma + b)\right\}\right], \qquad (35)$$

$$\underline{Y} = \tfrac{1}{2} - \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\exp\{-j(\gamma/2 + b)\} + \left[\tfrac{1}{2} + \frac{\pi}{8}\frac{c}{\sin(\gamma/2)}\exp\{-j(\gamma/2 + b)\}\right]\exp(-j2\alpha). \qquad (36)$$

In the previous treatment the powers and currents have been normalized to the quantities P_{om} and V. When γ varies, these also vary. This does not affect the determination of the efficiency, but must be taken into account when calculating the actual power or current.

The control of the switching angle γ can be effected most simply by electronic delay of the control signal for the switching intervals I_2 and I_4 (see Table I), for example by using a monostable circuit.

Summary. In electronically commutated d.c. machines the periodic reversal of the armature current can be controlled electronically, e.g. by varying the short-circuit time or by interrupting the current. In this article a theory applicable to an important class of such machines is outlined; some characteristic features of this class are that the excitation flux in the stator coils is sinusoidal and that the electronic control is limited to a variation of the short-circuit time. For this case the behaviour of the machine can be set out in expressions in closed form. The torque-speed characteristics and expressions for the power are derived; in particular, a very simple graphical procedure is described for determining the efficiency from the position of the current vector in the complex plane and a general plot of the losses due to higher harmonics.

Scaling laws for electric motors

E. M. H. Kamerbeek

Economic comparison of different types of motor

In deciding which type of motor to use in a particular application the economics may be as important as the engineering. The operating costs, which normally have to be kept as low as possible, are mainly determined by the depreciation and the cost of energy. The first requires a low price per kilowatt of rating and a long life, the second requires a high efficiency.

The relative standing of a given motor type in such a cost comparison is strongly dependent on the (mechanical) output power required. Scaling-up is an advantage for all types, but for some types more than for others; moreover, each type has its own limitations on the range of scaling. One fairly general limitation is the magnetic saturation; this has its greatest effect in reluctance and hysteresis motors [1], in which it determines the limit beyond which scaling up is no longer beneficial. In motors with windings on the stator and rotor the magnetic saturation is also of significance, but effective cooling of the windings is a major problem. In motors with permanent magnets, the demagnetization effect comes into play when the scale is increased. If the permanent magnets are replaced by superconducting magnets the range of possibilities is considerably extended, but the high costs only permit an acceptable price per kilowatt of converted power in very large systems.

This article will present a very general comparison of the various types of motor [2], based on two figures of merit. When certain simplifications are introduced, the relation between these figures of merit and the various motor parameters can be expressed in simple formulae. These formulae also indicate the dependence of the figures of merit on the dimensions of the motor, and can therefore be regarded as scaling laws for the various types of motor.

Two figures of merit

As stated earlier, the user's main costs in the operation of a motor are the depreciation and the energy costs. The annual depreciation is equal to the initial price K divided by the number of years T over which the motor is written off. The annual energy costs can be written in the form $\beta t P_d$, where β is the energy cost per kilowatt hour, t is the number of operating hours per annum and P_d is the power loss in kilowatts (this part of the electrical input power is dissipated in the motor; the rest is converted into mechanical power, and is made available at the shaft). The annual costs of using the motor are therefore:

$$G = \frac{K}{T} + \beta t P_d.$$

For simplicity we assume that the initial price of a motor is proportional to its volume V:

$$K = \alpha V.$$

The value of the constant α is not the same for all types of motor. To compare motors of different power ratings, it is necessary to determine the costs per kilowatt of output power:

$$\frac{G}{P} = \frac{\alpha}{T}\frac{V}{P} + \beta t \frac{P_d}{P},$$

where P is the mechanical output power. The factors α/T and βt are independent of the dimensions; the scale dependence is contained in the factors V/P and P_d/P. Now let us introduce two figures of merit equal to the reciprocal of these factors:

$$q_1 = \frac{P}{V} \quad \text{and} \quad q_2 = \frac{P}{P_d}.$$

The larger these figures of merit, the lower the motor costs per kilowatt of converted power. The figure of merit q_1 is equal to the power-to-volume ratio; q_2 is related to the efficiency η of the motor by the equation $q_2 = \eta/(1-\eta)$.

Instead of q_1 and q_2, we can also use the product $q = q_1 q_2$ and the ratio $\lambda = q_1/q_2$ as figures of merit. The product q is a (sometimes notional) maximum electromechanical power per unit volume, as determined by the electrical, magnetic and mechanical characteristics of the motor, since from the above

$$q = \frac{P^2}{V P_d}.$$

Let us consider a d.c. motor with permanent-magnet

Dr Ir E. M. H. Kamerbeek is with Philips Research Laboratories, Eindhoven.

[1] An introductory survey is given in E. M. H. Kamerbeek, Electric motors, Philips tech. Rev. 33, 215-234, 1973.

[2] C. Rioux, Théorie générale comparative des machines électriques établie à partir des équations du champ électromagnétique, Rev. gén. Electr. 79, 415-421, 1970.

excitation as an example. If we neglect the frictional and iron losses, the output power is given by

$$P = I_r E_r,$$

where I_r is the rotor current and E_r is the speed voltage or back e.m.f.[1]. Dissipation occurs only in the resistance R_r of the rotor circuit, so

$$P_d = I_r^2 R_r.$$

It follows that

$$q = \frac{I_r^2 E_r^2}{V I_r^2 R_r} = \frac{E_r^2}{V R_r}.$$

The figure of merit q is thus equal to the short-circuit power E_r^2/R_r per unit volume; the short-circuit power is the power dissipated in a motor when its terminals are short-circuited and it is driven externally, and is a characteristic quantity of the motor. If we write $E_r \propto \omega_r \Phi_{rs}$, then q contains the speed of the motor (which is proportional to ω_r) and the magnetic flux Φ_{rs} that is produced by the permanent magnet and penetrates the rotor. The figure of merit

$$\lambda = \frac{P_d}{V},$$

which represents the power dissipated per unit volume, includes the thermal characteristics of the motor; λ depends on the permissible temperature rise and the cooling.

For a given value of q, there is a certain value of λ corresponding to a minimum in the cost of operating the motor, since the expression for G/P given above, which can also be written in the form

$$\frac{G}{P} = q^{-\frac{1}{2}}\left(\frac{\alpha}{T}\lambda^{-\frac{1}{2}} + \beta t \lambda^{\frac{1}{2}}\right),$$

has a minimum at

$$\lambda = \frac{\alpha}{\beta t T}.$$

The annual depreciation and the energy costs are exactly equal at this minimum. The quantity q determines the magnitude of the costs per kilowatt at this minimum, given the values of the external parameters α, T, β and t.

The figures of merit q_1 and q_2 depend on a number of parameters:

— the properties of the magnetic materials and the conductors,

— the angular velocity ω_r of the motor shaft,

— the permissible increase in conductor temperature above ambient ΔT,

— the cooling coefficient c_k and

— the linear dimension L.

In this article the relations between the figures of merit and these parameters will be given in the form of proportionalities. For L, these relations can be regarded as scaling laws. A linear increase in scale, with no change in the relative proportions, is assumed here; this is not completely in accordance with reality, but the changes in the relative proportions during scaling may be regarded as a second-order effect. The areas are therefore proportional to L^2, and the volumes to L^3.

This is of significance for the cooling. The simplified thermal balance applies here:

$$P_d = c_k \Delta T . A_k,$$

where A_k is the area of the cooling surface. If the power dissipated per unit volume λ is kept constant as the scale is increased, the requirement

$$c_k \Delta T \propto L$$

follows. In practice, this requirement is at least partly met. The cooling coefficient c_k depends strongly on the cooling method used (radiation, natural or forced air cooling, water cooling). For small air-cooled motors c_k could be 20 W/m² °C, while for large turbogenerators with conductors with internal water cooling it could be 300 W/m² °C.

The two figures of merit q_1 and q_2, and their product q, do not vary in the same way with the material properties and the dimensions for all motor types. Motors are therefore divided into five classes according to construction and materials used, and the relation between the figure of merit and the parameters presented separately for each class.

For simplicity it is assumed here that the frictional and iron losses can be neglected in comparison with the losses in the conductors.

Motors with a superconducting winding (class S)

Although motors with a superconducting winding are still largely at the development stage, they should be mentioned first because their figures of merit are far superior to those of any other type of motor. The superconducting winding generates a constant magnetic field, in which a magnetic flux density of from 2 to 5 teslas can be reached; in non-superconducting machines, the effective flux density in the air gap is at most 0.8 to 1 tesla. The absence of iron in the superconducting magnet also makes the machines smaller and lighter; this is one of the reasons why they are being considered for marine applications. A type of motor that is very suitable for the application of superconductivity is the homopolar d.c. motor (see *fig. 1*). As mentioned earlier, only large units can be considered here, because of the expense of the construction and

the cooling system. Superconducting generators for a.c. current are being considered for powers exceeding 2 GVA, which could not be supplied from a conventional generator [3].

In the derivation of the relations for q_1 and q_2, the superconducting magnet can be simply characterized by the flux density B_0 at the active conductors. The energy taken by the cooling system is not taken into account. The active, uncooled conductors are characterized by the conductivity σ of the metal from which they are made. The following proportionalities are found [4]:

$$q_1 \propto \omega_r B_0 \sigma^{\frac{1}{2}} (c_k \Delta T)^{\frac{1}{2}} L^{\frac{1}{2}},$$

$$q_2 \propto \omega_r B_0 \sigma^{\frac{1}{2}} (c_k \Delta T)^{-\frac{1}{2}} L^{\frac{1}{2}},$$

$$q \propto \omega_r^2 B_0^2 \sigma L^2.$$

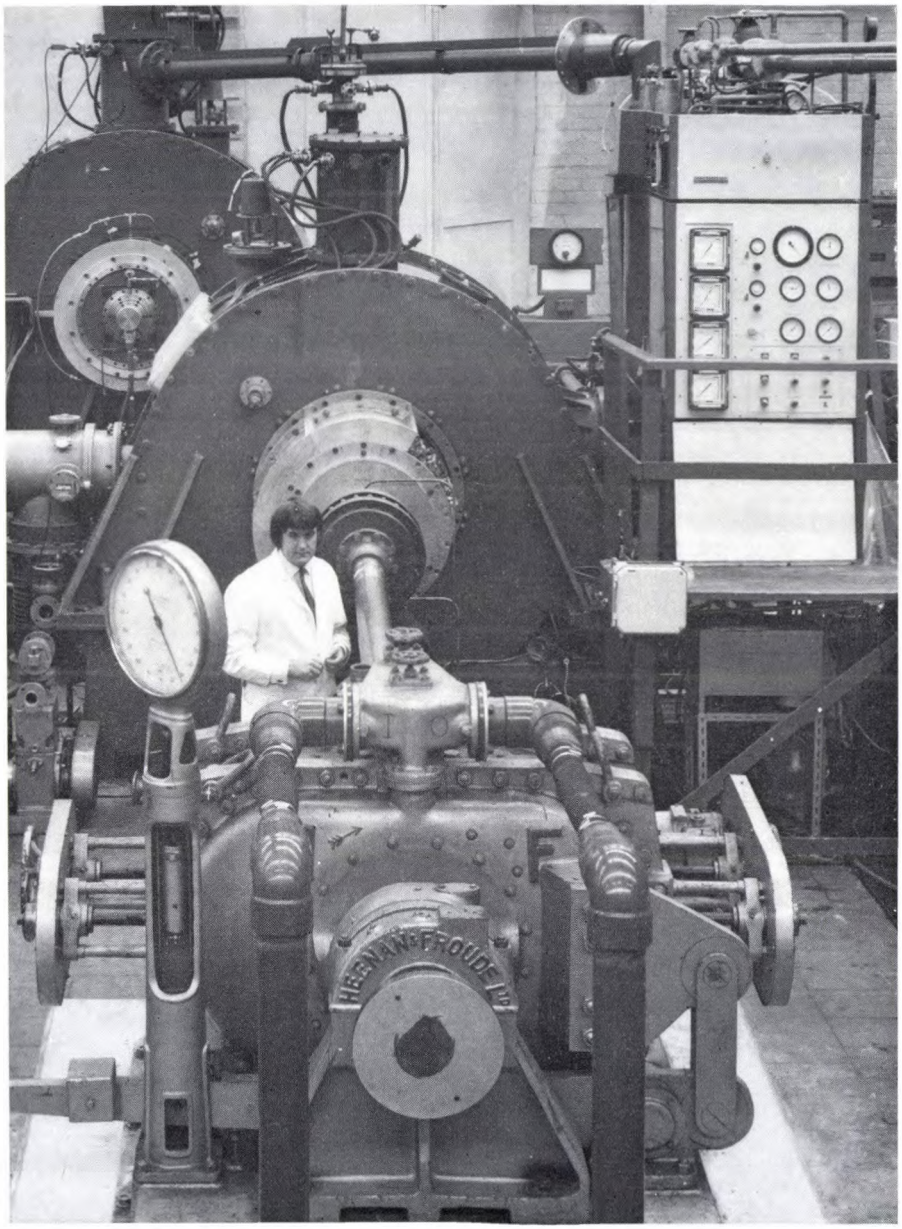


Fig. 1. Superconducting motor for marine use, supplied from a superconducting generator (in background). Both machines are of the homopolar type; the power is 1 MW. In this test arrangement the load is a water brake (in foreground). The cooling unit is in the white cabinet on the right. (Photo: International Research & Development Co. Ltd., Newcastle upon Tyne, England.)

If the dissipation per unit volume λ, which is equal to q_1/q_2, as we saw earlier, is assumed to be constant, then $c_k \Delta T \propto L$ and the following simplified relations are obtained:

$$q_1 \propto \omega_r B_0 \sigma^{\frac{1}{2}} L, \quad q_2 \propto \omega_r B_0 \sigma^{\frac{1}{2}} L, \quad q \propto \omega_r^2 B_0^2 \sigma L^2.$$

These relations show how important it is to have a high circumferential velocity ($\propto \omega_r L$). A limit is set to this velocity by the strength of the material. The high value of B_0 obtainable with a superconducting magnet is also seen to be important.

Motors with permanent magnets (class P)

Motors with permanent magnets (which include synchronous motors and d.c. commutator motors) are excited by a constant magnetic field, like motors of class S. The same expressions can therefore be used for the q's, with the difference that B_0 is now replaced by B_p, the flux density produced at the active conductors by the permanent magnet.

It is found in practice that there is a limitation with the larger permanent-magnet motors because the magnets are demagnetized when too-large reverse fields are generated by the active winding. For this reason, large d.c. machines are provided with compensating windings that oppose the rotor field. The limit to the increase in scale in the absence of a compensating winding is denoted by L_{max}.

The magnetic material is characterized by the second quadrant of the hysteresis curve, the 'demagnetization curve' (*fig. 2*). As a result of the demagnetizing effect of the ends of the magnet, the magnetic field H in the magnet in the absence of an externally applied field is negative; the magnet therefore has its operating point (H,B) in the second quadrant and it can be shown that:

$$B_p \propto (-HB)^{\frac{1}{2}}.$$

In static situations, as with loudspeaker magnets, the ideal operating point is one for which the product $-HB$ (the magnetic energy per unit volume) has a maximum value, so that the amount of magnetic material required can be minimized.

With motors, however, this is not the optimum situation, because the external fields are appreciable here. This is illustrated in *fig. 3* with reference to a schematic developed view of a d.c. motor with a permanent-magnet stator. An external field H_{ext} is superimposed on the field H in the magnet, and this reinforces the demagnetizing field at some points. If the demagnetizing field exceeds the limiting field-strength H_{rev}, the magnetic polarization J is reduced (see fig. 2, dashed line). When the external field is removed, the polarization does not return to its original value, and the original flux density is no longer obtained.

This can be prevented if the condition

$$|H_{ext}| \leqslant H - H_{rev}$$

is satisfied. Now $H_{ext} \propto J_r \tau$, where J_r is the current

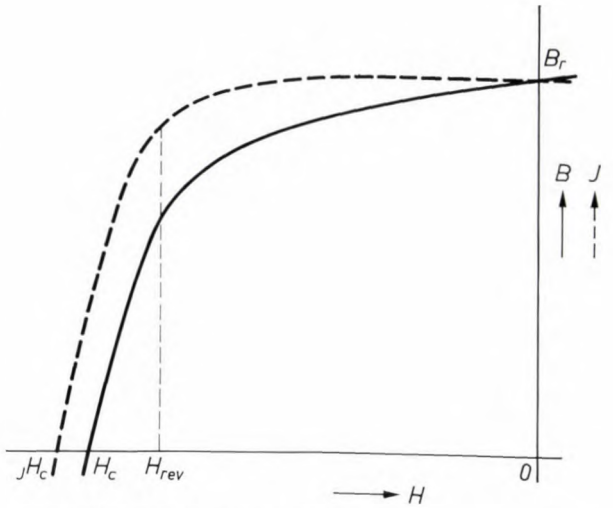


Fig. 2. Solid curve: demagnetization curve for a permanent magnet. B_r remanence. H_c coercive force. Dashed curve: corresponding variation of the magnetic polarization J. $_JH_c$ field-strength at which the polarization J becomes zero. H_{rev} limiting field-strength for reversible demagnetization.

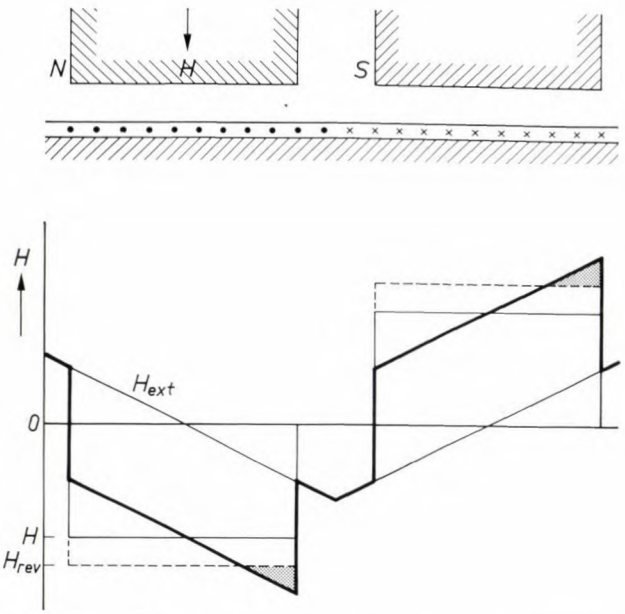


Fig. 3. Schematic developed view of a d.c. motor with permanent-magnet stator. N north pole. S south pole. The magnetic field H_{ext} excited by the rotor is added to the magnet's own demagnetization field H in the magnet. In the shaded region, the sum of these two fields exceeds the limiting field-strength H_{rev}.

[3] M. E. G. Hadlow, J. A. Baylis and B. C. Lindley, Superconductivity and its applications to power engineering, Proc. IEE **119**, 1003-1032, 1972.
J. C. Bittence, The superconductors are coming, Machine Design **45**, 26 July 1973, pp. 76-81.
H. Sequenz, Elektrische Maschinen mit supraleitenden Wicklungen, Elektrotechnik und Maschinenbau **91**, 103-109, 1974 (No. 2).

[4] These relations are derived in the appendices to: E. M. H. Kamerbeek, Comparison between permanent-magnet motors and other types of electric motor, Proc. 3rd Eur. Conf. on Hard magnetic materials, Amsterdam 1974, pp. 214-224.

density in the conductors and τ is the pole pitch. H_{ext} can therefore be limited by reducing the current density in the rotor or by choosing a larger number of poles.

This is related to the question of the optimum operating point (H,B) for motors. Various opinions have been expressed in the literature [5], but the problem has not yet been completely solved. The important point is the criterion chosen for the optimization, e.g. a maximum for q_1 or a minimum for α. If a maximum for q_1 is taken as the objective, then the basic relations are:

$$q_1 \propto J_r L B_p \propto (H - H_{rev})\,(-HB)^{\frac{1}{2}} L/\tau,$$

from which it follows that for a given L and τ, the maximum value of q_1 is found at an operating point to the right of $(-HB)_{max}$.

The maximum linear dimension for uncompensated d.c. motors then follows from the restriction on $J_r\tau \propto H - H_{rev}$:

$$L_{max} \propto p^2 H_{rev}^2/\sigma c_k \Delta T,$$

where the number of pole pairs p has been used instead of the pole pitch. With given material properties and cooling, the dimensions or torque of the machine can therefore only be increased by increasing the number of pole pairs. However, there is a limit to this increase, since on reducing the pole pitch more lines of force will pass directly from pole to pole without going through the air gap twice.

Conventional motors with wound stator and rotor (class C)

The conventional motors with current-carrying conductors on both sides of the air gap include induction motors and most large synchronous and commutator motors. They all contain soft iron. The radial flux density in the air gap is limited by the magnetic saturation there. As long as the flux density has a value lower than the saturation value B_{sat}, the following relations hold for the q's [6]:

$$q_1 \propto \omega_r \sigma (c_k \Delta T) \frac{\tau}{\delta} L,$$

$$q_2 \propto \omega_r \sigma \frac{\tau}{\delta} L^2,$$

$$q \propto \omega_r^2 \sigma^2 (c_k \Delta T) \left(\frac{\tau}{\delta}\right)^2 L^3.$$

Here the ratio of pole pitch τ to air gap δ is an important quantity.

The saturation flux density is reached when the linear dimension exceeds the critical value:

$$L_c \propto B_{sat}^2/\sigma (c_k \Delta T) \left(\frac{\tau}{\delta}\right)^2.$$

At saturation, other proportionalities hold for the q's:

$$q_1 \propto \omega_r B_{sat} \sigma^{\frac{1}{2}} (c_k \Delta T)^{\frac{1}{2}} L^{\frac{1}{2}},$$

$$q_2 \propto \omega_r B_{sat} \sigma^{\frac{1}{2}} (c_k \Delta T)^{-\frac{1}{2}} L^{\frac{3}{2}},$$

$$q \propto \omega_r^2 B_{sat}^2 \sigma L^2.$$

These expressions are surprisingly similar to those for motors of classes S and P. Even when the saturation flux density is reached it is found that the figures of merit still increase with the dimensions so that scaling-up is still worth while.

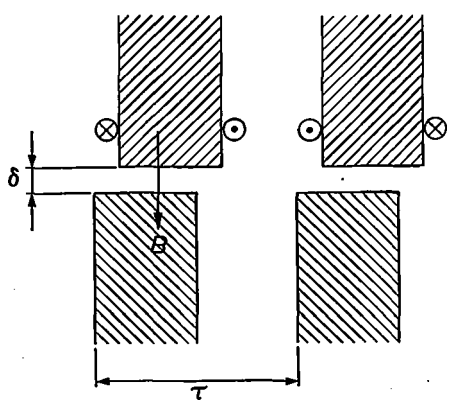


Fig. 4. Stator and rotor teeth of a reluctance motor (developed view). τ pole pitch. δ air gap.

Reluctance motors (class R)

Reluctance motors have soft magnetic material on both sides of the air gap, but a winding only on one side. The part without a winding (generally the rotor) has projecting poles on which the field in the air gap exerts forces that combine to give a torque. Reluctance motors are synchronous; they are often used as stepping motors.

Fig. 4 shows part of a developed view of a reluctance motor. It is assumed that the pole width is equal to half the pole pitch, and that all stator poles are simultaneously excited and form an alternating series of north and south poles. Reluctance motors — stepping motors in particular — generally have a large number of poles; the torque is at a maximum when $\tau/2\delta \approx 4$, and it is therefore assumed below that the ratio τ/δ is constant.

If the system is not magnetically saturated, the following proportionalities can be derived:

$$q_1 \propto \omega_r \sigma (c_k \Delta T) L,$$

$$q_2 \propto \omega_r \sigma L^2,$$

$$q \propto \omega_r^2 \sigma^2 (c_k \Delta T) L^3.$$

These are the same relations as for class C if τ/δ is assumed to be constant.

Magnetic saturation occurs above the critical dimension

$$L_c \propto B_{sat}^2/\sigma c_k \Delta T.$$

This proportionality is derived in the same way as for conventional motors, with the reservation mentioned above. At magnetic saturation:

$$q_1 \propto \omega_r B_{sat}^2,$$
$$q_2 \propto \omega_r \sigma L^2,$$
$$q \propto \omega_r^2 B_{sat}^2 \sigma L^2.$$

It can be seen that the power-to-weight ratio q_1 is independent of the size of the motor; scaling-up therefore gives no improvement in q_1. It is in fact easy to see that the torque T_e of a reluctance motor is proportional to B^2L^3 for a constant value of the geometrical parameter τ/δ, since the force acting on the rotor is proportional to the Maxwell stress [7] and hence to B^2. The force is also proportional to the surface area of the rotor and since the torque is proportional to the radius of the rotor as well, it must be proportional to L^3. If the constant value B_{sat} is now taken for B, then the power delivered $\omega_r T_e$ is proportional to the volume.

Hysteresis motors (class H)

Hysteresis motors do not usually have a winding on the rotor either. In this case, the rotor is made of material with a high magnetic hysteresis; the shape of the hysteresis loop has a great effect on the characteristics of the motor. The hysteresis motor is generally used as a synchronous motor, but it also develops a torque at speeds below the synchronous speed; it is self-starting.

The magnetic-hysteresis effects in a hysteresis motor are rather complicated: the magnetization vector not only alternates, but also rotates. Under these conditions, the hysteresis cannot be exactly calculated. In general, however, the rotor consists of a ring of hysteresis material surrounding a core. If the core is of a non-ferromagnetic material, then the lines of force in the ring are mainly parallel to the surface, so that the direction of the field is fixed and only its strength and polarity can vary; the magnetization is almost entirely alternating. If on the other hand the core is made of a material with a much higher permeability than the hysteresis material, the lines of force are largely perpendicular to the surface of the ring; here again the direction of the field is fixed, and the magnetization alternates. Both constructions are found in practice; their common characteristic is that the magnetization produced is mainly alternating, and we shall assume in the following that this is the case.

The losses in the rotor will then be proportional to the area enclosed by the hysteresis loop. If the material is saturated this area is A_{hmax}. If not, the hysteresis describes a smaller loop; if the magnetic field-strength has a peak value $\hat{H}$, the area of the loop will be given approximately by $A_h \approx A_{hmax} \hat{H}^2/H_{max}^2$. The area A_{hmax}, which is characteristic of the material, may be assumed to be proportional to the product $B_r H_c$, and H_{max} proportional to H_c (see *fig. 5*).

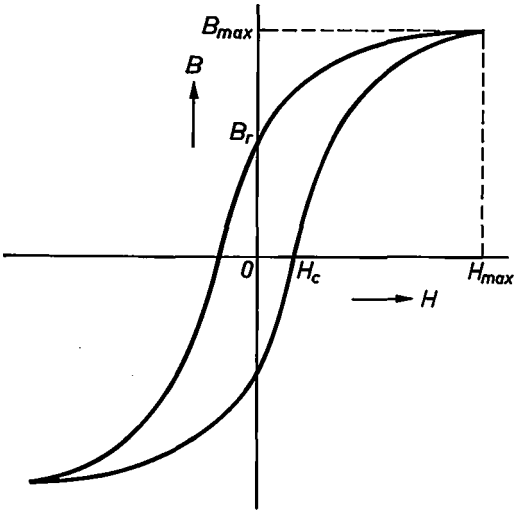


Fig. 5. Hysteresis loop, with the quantities B_{max}, B_r, H_{max} and H_c.

If the stator field is sinusoidal, as we assume here, then the torque of the hysteresis motor is proportional to the hysteresis losses in the rotor [1]. It follows that if magnetic saturation does not occur, the following relations apply for the figures of merit of a two-pole motor:

$$q_1 \propto \omega_r \sigma (c_k \Delta T) \frac{B_r}{H_c} L,$$
$$q_2 \propto \omega_r \sigma \frac{B_r}{H_c} L^2,$$
$$q \propto \omega_r^2 \sigma^2 (c_k \Delta T) \left(\frac{B_r}{H_c}\right)^2 L^3.$$

[5] B. W. Wentworth and E. L. Ellis, Stabilization prediction for permanent-magnet field motors and generators, J. appl. Phys. 37, 1143-1146, 1966.
J. R. Ireland, New figure of merit for ceramic permanent magnet material intended for dc motor applications, J. appl. Phys. 38, 1011-1012, 1967.
R. Fischer, Probleme der inneren und äußeren Entmagnetisierung im Felde permanenterregter Gleichstrommotoren, Elektrotechn. Z. A 89, 266-269, 1968.

[6] The figure of merit q_2 is identical with the 'goodness factor' introduced by E. R. Laithwaite: The goodness of a machine, Proc. IEE 112, 538-541, 1965.

[7] E. M. H. Kamerbeek, Torque measurements on induction motors using Hall generators or measuring windings, Philips tech. Rev. 34, 153-162, 1974 (No. 7).

If the material is saturated, these relations become:

$$q_1 \propto \omega_r B_r H_c,$$

$$q_2 \propto \omega_r \sigma (B_r/H_c) L^2,$$

$$q \propto \omega_r^2 \sigma B_r^2 L^2.$$

This happens when a critical dimension

$$L_c \propto H_c^2/\sigma c_k \Delta T$$

is exceeded. As in the case of the reluctance motor, the power-to-weight ratio q_1 of the hysteresis motor does not increase with the volume. This is easy to understand, since the hysteresis losses, and hence the torque, are proportional to the volume of the rotor. However, the power-to-weight ratio does increase with the area of the hysteresis loop. To obtain a large value of q_2, it is found to be advantageous to use a material with a large value of B_r/H_c, i.e. with a tall and thin hysteresis loop.

Conclusions

To provide a good overall picture of the situation, the scaling laws for the three q's are plotted in *fig. 6a*, *b* and *c*, on the assumption that $c_k \Delta T \propto L$ (λ constant); the scales are logarithmic. This figure is idealized, and is only intended to give a qualitative picture. The discontinuities in some of the curves occur at the critical dimension L_c above which magnetic saturation occurs; in practice, this will happen at an external motor diameter of between 5 and 15 cm. The discontinuity in the curves for class C is situated more to the right than for class R, because the critical dimension is proportional to $(\delta/\tau)^2$, and conventional motors often have a larger air gap then reluctance motors. The discontinuity in the curves for hysteresis motors is even further to the right, because the hysteresis material has a much lower permeability than soft iron, so that the flux density does not reach saturation so quickly. However, it should be pointed out that these are general considerations; there are exceptions. The initial and final points of the curves give an approximate idea of the power ranges for which the different types of motor are best suited.

A number of conclusions can be drawn:

— For all classes of motor, $q_1 \propto \omega_r$, $q_2 \propto \omega_r$ and $q \propto \omega_r^2$. Increasing the speed is thus an important way of reducing the costs of the energy conversion. The circumferential speed is limited in practice to about 200 m/s; this limit is only approached in turbogenerators, but then the rotor has to be forged in a single piece to make it strong enough. As might be expected, bearing problems arise at these high speeds.

— Classes S and C are superior for very large motors. The q-values for class S are higher than for class C, because of the high flux density B_0 (2 to 5 T).

— All classes except S are of interest for small motors. For small and very small motors, the power-to-weight ratio and efficiency are better in class P than in classes C, R and H.

The scaling laws for class P resemble those for class S. This is not really surprising, since a permanent magnet can be regarded as a superconducting magnet

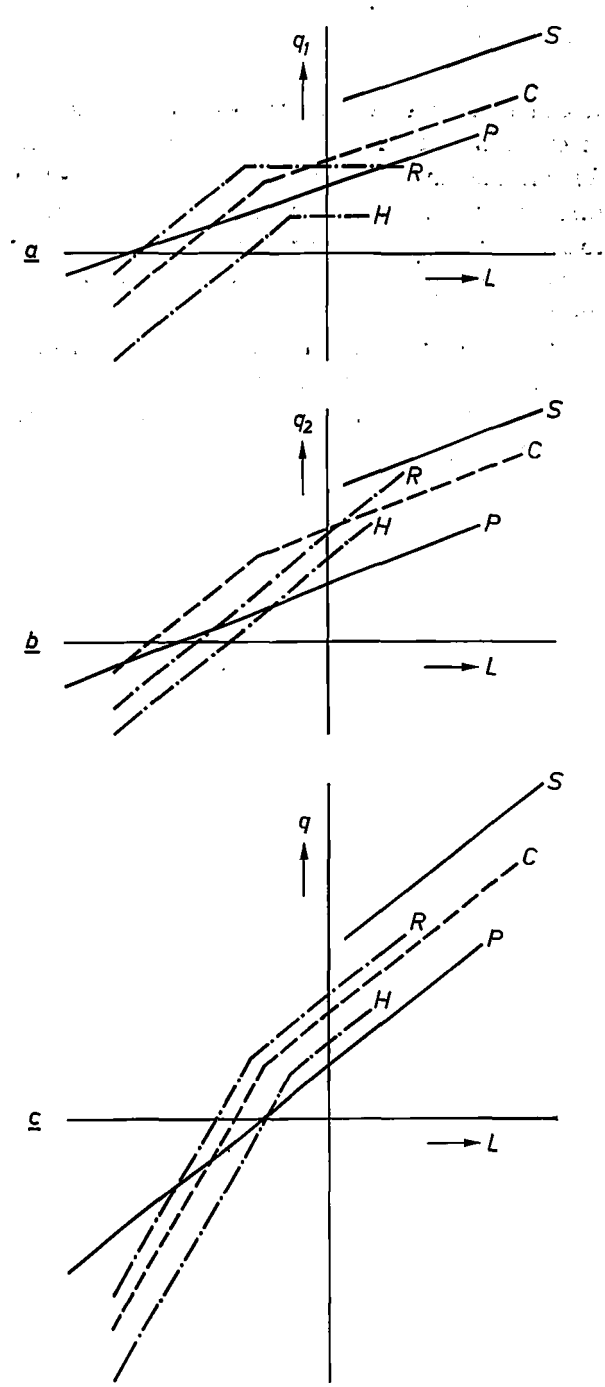


Fig. 6. Variation of the figures of merit with the linear dimension L. The discontinuity in some of the curves occurs at the critical dimension L_c at which magnetic saturation starts. *a)* The power-to-weight ratio q_1, *b)* the figure of merit q_2, which is related to the efficiency, *c)* q, the product of q_1 and q_2.

at room temperature. However, class P motors are subject to magnetic limitations: it would be desirable to have higher values of B_r, H_c and H_{max} without too great an increase in the cost of manufacture. If this were possible, uncompensated class-P motors could be made even larger. It is true that some of the compounds of cobalt with the rare-earth metals do have excellent magnetic properties, but these compounds are still so expensive that they can only be considered for miniature motors.

— Class R motors have a high power-to-weight ratio q_1 at only moderately large dimensions. If they are given a large number of pole pairs for optimum design, they require high-frequency a.c. current for normal speeds. The savings arising from the high power-to-weight ratio are completely or partly absorbed by this requirement. These motors are widely used as stepping motors, where a more complicated power supply is necessary anyway.

— The hysteresis motor has a reasonably high efficiency, but its power-to-weight ratio is always lower than that of a permanent-magnet motor. While a hysteresis rotor does behave like a permanent magnet at the synchronous speed, it is a permanent magnet with a lower H_c than that of hard magnetic material. This means that more magnetic material is necessary for a given output power.

— In motors of classes R and H, but not in the other types, the power-to-weight ratio q_1 does not increase above the critical dimension L_c. This is because these motors only have a winding on one side of the air gap. The forces on the part without a winding depend only on the flux density in the rotor material, which reaches a limit at the saturation flux density. On scaling-up beyond L_c, the ratio q_1/q_2 becomes steadily worse; the motors become very heavy compared with those of other types. This makes the purchase price high, so that these motors are uneconomical, despite their high efficiency.

Summary. Besides the technical requirements, costs are also important in determining the best type of motor for a given application. The costs can be divided into two groups, one relating to purchase, the other to consumption of energy. Two figures of merit corresponding to the two groups are defined. The motors are divided into five classes, and the variation of these two figures of merit with a number of parameters, in particular the dimensions of the motor, is shown for each class. The scaling laws derived in this way are different for the different classes of motor: some motors give optimum performance at large dimensions, others at small ones. In some classes, magnetic saturation of the iron or demagnetization of permanent magnets sets a limit to the scaling range.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or cooperate 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, 3 avenue Descartes, 94450 Limeil-Brévannes, 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*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.)	*N*

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. Bäuerle, Y. Yacoby (Hebrew University, Jerusalem) **& W. Richter** (Max-Planck-Institut für Festkörperforschung, Stuttgart): Soft mode behavior in $PbTi_{1-x}Zr_xO_3$.
Solid State Comm. **14**, 1137-1139, 1974 (No. 11). *A*

R. Bernard, M. Boothroyd, H. Vainsel, M. Demeester (all with Hôpital St. Pierre, Bruxelles) **& W. J. Rey:** Le monitoring des arythmies en temps réel.
J. Inform. méd. IRIA, Toulouse 1974, pp. 143-150. *B*

D. Bois: Etude expérimentale de la relaxation de réseau des centres profonds dans GaAs.
J. Physique **35**, C3/241-246, 1974 (Colloque C3). *L*

K. H. J. Buschow: On the eutectoid decomposition of $CaCu_5$-type rare earth - cobalt phases.
J. less-common Met. **37**, 91-101, 1974 (No. 1). *E*

F. L. Engel (Institute for Perception Research, Eindhoven): Visual conspicuity and selective background interference in eccentric vision.
Vision Res. **14**, 459-471, 1974 (No. 7).

C. T. Foxon, M. R. Boudry & B. A. Joyce: Evaluation of surface kinetic data by the transform analysis of modulated molecular beam measurements.
Surface Sci. **44**, 69-92, 1974 (No. 1). *M*

G. C. Groenendaal, L. O. G. Kristiansson (Royal Institute of Technology, Stockholm) **& J. B. H. Peek:** The behavior of a hold-and-forward concentrator for a type of dependent interarrival statistics.
IEEE Trans. **COM-22**, 273-278, 1974 (No. 3). *E*

E. E. Havinga & M. H. van Maaren: Brillouin zone effects on the critical temperature of superconductors and some related normal metal properties.
Physics Repts. **10C**, 107-150, 1974 (No. 3). *E*

B. Hoekstra: Magnetostriction and magnetic anisotropy of $CdCr_2S_4$.
Phys. Stat. sol. (b) **63**, K 7-11, 1974 (No. 1). *E*

J.-P. Krumme: The structure of the compensation wall in ferrimagnets.
Phys. Stat. sol. (a) **23**, 33-41, 1974 (No. 1). *H*

G. Landvogt & G. Schröder: Frequenzanaloge Signale für die Prozeßinstrumentierung.
Regelungstechn. Praxis & Prozeß-R. **16**, 88-92, 1974 (No. 4). *H*

F. Meyer: Surface-state densities on clean semiconductor surfaces measured by ellipsometry.
Phys. Rev. B **9**, 3622-3626, 1974 (No. 8). *E*

R. I. Pedroso: Thermal analysis of laminar fluid film under side cyclic motion.
Trans. ASME C (J. Heat Transfer) **96**, 100-106, 1974 (No. 1). *N*

F. Rondelez: Measurements of diffusion anisotropies in liquid crystals by use of dyes.
Solid State Comm. **14**, 815-819, 1974 (No. 9). *L*

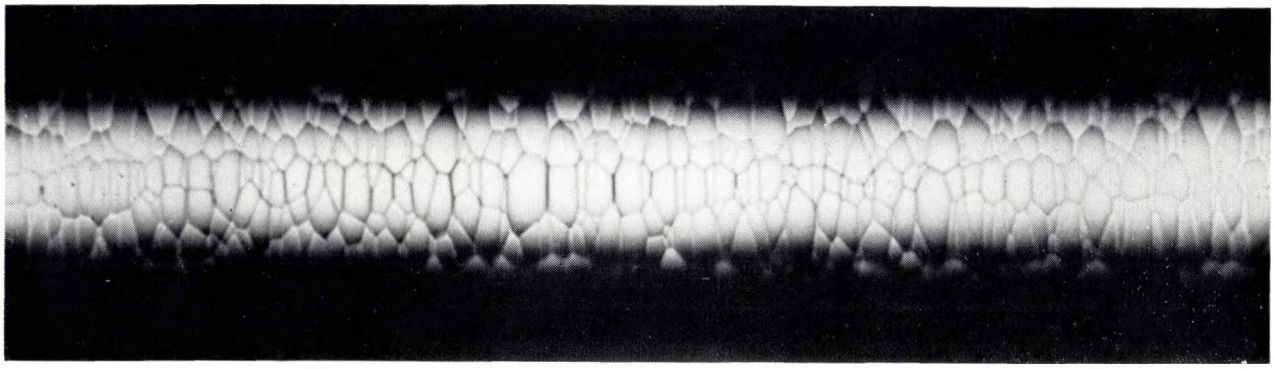
O. Schob & W. Vrieze (Philips Lighting Division, Eindhoven): Der Verschleiß von Diamantziehsteinen in Abhängigkeit von der kristallographischen Orientierung.
Draht **25**, 288-297, 1974 (No. 6).

J. M. Shannon: Increasing the effective height of a Schottky barrier using low-energy ion implantation.
Appl. Phys. Letters **25**, 75-77, 1974 (No. 1). *M*

H. W. Werner: A study on mass discrimination in a magnetic sector mass spectrometer.
Int. J. Mass Spectrom. Ion Phys. **14**, 189-203, 1974 (No. 2). *E*

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 5



Boron filament: a light, stiff and strong material

A. C. van Maaren, O. Schob and W. Westerveld

Philips produce tungsten wire in large quantities for their incandescent lamps. Tungsten wire is also ideally suited as a core for boron filament, which has attracted increasing attention in recent years as a basis for strong, stiff and light structural materials. It is therefore not surprising that Philips have also given serious thought to the production of boron filament. Research carried out at the Central Laboratories of the Lighting Division has created knowledge that makes it possible to produce boron filament of very high quality. Boron filament is approximately as strong as the strongest steel wire, twice as stiff and three times as light. By embedding it with the appropriate orientation in an epoxy resin, a structural material is obtained which combines high strength and stiffness with low weight. The boron wire made by Philips has been used in an experimental section for the cargo-hold door of a future space shuttle.

Introduction

Houses, cranes, gears and innumerable other large and small structures have to be not just strong but also stiff. The suspension bridge built over the Menai Strait in Wales between 1819 and 1826 is strong but was originally not stiff enough: it swayed so much in gales that the horses pulling the mail-coach could not keep their feet. It was later stiffened and is still in use, carrying modern traffic [1].

Strong and stiff structures require strong and stiff materials such as steel. In the last twenty or thirty years engineers have been concerned with discovering or devising materials that are also light. The importance of this last property becomes obvious when we consider that in many bridges, cranes and vehicles the main load carried is the actual weight of the structure itself; they are in fact much heavier than their useful load. The need for light and strong materials is even more vital in aircraft and spacecraft.

It is now possible to make materials that are just as strong and stiff as steel but considerably lighter. These are the modern reinforced materials sometimes referred to as fibre or filament composites. They consist of oriented strong and stiff filaments of one material in a matrix of another tougher and more flexible material. The situation is thus the reverse of that in traditional reinforced materials such as reinforced concrete and

A. C. van Maaren and Dr O. Schob are with the Philips Lighting Division; W. Westerveld was also with the Lighting Division until his retirement in 1974.

[1] J. E. Gordon, The new science of strong materials, Penguin Books, 1968. This paperback gives a very clear general account of materials science.

glass-fibre-reinforced polyester, where the bars and fibres are in fact tougher than the matrix.

Only a few types of fibre or filament are suitable for modern composites, and boron filament is one of these. In its single-crystal form boron is, after carbon (diamond), the element with the highest strength and stiffness. Like diamond, however, it is also brittle and cannot be drawn into filaments. Boron filament is therefore a composite (a 'clad wire'), formed by depositing boron on a wire core. Because of the high temperature necessary for this process, the choice of core material is very limited; as a rule, tungsten is used.

Philips are interested in boron filament, partly because the company itself produces large quantities of tungsten wire and partly because of the importance that Philips also attach to materials that are strong, stiff and light. Some of the components in production machinery, for example, are subjected to large forces and accelerations. To acquire experience in this field we undertook an investigation into the making of boron filament with a tungsten core. The knowledge gained has enabled us to produce boron filament of previously unequalled quality quickly and reproducibly on a laboratory scale.

This investigation forms the chief topic of the present article. It is preceded by an introduction dealing with the concepts of strength and stiffness, the effect of faults in the material on strength, and strong filaments and filament composites. The article ends with a description of how boron filament is made into a structural material and some examples of its application.

Strength and stiffness

Strength and stiffness are easily defined in the case of a test bar in *tension*. If the bar is subjected to a steadily increasing tensile load, strain first occurs, followed eventually by fracture (*fig. 1*). To begin with, the deformation is elastic: the bar returns to its original state when the load is removed. The strain ε, i.e. the relative extension $\Delta l/l$ of the rod, and the tensile stress σ are proportional to each other (Hooke's law):

$$\sigma = E\varepsilon. \qquad (1)$$

The proportionality constant E is the 'stiffness' or 'Young's modulus' of the material. Beyond a certain extension the deformation becomes plastic: the bar no longer returns to its original state when the load is removed. The tensile strength of the material (σ_t) is the tensile force at which the bar finally breaks, divided by the original cross-sectional area.

A material can also be subjected to a shear stress, and a stiffness constant and a strength can likewise be defined for that situation.

The important point about the new materials, just as for steel, is that they can be loaded heavily in *tension*. Materials that can accept *compressive* stress — such as stone or concrete, which are strong in compression but weak in tension — have been freely available since ancient times. We shall therefore confine the rest of our discussion to tensile-load situations.

Strength of materials in theory and practice

The strength found in a perfect single-crystal bar subjected to a homogeneously distributed tensile stress ought to depend entirely on the strength of the chemical bond between the atoms. A measure of this bond strength is the *surface tension* γ (surface energy per unit area). It can be shown that the 'theoretical tensile strength' σ_{th} is given approximately by:

$$\sigma_{th} = \sqrt{E\gamma/a}, \qquad (2)$$

where a is the distance between two adjacent crystal planes, at right angles to the direction of tension, which will be separated when the bar breaks.

The surface tension — a familiar concept and an easily measured quantity for liquids — is a quantity that can also be estimated reasonably well for solid substances [2].

In the simple theory from which (2) was derived γ is proportional to E, so that σ_{th} is proportional to $E/\sqrt{a}$. Materials that are both stiff and light (E large, a small) should therefore also be strong — an interesting conclusion in view of the need for strong and stiff mat-

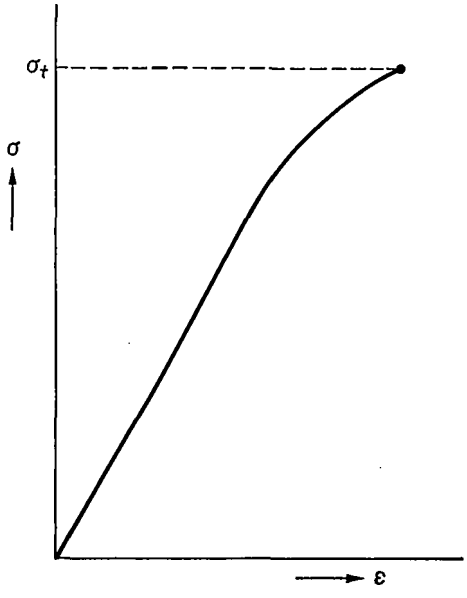


Fig. 1. The tensile stress σ as a function of the strain ε (the 'stress-strain curve') of a test bar with a tensile load (schematic). The ratio σ/ε in the linear initial part of the curve is the stiffness or Young's modulus E of the material. The tensile force at which the bar breaks, divided by the original cross-sectional area, is known as the tensile strength (σ_t).

erials that are also light. There are a number of elements in the second and third periods of the periodic table, or compounds of these elements, e.g. B, C, B_4C, SiC, Si_3N_4, BN, AlN and Al_2O_3, which are light and very stiff. The strength of the chemical bond in these substances is also apparent from their high melting points.

Equation (2) can be derived from some very simple arguments [3] based on principles put forward by A. A. Griffith in about 1920. If a bar is stretched, we add to it a strain energy of $\int\sigma d\varepsilon$ per unit volume. Since we are only concerned with an order-of-magnitude calculation here, we shall assume for simplicity that Hooke's law continues to apply until the bar breaks. We then have: $\int\sigma d\varepsilon = \frac{1}{2}\varepsilon\sigma = \frac{1}{2}\sigma^2/E$. The strain energy between two crystal planes is thus $\frac{1}{2}\sigma^2 a/E$ per unit area. As soon as this energy starts to exceed the surface energy 2γ of two fracture planes the bar will break. From this we obtain equation (2) but with a factor of 2 before the root sign. Closer examination shows that the correct factor in front of the root sign also depends on the type of bond in the material [4].

In spite of the theoretical predictions, the light and stiff materials B, C, B_4C, etc., in their normal forms are practically incapable of withstanding tensile loads. The strengths of other materials measured in practice also remain far below the theoretical values given by equation (2). The theoretical strength σ_{th} for steel is about 30 000 N/mm^2. The strength of ordinary commercial steels is only 500 N/mm^2, and that of very strong steels (piano wires) 3000 N/mm^2.

In about 1920 A. A. Griffith [5] showed from an experiment on thin glass fibres that the theoretical strength is not simply a fictitious quantity. For the theoretical strength of glass he obtained a value of about 15 000 N/mm^2. That of ordinary glass is about 200 N/mm^2. Thin, newly drawn glass fibres, however, are far stronger, and Griffith found that as his fibres got thinner they also got stronger. Extrapolation to zero thickness gave a value of 12 000 N/mm^2, which is not so very far below the theoretical value.

The weakness of *brittle* materials such as glass, boron and sapphire in their usual form must be ascribed to the *notch effect* from scratches on the surface, cracks on the inside or, more generally, any irregularity in the material. If a bar in tension contains a notch (*fig. 2*), the stress is not homogeneously distributed but is much greater at the edge of the notch than the average stress (stress concentration). Once the theoretical strength is reached at this edge, the material breaks there, the stress concentrates at the new edge, the material breaks there again, and so the process continues. Griffith calculated that a brittle bar with internal disc-shaped cracks at right angles to the direction of tension has a strength σ_c given by:

$$\sigma_c = \sqrt{E\gamma/c}, \qquad (3)$$

where c is the radius of the largest disc [6]. If c is one atomic spacing, σ_c is equal to the theoretical strength σ_{th}. Cracks of only 100 atomic spacings in Griffith's model will decrease the strength by a factor of 10.

In glass the sensitive place is its surface: if the surface is carefully protected from damage, glass can be

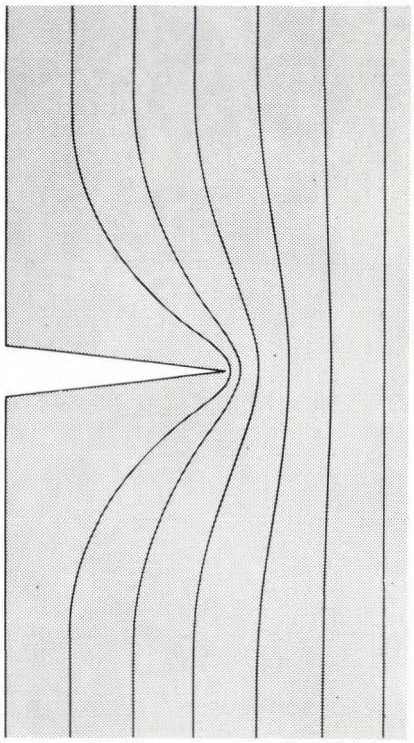


Fig. 2. 'Lines of force' in a bar with a notch (schematic). The direction and density of these lines indicate the direction and size of the main stress. The bar has a tensile load in the longitudinal (vertical) direction. There is a concentration of the lines of force at the edge of the notch.

stronger than the strongest steel. Extreme strengths are found in *whiskers*. These are thin, needle-like single crystals (e.g. of Al_2O_3, SiC or graphite) in which, because of the reduced volume, the probability of irregularities is extremely small.

In metals an entirely different mechanism is responsible for failure to attain the theoretical strengths. Since the metal bond is relatively omnidirectional in nature, atomic surfaces can slide relatively easily in relation to each other through the movement of dislocations. Under load, metals therefore often exhibit marked plastic deformation (flow) before they break: they are to a greater or lesser degree ductile. Metal whiskers are also relatively strong owing to the absence of dislocations. In boron, graphite and the other light, stiff, non-metallic materials mentioned, dislocations cannot move

[2] See for example B. F. Ormont, Dokl. Akad. Nauk SSSR **106**, 687, 1956, and Zh. n'eorg. Khimii **3**, 1281, 1958.
[3] See the book by J. E. Gordon [1], pp. 68 ff.
[4] A. Kelly, Strong solids, Clarendon Press, Oxford 1966, p. 11.
[5] A. A. Griffith, Phil. Trans. Roy. Soc. A **221**, 163, 1920.
[6] See A. Kelly [4], p. 45.

because of the covalent and hence markedly directional nature of the atom bonds.

However, while ductility is the weak feature of a metal, it is also its great advantage, since the cracking process is halted as soon as it occurs because plastic deformation evens out the stress concentrations at the edge of the crack before local fracture occurs.

Plastic deformation (flow) and brittle fracture (propagation of cracks) are theoretically possible in any material. Which of the two processes occurs in a particular case will largely depend on the temperature and the nature of the deformation.

of the types now under investigation. To illustrate the properties of boron filament we have compared its strength σ_t and stiffness E with those of some other strong filaments and fibres in *Table I*. The *specific* strengths σ_t/ϱ and *specific* Young's moduli E/ϱ are also shown (ϱ is the density). These are appropriate figures of merit for assessing the quality of strong, light and stiff materials.

Steel is a good, if not the best, representative of metal filaments. It is strong and stiff but heavy. Tungsten and molybdenum are just as strong, considerably stiffer but also much heavier. Of the metal filaments, beryllium [7]

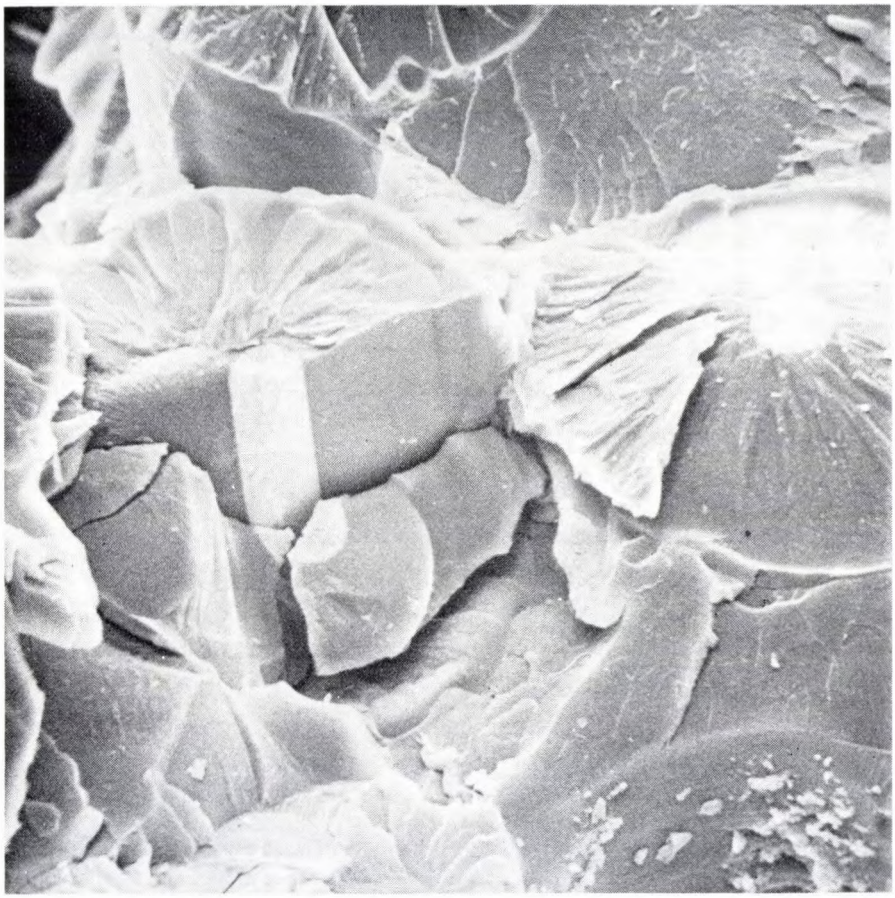


Fig. 3. Fracture surface of a boron-epoxy composite. The fracture clearly started at the filaments. It can also be seen that the adhesion between the filaments and the matrix is very good.

Strong filaments and filament composites

The glass fibres and whiskers referred to above furnish striking examples of the fact that materials attain their greatest strength in the form of thin filaments. In a nutshell, this is because it is only in this form that the perfect state can be very closely approached. This more or less perfect state may be monocrystalline (whiskers), polycrystalline (graphite filament) or amorphous (glass, boron filament).

The production of very strong materials must therefore be based on filaments and boron filament is one

Table I. Tensile strength σ_t, Young's modulus E and density ϱ of some filaments and wires, together with their specific strength σ_t/ϱ and specific stiffness E/ϱ.

	σ_t 10^3 N/mm^2	E 10^5 N/mm^2	ϱ 10^3 kg/m^3	σ_t/ϱ 10^5 Nm/kg	E/ϱ 10^7 Nm/kg
Steel	4.2	2.1	7.8	5.4	2.7
Beryllium	1.3	2.4	1.8	7.0	13
Quartz glass	6.6	0.73	2.2	30	3.3
Graphite	2.5	2.1	2.0	13	11
Boron	3.0	3.8	2.6	12	15
SiC whiskers	21	4.9	3.2	66	15

also merits some attention because it is so light; unfortunately it is toxic and very expensive. Glass fibres are strong and light but not stiff. The quartz-glass fibre included in the Table is the strongest of the glass fibres.

Of the filaments that are strong, stiff and light, graphite filament and boron filament are the only ones now being tested on a wide scale for various aeronautical and aerospatial purposes. Like boron, graphite cannot be drawn into filament. Graphite filaments are obtained by heating filaments of organic material in an inert atmosphere under a tensile load and thus carbonizing them [8].

Whiskers are by far the strongest form of any material. However, unlike steel or boron filament, they cannot be made in lengths exceeding several centimetres; this means that special processing is required, which will not be discussed here.

Construction with individual filaments is difficult. Structural elements — strips, sheets, struts, tubes — can be produced by embedding the filaments in another material, the carrier or matrix. In the resultant 'composite' the matrix protects the filaments or fibres from damage and also ensures that forces are distributed evenly over the filaments. The maximum advantage can be derived from the filaments if they are kept parallel and as close as possible to each other over the entire length of the structural element by a strongly adhesive matrix. The resultant elements are strong and stiff in one direction and can be combined to form units that are strong and stiff in a number of directions. Epoxy resin is a suitable matrix for boron filament, largely because the adhesion is so good (*fig. 3*).

The similarity between modern composites and the older ones such as reinforced concrete or glass-fibre-reinforced polyester is only superficial. The now generally accepted designation of 'reinforced materials' for the modern composites is not in fact a very appropriate one, since these are not really materials that are reinforced by filaments — like reinforced concrete — but very strong filaments embedded in a matrix. Moreover, the wires or bars in older materials often act as 'crack-stoppers': cracks in the matrix stop short at the filaments. The original rather brittle materials thus acquire a certain toughness. In modern composites, on the other hand, the matrix is usually tougher than the filaments which are not intended to stop cracks. This difference can clearly be seen if we attempt to break glass-fibre-reinforced polyester and a boron-filament/epoxy composite. Since glass has a relatively large strain and boron filament relatively little strain, the fracturing process starts from the matrix in the first material and from the filaments in the second (see fig. 3).

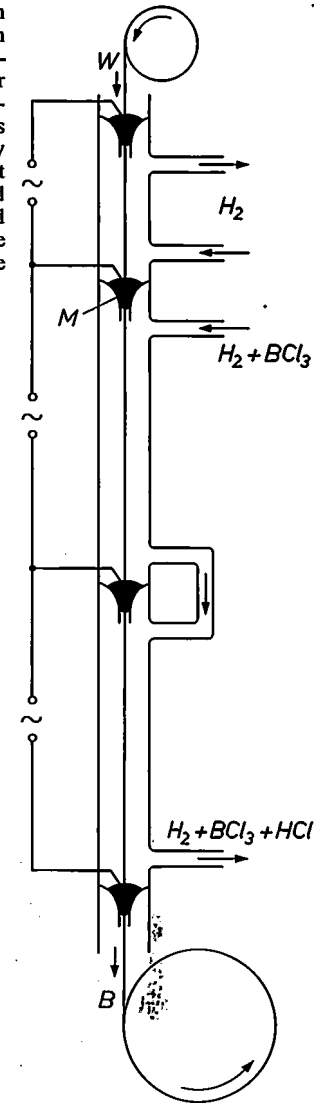


Fig. 4. Schematic representation of the production of boron filament. Tungsten wire (*W*) passes through a reduction chamber and then through two boron-deposition chambers. The wire is electrically heated; the mercury traps (*M*) also provide contact with the wire. The wire is cleaned in the reduction chamber and boron is deposited on it in the subsequent two chambers. *B* the finished boron filament.

The manufacture of boron filament

Boron filament is made by heating a tungsten wire electrically in a gas mixture containing a boron compound. The gas reacts at the wire and deposits boron on it. In our investigations we used wire temperatures ranging from 950 °C to 1320 °C and a gas mixture of H_2 and BCl_3. The reaction at the wire produces both solid boron and gaseous HCl. Only about 10% of the expensive BCl_3 is converted to boron in this reaction. Efficient recovery of the unused BCl_3 lowers the cost of boron filament considerably.

Our best results were obtained with a continuous production process in which the tungsten wire first passes through a reduction chamber and then through a number of boron-deposition chambers (*figs. 4* and *5*). The reduction chamber is included for cleaning the

[7] N. P. Pinto and J. P. Denny, Metal Progress **91**, No. 6, 107, June 1967.

[8] M. M. Tang and R. Bacon, Carbon **2**, 211 and **221**, 1964. W. Johnson and W. Watt, Nature **215**, 384, 1967, and New Scientist **41**, 398, 1969.

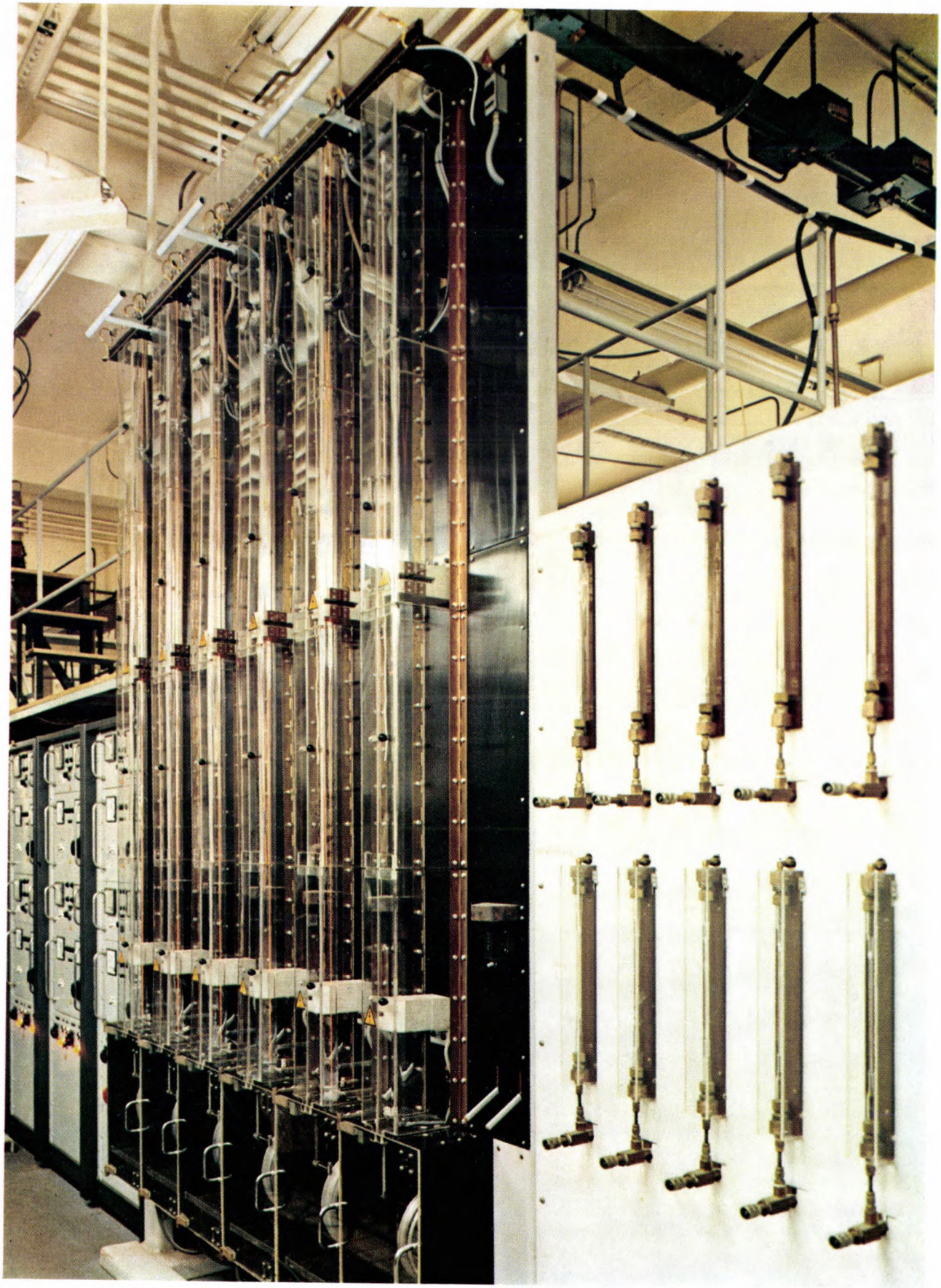


Fig. 5. A battery of boron-deposition reactors in which boron filament was produced continuously for several months.

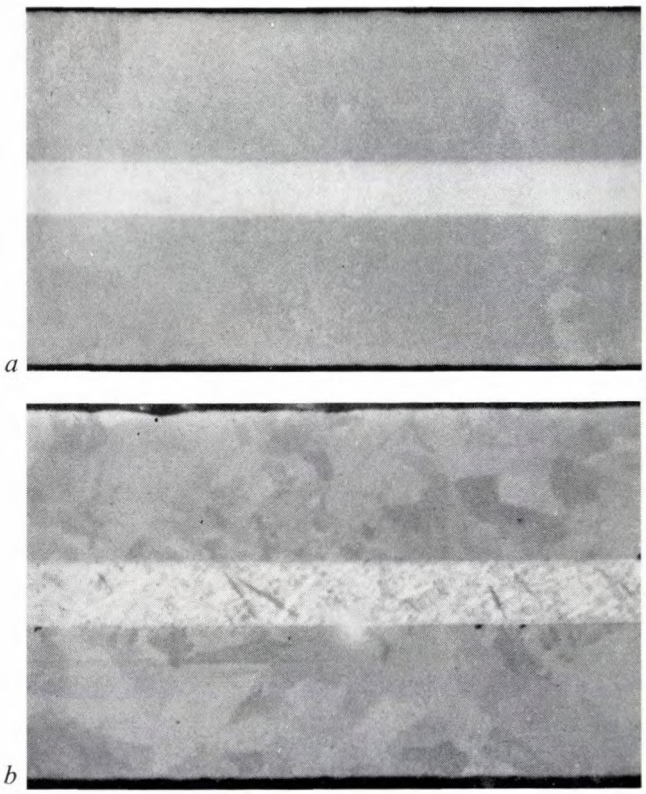


Fig. 6. *a*) Longitudinal section through a boron filament in which the boron is amorphous. *b*) As (*a*), but with crystalline boron. Strong filaments are invariably amorphous. Filaments of crystalline boron are weak. The sections were photographed with a polarizing microscope, which makes the structure in (*b*) visible.

surface of the tungsten wire. The chambers are sealed to the gas mixtures by mercury traps through which the tungsten wire passes. These traps also act as the electrical connections for the heating current. The current, and hence the temperature of the wire, can thus be adjusted separately for each chamber. The diameter of the incoming tungsten wire is usually 12.5 μm and that of the outgoing boron filament 100 μm.

There are three vital factors involved in this process: the variety of the boron modifications, the manner in which the boron coating begins to form and grows, and the formation of tungsten borides in the core. We shall now refer to these factors in explaining how to obtain the best boron filament.

Boron modifications

The surface of a strong boron filament has a 'corn-on-the-cob' structure (see the photograph on the title page). Examination of the boron wire by X-ray diffraction shows that the boron is practically amorphous, with no diffraction lines. This modification is metastable, tending to change to a crystalline modification with poor mechanical properties (*fig. 6*). Below 300 °C the rate of change is completely negligible, so that there is no danger of recrystallization if the finished

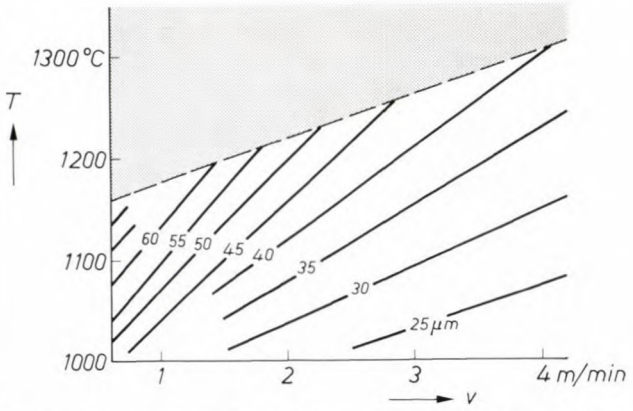


Fig. 7. *T-v* diagram (*T* temperature, *v* wire speed) with lines showing where filaments with a diameter of 25, 30, ... 60 μm occur in an experimental boron-deposition chamber of length 25 cm. The lines were obtained by interpolation between diameters found at four wire-feed rates and six temperatures. Filaments formed at temperatures in the grey region above the dashed line exhibit a surface containing crystalline regions (fig. 8) and are always relatively very weak.

filament is not used above that temperature. However, at a temperature of 1000 °C the filament would recrystallize within a minute.

Our investigations revealed that in the production of boron filament there is a critical temperature for the wire. Below this temperature only the desired amorphous form occurs, while above it the unwanted crystalline modification also occurs. With the wire stationary the critical temperature is 1000 °C. In a system in which the wire moves the critical temperature is higher and increases with the speed of the wire. *Fig. 7* shows the combinations of wire temperature and processing rate that were necessary to produce filament of a particular diameter in one of our experimental boron-deposition chambers. Filaments formed in the region above the dashed line are relatively weak and include patches of an abnormal structure (*fig. 8*) where the material has recrystallized. The explanation for this connection between critical temperature and wire speed is evident: the boron is obviously deposited in the amorphous condition and the faster it is removed from the chamber, the higher the permissible temperature.

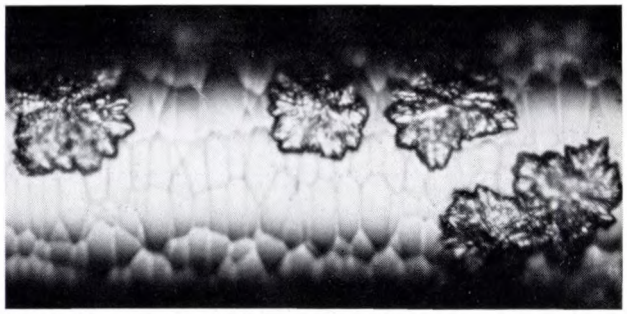


Fig. 8. The surface of a filament formed at a temperature above the critical temperature (the grey region in fig. 7). The boron in the patches of different structure is crystalline (β-rhombohedral).

In our tests with various boron-deposition chambers we found no conditions in which strong amorphous boron could be obtained at a wire temperature exceeding 1320 °C.

The fact that the critical temperature increases with increasing wire speed is of great importance in practice, because the strongest wires are obtained at the highest temperature (below the critical temperature, of course), and the temperature may clearly be high if the wire speed is high enough. Running the wire faster will also speed up the production and hence reduce costs.

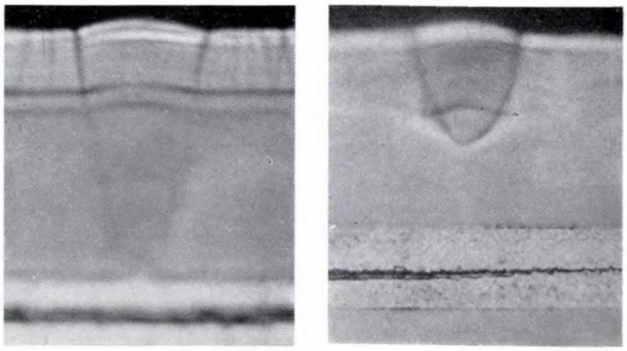
Nucleation and growth of nodules

The creation of a corn-on-the-cob structure like that shown in the title page, with large nodules, probably takes place in the following way. When a particular part of the wire enters the boron-deposition chamber, boron nuclei form at its surface. These grow in all directions until they cover the entire surface of the wire, and then they all grow in a radial direction until that part of the wire leaves the chamber. Every nodule is derived from a single nucleus.

Impurities in the surface of the wire or in the gas mixture lead to irregular formation of nuclei and hence to 'activated' growth of nodules (*fig. 9*). Appreciably larger smooth nodules then form at the surface (*fig. 10*). Filaments in which this development is observed are always relatively weak (σ_t less than 1500 N/mm^2). Flaws of this kind can be avoided by carefully cleaning the wire and the gas mixture beforehand.

In general, a higher formation temperature produces a stronger filament. The filament in *fig. 11a* was made at 1230 °C and has a tensile strength of 4000 N/mm^2. The one in fig. 11*b* was made at 970 °C and its tensile strength is 1500 N/mm^2. The weaker of these two filaments does not have the large perfect nodules of those seen on the title page, but a finer structure. The explanation is simple. It can be seen in fig. 7 that for the same processing time thicker filaments are formed at higher temperatures, i.e. the growth rate increases with temperature. If the number of nuclei formed per second is approximately constant, the growth of nodules at a low temperature will not keep pace with the formation of nuclei and growth will therefore keep starting again from new nuclei. If the gas mixture contains too much BCl_3, filaments of the type illustrated in fig. 11*b* are also formed at high temperatures; the rate at which nuclei form is evidently faster in that case.

The filament temperature must therefore be as high as possible provided it does not exceed the critical temperature; in other words, it should be just below the critical temperature. If for example a boron filament of



a *b*

Fig. 9. Longitudinal sections through filaments with activated growth of nodules, *a*) starting from the tungsten substrate wire, *b*) from an imperfection in the boron layer.

diameter 100 μm is to be made from a tungsten wire of diameter 12.5 μm, it is impossible to achieve this, even approximately, for all parts of the growing filament if the process is limited to a single boron-deposition chamber. This is because boron is a conductor at the high temperatures necessary here, so that the current density and hence the temperature in the thick filament at the lower end of the chamber would be much lower than in the thin filament at the top. It would therefore be impossible to deposit good-quality filament at the top and bottom of the chamber simultaneously. This is why the process is split up between two boron-deposition chambers (see fig. 4); the layer typically grows in diameter from 12.5 to 40 μm in the first and from 40 to 100 μm in the second.

Formation of tungsten borides in the core

The strength of a filament in which these various defects (figs. 8, 9, 10 and 11*b*) have been successfully avoided is determined by the core. During the production of the filament a transition area of tungsten borides is formed between the tungsten core and the boron layer, and in a broken boron filament it

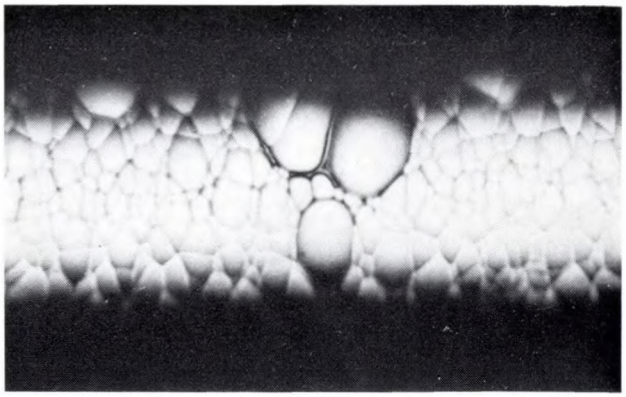


Fig. 10. Surface of a boron filament in which activated growth (fig. 9) has occurred.

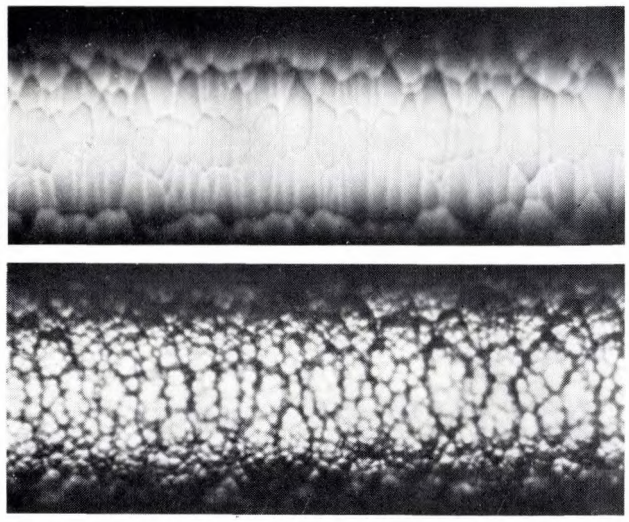


Fig. 11. *a*) Surface of a boron filament made at 1230 °C. *b*) As (*a*), but at 970 °C. The tensile strengths of these filaments were 4000 and 1500 N/mm² respectively.

is often possible to distinguish radial cracks that have obviously been initiated in this region (*fig. 12*). The negative influence of the core has been clearly demonstrated by tests in which extremely high tensile strengths (7000-11 000 N/mm²) have been attained by etching the core away [9].

By etching layer after layer from filaments and carrying out X-ray-diffraction analyses after every etching, we have found that the transition zone contains the entire range of tungsten borides —W_2B, WB, W_2B_5 and WB_4. As would be expected, the W_2B occurs close to the tungsten core, the WB_4 close to the boron layer.

The boride W_2B behaves rather differently from the others. If 2W is made to react with B to form W_2B, the atomic volume *increases* by no less than 50% [10]. Thus when W_2B forms, the core expands and easily causes cracks in the boron coating. When the other borides are formed, the atomic volume *decreases* by about 20% or less. Here the contraction of the core tends to pull the boron coating more tightly together, until it separates: this improves the strength of the filament. The absence of W_2B in the cores of very strong filaments confirms this.

As long as there is some tungsten remaining in the core the whole series of borides will continue to be formed during the production of the filament. Consequently, to eliminate the W_2B, the formation of borides must be allowed to continue until first the tungsten and then the W_2B have completely disappeared. This means that the process temperature must be made as high as possible and the processing time as long as possible within the range permitted by the conditions discussed above.

In principle, the elimination of the W_2B can be made more effective by starting with a very small quantity of tungsten, i.e. with a very thin wire. We made boron filaments starting from tungsten wires of various diameters and tested the resultant tensile strengths. The thickness of the finished boron filament was chosen in each case to give a density of 2.6 g/cm³. The results are shown in *Table II*. The filaments are indeed stronger with thinner substrate wire. Unfortunately, 6-μm diameter tungsten wire is so much more expensive than the usual 12.5-μm wire that the increased strength can only be obtained at appreciably higher cost.

Table II. Tensile strength σ_t of boron fibre as a function of the diameter of the tungsten substrate wire ($Ø_W$). The diameter of the boron filament ($Ø_B$) is selected in each case to make the density of the finished filament equal to 2.60×10^3 kg/m³.

$Ø_W$ (μm)	$Ø_B$ (μm)	σ_t (10^3 N/mm²)
6	50	4.75
8.9	70	4.09
12.5	100	3.75
18	150	2.90

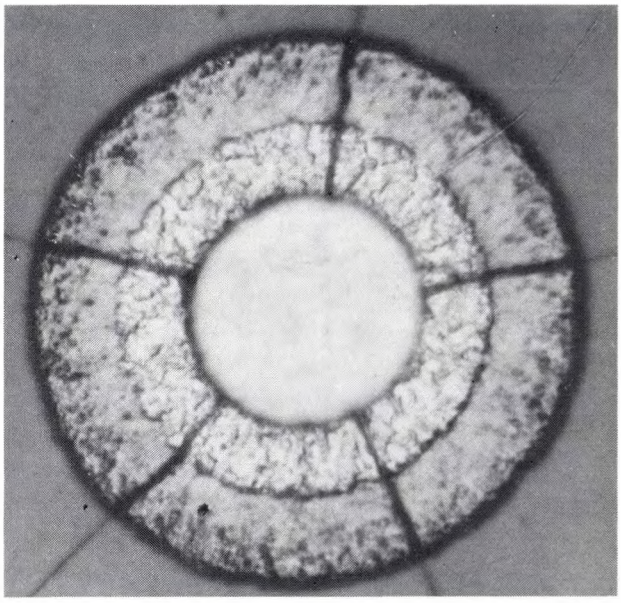


Fig. 12. Cross-section of the core of a boron filament. Starting at the centre, we have first tungsten, then various tungsten borides and finally boron. Radial cracks spread from the core, through the transition region consisting of borides, into the boron.

We also approached the boride problem in quite a different way — by depositing a layer of SiC on the tungsten wire ahead of the boron, thus forming a diffusion barrier between the tungsten and the boron. We did indeed manage to eliminate boride formation completely, but we were unable to deposit a layer of good-quality boron on the SiC.

[9] R. Kochendörfer and H. Jahn, Kunststoffe **59**, 859, 1969.
[10] R. P. I. Adler and M. L. Hammond, Appl. Phys. Letters **14**, 354, 1969.

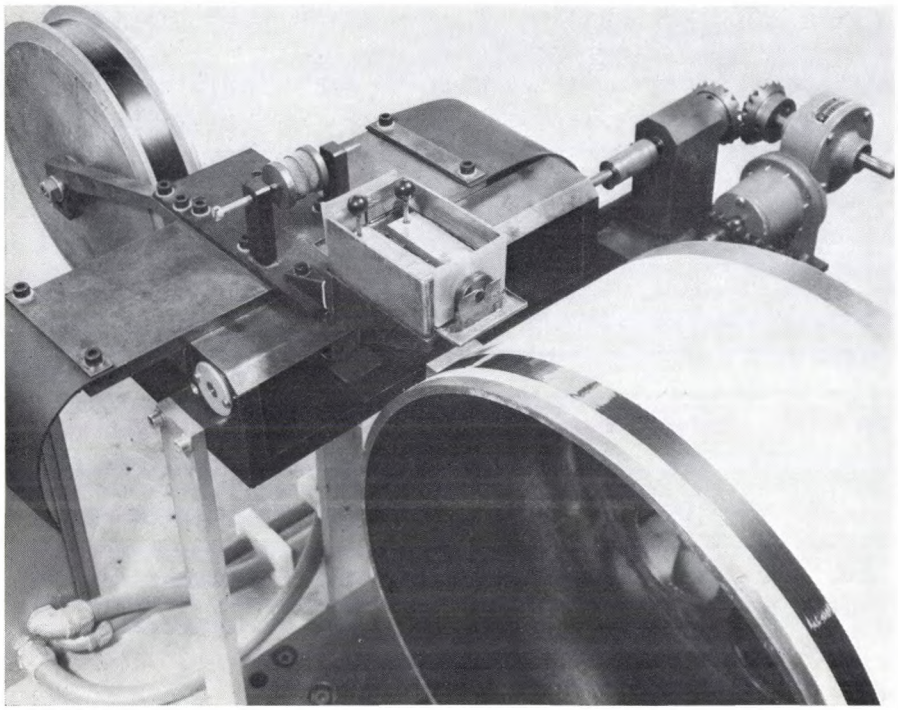


Fig. 13. Making in strips one filament thick ('pre-pregs') a boron-epoxy composite.

The strength of boron filament

The tensile strengths that we find in the filaments we have produced can be divided into three ranges clearly associated with the factors discussed above. The very weak filaments containing crystalline boron (fig. 8) will not be considered here.

We find tensile strengths of less than about 1500 N/mm^2 mainly in filaments in which activated, irregular growth has occurred (fig. 9). This appears as large smooth nodules at the surface (fig. 10), and can be attributed to irregularities such as thickness variations in the substrate wire or contamination on the substrate or in the gas.

We find tensile strengths between about 1500 and 3200 N/mm^2 mainly in filaments of the type illustrated in fig. 11*b*, in which, for some reason or other, nodule growth has lagged behind the formation of the nuclei, so that a perfect corn-on-the-cob structure is not obtained.

Finally, in filaments that are stronger than 3200 N/mm^2 the strength is limited principally by crack formation starting from the core as a result of the formation of tungsten borides. In the strongest filaments there is no W_2B in the core, as we saw earlier.

To summarize, we have successfully produced strong boron filament of homogeneous quality, on a laboratory scale, in a continuous process in reactors with two boron-deposition chambers. Flaws of the kinds illustrated in figs. 8, 9, 10 and 11*b* were avoided in the process by starting from tungsten wire with a thoroughly clean and smooth surface, and by a correct choice of the temperature, speed of the wire and composition of the reaction gas. The strength of the filament then obtained is at least some 3200 N/mm^2. We can also make filaments of appreciably greater strength (about 5000 N/mm^2) reproducibly, but only at much greater cost because of the much more expensive 6-μm tungsten wire that then has to be used.

Incorporating boron filament in composites

As stated earlier, epoxy resin is a suitable matrix for boron filament. A composite can be made from these two component materials in two stages. The first consists in making strips one filament thick, known as 'pre-pregs'. These are formed by passing the filament through a resin solution and winding it round a drum with a spacing of about 20 μm between turns (*fig. 13*). When the desired number of turns have been formed on the drum, a thin piece of woven glass-fibre is applied to provide the necessary lateral adhesion and make the pre-pregs into strips that can be handled easily. The resin is then cured by gently heating the drum. Finally, the pre-preg is cut through at right angles to the filaments and removed from the drum.

The next stage is to combine pre-pregs to form laminates. They are stacked in the desired fashion and the resin fully cured under pressure. If the pre-pregs are stacked with the filaments all in the same direction,

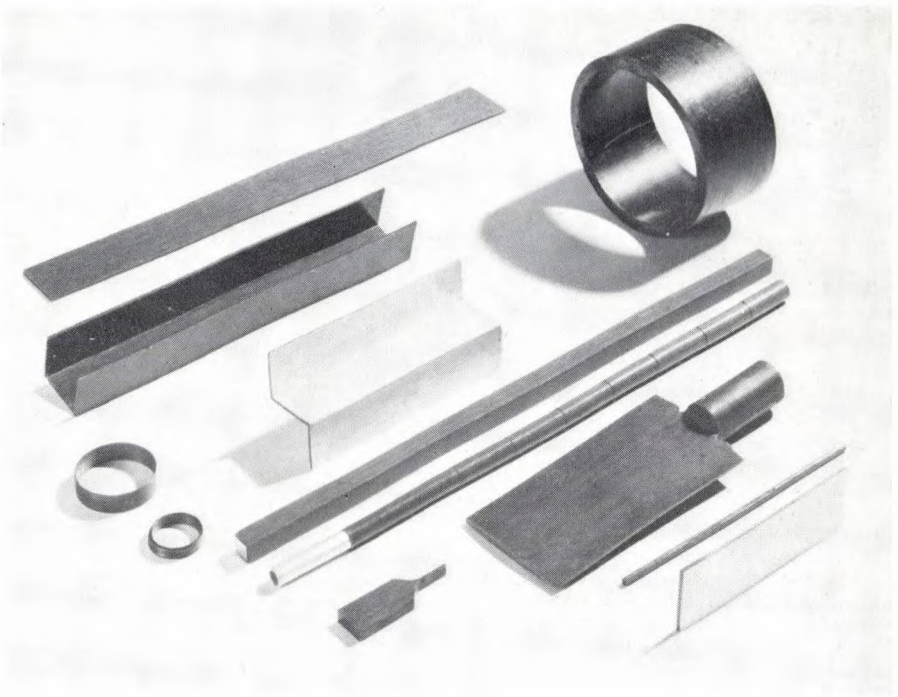


Fig. 14. Laminates of boron-epoxy composite and some structural elements reinforced with the composite.

a laminate is obtained that is extremely strong and rigid in one direction. The strength and stiffness can be 'spread' over a range of directions by stacking the successive pre-pregs with their filaments at an angle, usually a right angle. Since the pre-pregs are very flexible, laminates of a great variety of shapes can be made from them (*fig. 14*).

If it is desired to make objects with circular symmetry such as pressure cylinders, rods or shafts from composite materials, the resin-impregnated filament is sometimes wound directly in a number of successive layers on a jig of the desired shape (filament winding).

Examples of application

We will now briefly describe a few applications of our boron filament.

Interest in modern filament composites is almost entirely confined to the aerospace industry, because of their high cost. Most of our filament has also been made for that application. The international space-research programme includes the construction of a space laboratory or 'sky lab'. Its link with Earth will be maintained by a space shuttle. This vehicle will land like an aeroplane on return to Earth and will therefore be able to make the two-way trip many times. Each time it re-enters the atmosphere it will be subjected to very large forces. Under a contract with ELDO (the European Space Vehicle Launcher Development Organization), Fokker-VFW have been engaged in a study for the construction of the cargo door for the space shuttle. The door is to be about 20 m long and 5 m wide. An experimental model of a representative piece of the door, measuring about 55 by 75 cm, has been built and tested. It consisted of boron-filament composite coated with titanium on both sides. In view of our experience in this field Fokker delegated the development and production of the required composite to Philips. We have been able to meet the specification in full. Because of the frictional heat developed on re-entry the material has to be able to withstand a temperature of 350 °C for a total of 100 hours. For this reason the material chosen for the matrix was not the usual epoxy resin but polyimide resin, which is resistant to higher temperatures.

An entirely different type of application is in the wire-tension meters used in winding coils of copper wire. In automatic winding machines the wire feed has to be adjusted to make the tension in the wire almost constant. This is usually not difficult for cylindrical coils but for other shapes of coil — e.g. square ones — the wire speed varies so greatly that the feed is difficult to control. Wire-tension meters were made by Philips for determining just how the wire tension does in fact vary. In these meters the wire travels over the end of a leaf spring, whose deflection is converted into an electrical signal by strain gauges. To enable wire-tension changes to be transmitted almost instantaneously the spring has to be extremely stiff and light. In effect this means that its resonant frequency must

be much higher than the frequencies of the variations in the wire-tension, which are often of the order of 1000 Hz. The resonant frequency is inversely proportional to the square root of the specific stiffness. By making the spring of boron-filament composite the resonant frequency was raised to 6000 Hz. This case is characteristic of the many problems in which mechanical changes have to be transmitted very quickly. Such problems are enconntered in computer peripherals.

A final example is that of a long thin support arm that had to be stiffened. In the Philips Industrial Equipment Division the field of a cyclotron magnet had to be measured accurately with a Hall element in the course of some experiments. Because the area to be covered was large and there was limited vertical clearance, the Hall element had to be supported by a long thin non-magnetic arm with a minimum of deflection. These conditions were satisfied by employing a square aluminium tube 4 m long and with a cross-section of only 25×25 mm, stiffened with boron-epoxy laminates. Before stiffening, the arm deflected by about 35 mm under its own weight; the stiffening reduced this to about 3 mm.

Summary. Boron is a light and very stiff element and should theoretically also be very strong. Like other brittle substances, however, it will not take an appreciable tensile load in its normal form. This weakness is due to the notch effect of scratches, cracks or other imperfections. It most nearly approaches the perfect condition in the form of thin filaments. The strength of filaments is employed in modern composite materials, by embedding the filaments in a matrix such as epoxy resin. Of the known strong, stiff and light filaments — whose properties make them specially attractive for aeronautical and space applications — only boron and graphite filament have been widely tested.

In the investigation reported here on boron filament an electrically heated tungsten wire is fed continuously through a number of reaction chambers containing a gas mixture that reacts with the hot wire, depositing boron on it. In strong filaments the boron is practically amorphous and its surface shows a regular pattern of nodules, the 'corn-on-the-cob' structure. If the temperature is too high, crystalline boron is formed. Impurities on the tungsten substrate or in the reaction gas lead to activated growth and hence to large nodules at the surface. If the temperature is too low or the gas composition incorrect, nuclei form too quickly for the growth and the nodules at the surface exhibit a substructure. In the absence of these defects the strength is determined by the core, which generally also contains tungsten borides in addition to the tungsten. The borides include W_2B, which has a relatively large atomic volume and consequently has an adverse effect on the strength. Tungsten and W_2B can be avoided in the finished filament by starting from a very thin tungsten wire.

Some of the boron filament has been used in the construction of an experimental test-piece for the cargo door of a space shuttle.

Spiral-groove bearing systems with grease

J. G. G. Bos

If bearings are filled with grease, it is almost inevitable that a certain amount of air will become entrapped. In combinations of two or more spiral-groove [*] bearings, which should be considered as a single grease-circulation system, this can have very serious consequences when air accumulates at a place where it cannot escape. By analysing the pressure distribution in the grease of such a bearing system, a design can be produced in which this difficulty is avoided.

Lubrication 'for life'

Spiral-groove journal and thrust bearings have a number of attractive features. These include great rigidity, a pumping action that permits self-sealing, and a long-term performance better than that of porous bearings made of sintered and oil-impregnated metal [1] [2] [3]. Moreover, a journal bearing with grooves is more stable in operation than one with no grooves in the bearing surfaces [4]. For smooth journal bearings, it is known that when the load is zero or almost zero, the notorious (half-speed) whirl appears, unless special precautions are taken, such as the addition of two eccentric cylindrical surfaces. Preventive measures of this type are expensive or technically difficult.

The promising results with spiral-groove patterns encouraged us to apply them in a variety of complicated bearings developed at Philips Research Laboratories in recent years, and built up from combinations of simple bearings with different functions. Grease was chosen as the lubricant because greases leak away much less than oils, especially when the bearing is stationary. In some recent applications it is particularly important to prevent leakage of the lubricant. The addition of lubricant afterwards is sometimes completely impossible, for example in unmanned spacecraft, or is inconvenient and expensive — as in domestic equipment. The rate of leakage may then determine the appliance's useful life.

A grease may be considered as a Bingham fluid, i.e. a medium that has a yield point. When the shear stresses are below this limit, it behaves as a solid, and if they are higher, it flows and behaves more or less like an oil (i.e. a Newtonian fluid). For these values of the shear stress, there is an approximately linear relation with the shear rate; the magnitude of the viscosity need not then differ very greatly from that of an oil.

J. G. G. Bos, formerly with Philips Research Laboratories, Eindhoven, is now with N.V. Pope, Venlo, the Netherlands.

While investigating these compound bearings we encountered a particular difficulty not found in single spiral- or helical-groove bearings: the local accumulation of air in the grease. This forms the subject of this article. When the bearing is filled with grease during assembly it is almost impossible to prevent some air from being trapped. Also, air can sometimes enter the bearing while it is running. Generally the air in a single spiral- or helical-groove bearing can be readily eliminated by the pumping action and will not cause difficulties. However, in compound bearings, trapped air can easily accumulate somewhere inside the bearing, for example at places where there is a local minimum in the pressure of the grease. The air cannot escape from such a 'pressure dip'. If the pressure outside the system falls — as in a spacecraft directly after launching — or the temperature of the grease increases, the expansion of the air can force grease out of the bearing, with all the consequences for the life. In designing a system of grooved bearings it is therefore important to ensure — while taking the other requirements into account, of course — that any pressure dips occur at places that incoming air can escape from.

[1] For a general treatment see E. A. Muijderman, New bearing types: the gas and the spiral groove bearing, Philips tech. Rev. 25, 296, 1963 and E. A. Muijderman, New possibilities for the solution of bearing problems by means of the spiral groove principle, in: Lubrication and Wear 4th Conv. (Proc. Inst. Mech. Eng. 180, part 3K), 174-183, 1966. The name 'greaseber' has since been proposed for spiral-groove bearings containing grease (*grease*-lubricated, *self*-acting and self contained *bearing*).

[2] G. Remmers, Grease-lubricated spiral groove bearings for a straight-through shaft, Philips tech. Rev. 27, 107, 1966.

[3] G. Remmers, Grease-lubricated helical-groove bearings of plastic, Philips tech. Rev. 34, 103, 1974 (No. 4).

[4] G. G. Hirs, The load capacity and stability characteristics of hydrodynamic grooved journal bearings, ASLE Trans. 8, 296-305, 1965.

[*] In this article the term 'spiral' also refers to three-dimensional curves such as the helix.

There are further reasons why the presence of air bubbles must be considered undesirable. Air in the grease will diminish the load-bearing capacity of the compound bearing. It has also been found that a spiral-groove bearing becomes unstable rather easily when the grease does not fill the grooves right to the outer edge of the bearing surface [5]. A compound bearing could thus become unstable due to the presence of air/grease interfaces, and start whirling. The bearing surfaces then come into direct contact and rapid wear will occur.

A pressure dip of the type described above arises because the effect of the system of bearings on the grease flow and pressure is that of a system of interconnected pumps and flow resistances. We have now been able to produce equivalent circuits for such systems. Once such a circuit has been set up, it can be analysed in the same way as an electrical network by applying Kirchhoff's first and second laws [6]. *Fig. 1* shows the elements needed to represent a groove pattern in such an equivalent circuit; further details will be considered in the example which follows.

This method of analysis can be applied during the design phase to predict approximately where the trapped air will collect. Systematic modification of the design based on the circuit and on practical experience will then give an arrangement in which the air can escape and no grease need be lost.

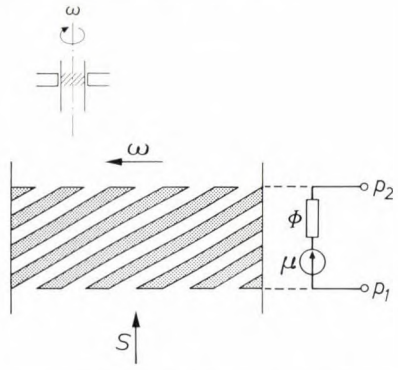


Fig. 1. The groove pattern in a spiral-groove journal bearing (schematic) and the equivalent circuit of such a bearing in relation to its pumping action. The circuit consists of a pressure source (pump coefficient μ) and a restriction (resistance coefficient Φ), connected in series. If the bearing surfaces rotate with a relative angular velocity ω, the bearing produces a pressure difference $p_2 - p_1$, which causes a mass flow S of the lubricant. The relation between these quantities is: $\mu\omega - \Phi S = p_2 - p_1$.

Air bubbles in grease

Before dealing with the example just mentioned, a little more should be said about the behaviour of air bubbles in the grease of a bearing. We have investigated this effect with transparent models of bearings. Sometimes very many small air bubbles appear in the grease; these bubbles are so small that they are carried along by the flows in the system. To eliminate these bubbles, it is sufficient to design the bearing in such a way that the flow — or some of it — crosses a point where the system is open. These very small air bubbles can then escape there without much leakage of the grease, provided that there is a physical mechanism for separating the air from the grease, such as a centrifugal force.

The grease can also contain air bubbles that are sufficiently large that they 'roll down' a pressure gradient (like the rising bubbles in a glass of beer) while the flow itself has little effect. These air bubbles are the ones most likely to accumulate at the pressure dip and cause the difficulties previously mentioned.

Apart from these two extremes there are of course many different types of air-bubble behaviour. Nevertheless grease leakage from a compound bearing can be almost totally eliminated by taking care that no pressure minima can occur in the system.

Bearing systems for 'flywheel' for attitude stabilization of space vehicles

A spiral-groove bearing for a stabilizing 'flywheel' (or momentum wheel) for space vehicles has been one of the cases to which we have applied our pressure-analysis method [7]. *Fig. 2* is a photograph of the flywheel and *fig. 3* gives a simplified cross-section. The wheel is connected to the bearing bush while the shaft remains fixed. One end of the shaft can be connected to the vehicle by a flange; the other end of the bearing is closed.

The bearing had to meet the following requirements:
— it must have a long life (7-10 years), and continuous operation without re-lubrication;
— the friction should be virtually constant;
— the positioning of the bearing bush should be accurate and invariable;
— it must take up, without wear, the extra load due to the slewing of the space vehicle;
— there should always be a continuous grease layer (before launching, too, for testing);
— it must be 'spaceworthy'.

Fig. 4 shows a cross-section of the first design of spiral-groove bearing for the flywheel. On testing this design a very small leakage of grease was found, which would have decreased the life of the bearing. *Fig. 5* shows the equivalent circuit of the design. It differs from an electrical network in that the mass flow represents the current and that there are three types of resistance: moreover, there are two different sources of pressure. The characteristic parameters and notation are summarized in *Table I*. From fig. 5 the mass flow S may be calculated from the equation

$$S\left\{\frac{\lambda_{23}}{|S|} + \phi_{23} + \Phi_{34} + \phi_{45} + \Phi_{51}\right\} = \\ = (\mu_{12}^* - \mu_{45}^*)\omega^2 + (\mu_{34} - \mu_{51})\omega.$$

Fig. 2. A 'flywheel' for attitude stabilization of a space vehicle. The shaft is fixed to the vehicle and the bearing bush supports the wheel. The drive motor is not illustrated.

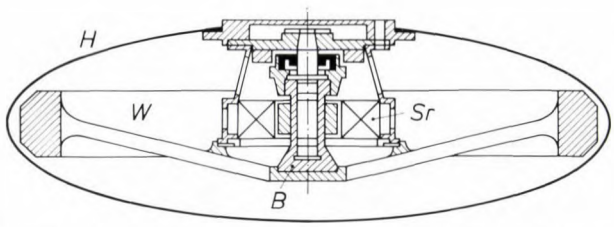


Fig. 3. A simplified transverse cross-section of the flywheel illustrated in fig. 2. *H* housing. *W* wheel, with bearing bush *B*. The stator *Sr* of the driving motor is concentric with the bearing bush.

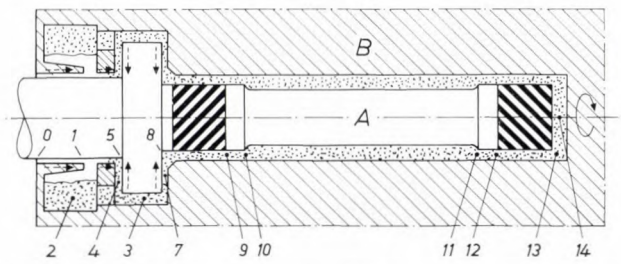


Fig. 4. A longitudinal cross-section of the first design for a spiral-groove bearing for the flywheel. *A* is the shaft, *B* the bearing bush. The lubricant is indicated by the dotted area. In addition to the two spiral-groove journal bearings indicated by the heavy stripes, there are also two spiral-groove thrust bearings (indicated by broken arrows in the pumping direction) and, on the extreme left, a radial-groove pattern that functions as the seal (broken arrows). The boundaries of the regions that function as pumps or restrictions have been denoted by a number (see text).

In this equation ω is the angular velocity of the bearing bush with respect to the shaft (fig. 4): the parameters μ, μ^*, Φ and ϕ are the pump coefficients — the asterisk indicates a centrifugal pump—the resistance coefficients for groove patterns and the coefficients for 'smooth' resistances respectively; the yield-stress resistance of the lubricant is expressed by the parameter λ. The

[5] J. Bootsma, the gas-to-liquid interface of spiral-groove journal bearings and its effect on stability, Trans. ASME F (J. Lubr. Technol.) **96**, 337-345, 1974 (No. 3).

[6] See for example P. Le Corbeiller, Matrix analysis of electric networks, Harvard Univ. Press, Cambridge (Mass.) 1950. The analogy with electrical circuits is already well known in a somewhat different form: see for example: A. G. M. Michell, Lubrication, Blackie, London 1950, in particular Appendix II, Investigation of bearing problems by means of physical analogies, pp. 292-298.

[7] The investigations of this bearing have been carried out as part of a contract for the European Space Research Organization (ESRO) in Paris and are described in J. G. G. Bos, Qualification tests on reaction flywheels supported on grease-lubricated bearings, ESRO-Contractor Report CR-87, July 1972, and J. A. C. Bollen, Further testing and design improvement of grease-lubricated spiral groove bearing momentum wheel, Final Report ESTEC-contract 1798/72 AA. An earlier version of the flywheel, with smooth journal bearings with eccentric cylindrical surfaces, has been described in J. P. Reinhoudt: A flywheel for the stabilization of spacecraft, Philips tech. Rev. **30**, 2, 1969. A detailed description of these bearings has been given in J. P. Reinhoudt: A grease-lubricated hydrodynamic bearing system for a satellite flywheel, Lubr. Eng. **26**, 95-100, 1970.

double indices give the positions (fig. 4) between which the elements may be localized for the computation of the parameters. Once the resulting mass flow is known at a given shaft speed, it is easy to determine the pressure as a function of position in the system. The mass flow is zero in the branch between positions *3* and *14* and the resulting (static) pressure distribution there should be considered as the analogue of the potential levels of a number of batteries in series but not delivering a current. The pressure distribution determined in this way for the system is shown in *fig. 6*. This pressure distribution indicates directly why the compound bearing of fig. 4 presented problems. Large air bubbles that have formed in the region *9-12* during assembly are quite likely to 'roll down' the right-hand side of the graph and arrive at the blind end of the bearing.

This fault has been corrected by the construction of a return channel that runs from *13-14* back to the vicinity of *5*, so that the air can escape from the blind end. A new analysis of the pressure distribution, with the return channel now taken into account, gave the dimensions of the channel.

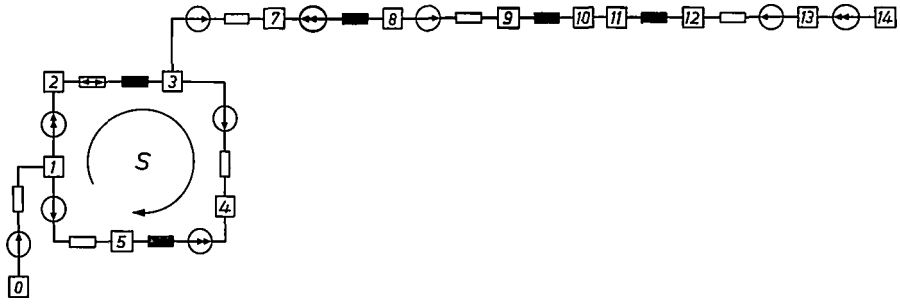


Fig. 5. The network of pressure sources and flow resistances forming the equivalent circuit of the bearing system shown in fig. 4; the numbers *0-14* in the squares correspond to the equivalent positions in fig. 4. Further details of the notation are given in Table I. In the branch between *3* and *14*, no yield-stress resistance has been included; this is permissible because its value is negligible compared with the flow resistance present there. Lubricant only flows in the ring part of the network, with a mass flow S; in the branch between *3* and *14* a static pressure distribution is established with zero mass flow. In the branches *3-5* and *3-8* there is some centrifugal pumping action besides the pumping action of the spiral grooves. The circuit has been modified to take account of this by including an extra pump in the branches *4-5* and *7-8*.

Table I. The active and passive components that can appear in a grease-lubricated spiral-groove bearing; the active components cause a pressure (pressure sources) and the passive ones a restriction (flow resistance). The arrow in the pressure source gives the direction in which the source pumps. S mass flow. ω the relative angular velocity of the bearing surfaces. The pump coefficients μ and μ^* and the resistance coefficients Φ and ϕ can be calculated for all components as a function of their dimensions and the viscosity of the grease. (These calculations, which were performed in part by G. Remmers, of these laboratories, were mainly based on the theoretical treatment in E. A. Muijderman, Spiral groove bearings, Philips Technical Library, Eindhoven, 1966.)

Symbol	Name	Equation [a]	Bearing component
	Active		
⊖ ; μ	hydrodynamic pump	$\Delta p = \mu\omega$	groove pattern
⊖ ; μ^*	centrifugal pump	$\Delta p = \mu^*\omega^2$	centrifugal seal
	Passive		
▭ ; Φ	groove resistance	$\Delta p = -S\Phi$	groove pattern
▬ ; ϕ	'smooth' resistance	$\Delta p = -S\phi$	'smooth' part of bearing
▭ ; λ	flow-point resistance	$\Delta p = -\frac{S}{\lvert S\rvert}\lambda$	grease

[a] the pressure difference $\Delta p = p_{right} - p_{left}$

A number of other necessary modifications were also made: point *8* was connected to the return channel just described, the pump region *1-5* was replaced by a

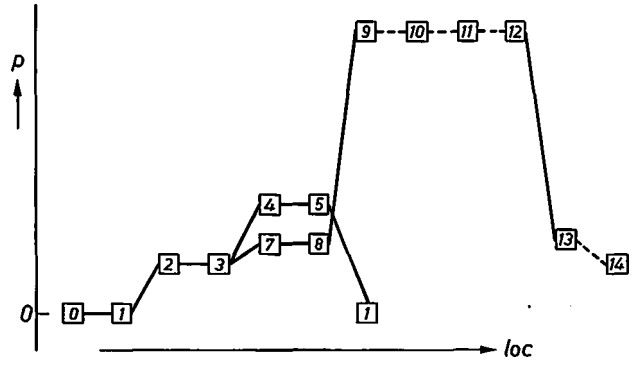


Fig. 6. The distribution of the pressure p in the bearing system of fig. 4, calculated with the aid of the equivalent circuit of fig. 5. The figures in squares correspond to the same locations (*loc*) in fig. 5. The pressure at position *0* is set equal to zero. In the regions where the air can accumulate from the grease, the pressure distribution is indicated by a dashed line. Separation of air and grease in the pump regions *8-9* and *12-13* can lead to the development of free edges (i.e. strips with no grease), with the risk that the 'half-speed whirl' will appear. In particular, the pressure drop at the closed end of the bearing (region *13-14*) is extremely undesirable.

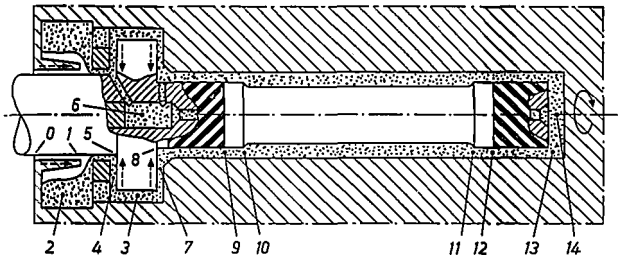


Fig. 7. Cross-section of the improved spiral-groove bearing system. It differs from the bearing illustrated in fig. 4 by the presence of a connection that allows the air that collects at *13-14* to escape to *5*.

gradient should occur not only at the nominal speed, but also at much lower speeds. If this is not the case, then eventually after each start and before each stop a situation will arise for some time in which the grease will behave locally as a solid; there will be no flow, and the consequences can be disastrous.

Fig. 7 shows a cross-section of the modified bearing and its equivalent circuit is given in *fig. 8*. The three mass flows S_1, S_2 and S_3 can be calculated using known methods with the aid of a 3×3 matrix [6]. The pressure distribution can again be determined from the

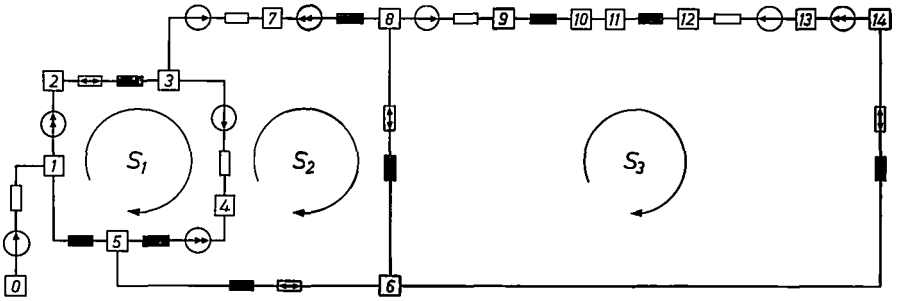


Fig. 8. Equivalent circuit of the bearing system of fig. 7.

smooth restriction, and finally the length of the pump region *8-9* was made larger than that of *12-13*, so that the first region receives the stronger pumping action once the flow pattern is defined.

These changes were necessary because the lubricant did not behave as a Newtonian fluid under all circumstances. In a grease-lubricated system of spiral-groove bearings, a suitable pressure distribution and flows in a straightforward way. This pressure distribution, which is given in *fig. 9*, approximates to the ideal situation. The only feature that might be questioned is the weak dip at point *3*. The flows there are so strong, however, that an accumulation of air at that point seems most unlikely.

To recapitulate the essentials briefly, the pumping action of a spiral-groove bearing in a properly designed compound bearing can simultaneously deliver the required load-bearing ability, rigidity and stability, with minimal wear (owing to the presence of the continuous lubricant layer), and provide an effective seal for the lubricant.

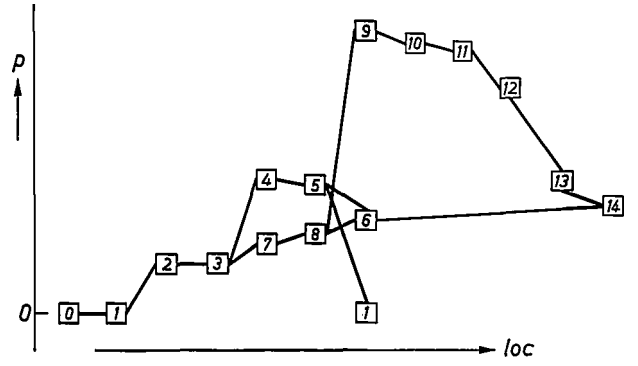
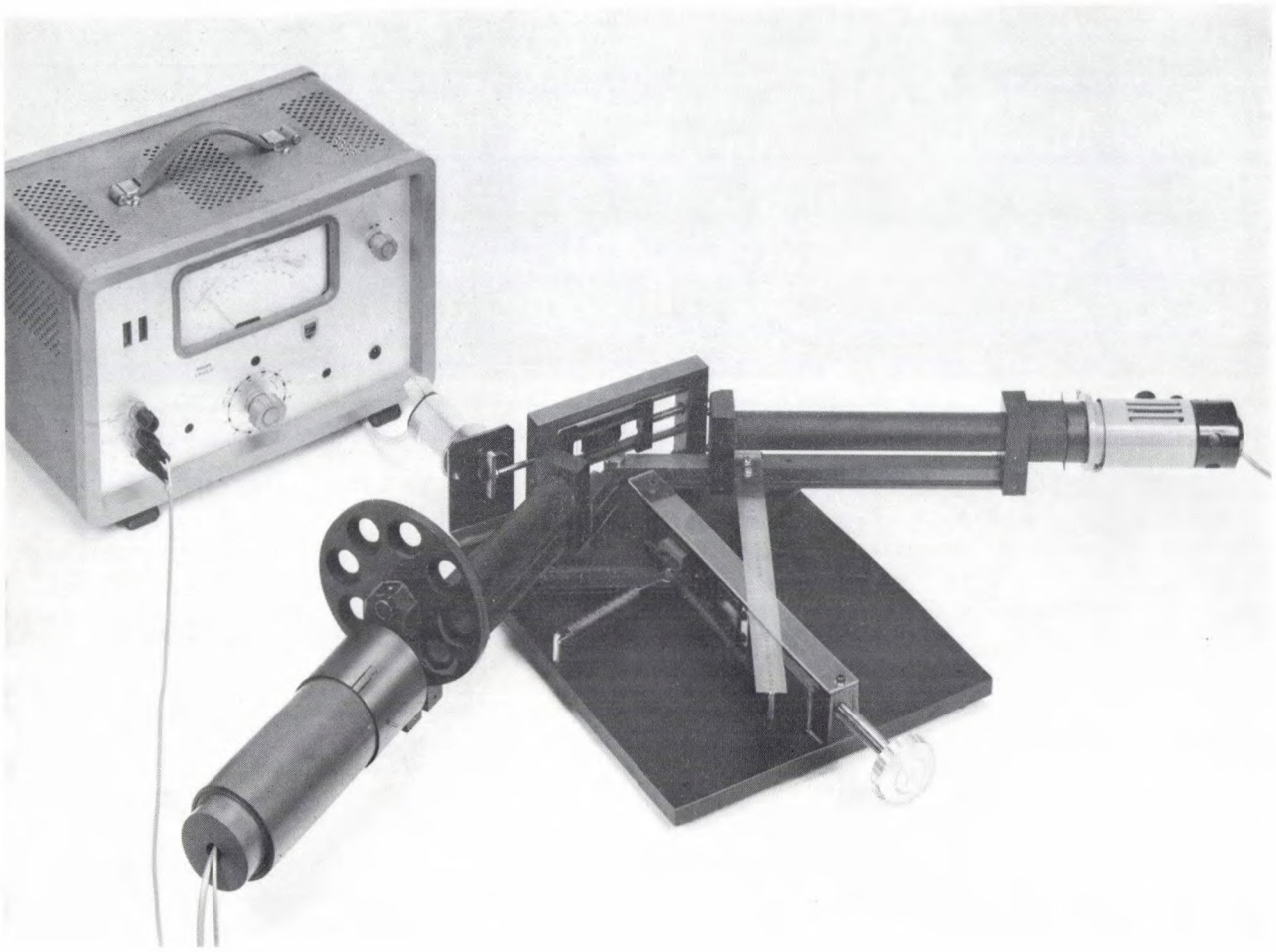


Fig. 9. Pressure distribution in the bearing system of fig. 7, calculated from the equivalent circuit of fig. 8. There is now no pressure drop at *13-14*, as there was in fig. 6. The slight drop at *3* does not present much of a problem, since the air will not be very likely to collect there because of the strong local flow.

Summary. A method of analysis has been developed that enables the pressure distribution — and consequently the air-removing characteristics — of a system of coupled, grease-lubricated spiral-groove bearings to be predicted. The system is considered as a number of pressure sources, connected by flow resistances. The mass flow and the pressure distribution may be calculated in the same way as the currents and potential differences in an electrical network. Even at the design stage, successive applications of the method yield improvements that ensure that any air in the lubricant can escape during operation without much loss of grease. Compound bearings designed in this way have a very long life and require no maintenance. As an example, a hydrodynamic bearing for a flywheel in a space vehicle (a project for ESRO) is described.

Philips tech. Rev. **35**, 142-143, 1975, No. 5



A fast refractometer for evaporated thin films

The measuring instrument shown above, which we have been using for several years, can be used to determine the refractive index of a non-absorbing material in the form of a thin film on a substrate at ten wavelengths between 0.2 and 2 μm. It takes less than five minutes to perform all ten measurements, and the uncertainty in each measurement is only a few units in the third decimal place of the measured value.

The method followed is a modified version of that due to F. Abelès [1]; the particular advantage of the new instrument is that it gives a short measurement time for the same accuracy. The method depends on the change in polarization obtained when light is reflected. If the film under investigation is illuminated with light incident at the Brewster angle, the reflected light only contains the component *s* for which the direction of polarization is parallel to the reflecting surface. The other component (*p*), polarized at right angles to *s*, then behaves as though the non-absorbing film was not present, and is reflected only by the surface of the substrate. The desired refractive index is equal to the tangent of the Brewster angle, as the theory of reflection and refraction shows [2]. The measurement therefore amounts to a measurement of the Brewster angle. Just how we do this in our refractometer — by a null method — will now be explained with the aid of *fig. 1*.

A substrate *S*, usually of glass, is partly coated with a film of the material under investigation [3]. The substrate with its film is then mounted in the refrac-

[1] F. Abelès, C.R. Acad. Sci. **228**, 553, 1949.

[2] See for example: F. Abelès, Methods for determining optical parameters of thin films, in: E. Wolf (ed.), Progress in optics II, 249-288, in particular p. 257, North-Holland Publ. Co., Amsterdam 1963.

[3] The part of the glass substrate that must remain uncoated is treated beforehand with magnesium oxide to prevent the film from adhering. This method was proposed by J. L. H. M. Wijgergangs of Philips Research Laboratories, Eindhoven.

[4] W. K. Burns and A. B. Lee, J. Opt. Soc. Amer. **64**, 108, 1974.

tometer in a kind of slide, which is moved slowly back and forth by the motor *Mo* and an eccentric. The light therefore falls alternately and regularly on the thin film or on the adjacent uncoated substrate, so that the photomultiplier *D* 'sees' alternately the coated and uncoated parts of the substrate. In the path of the reflected light there is an analyser *A*, adjusted to transmit only the component *p*. When the angle of incidence is made equal to the Brewster angle of the film material — by adjusting the set-screw *E* — the *p*-component of the reflected light will have the same intensity whether the incident light beam has been reflected by the film or the uncoated substrate. The pointer of the d.c. voltmeter *M* connected to the photomultiplier then remains stationary.

The *p*-component appears after reflection by the uncoated substrate because the beam is incident at an angle *different* from the Brewster angle of the substrate material.

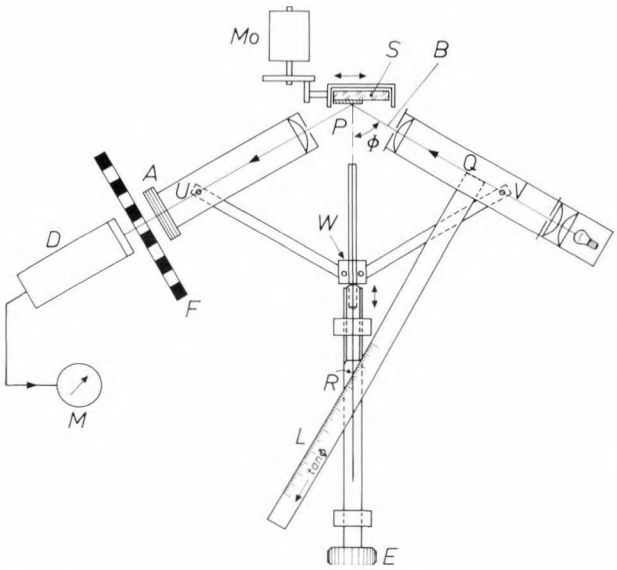
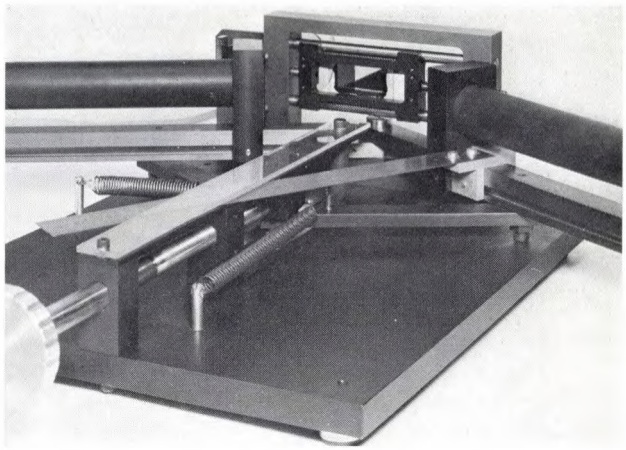


Fig. 1. The refractometer for determining the refractive index of material in the form of a non-absorbing film. A substrate (*S*), half of which is coated with a film of the material to be investigated, is placed in a slide; this slide is moved back and forth by the motor *Mo* and an eccentric during the measurement. The very fine parallel light beam *B* therefore falls alternately on the film and the adjacent uncoated substrate. The reflected light passes through the analyser *A* and one of the interference filters in the mount *F* to the photocathode of the detector *D*. The voltmeter *M* then usually gives an oscillatory indication. Only when the angle of incidence Φ is equal to the Brewster angle does the pointer become stationary. The tangent of the Brewster angle is equal to the refractive index of the film. Since *Q* is a right angle the refractive index can be read directly off the scale *L* at the point *R*. The adjustment of the angle of incidence by the knob *E* amounts to a symmetrical deformation of the rhombus *PUWV*, with the point *W* moving along the line *PR*. The deformation is kept free of backlash by tension springs. The mount *F* contains ten filters, permitting measurements at ten different wavelengths between 0.2 and 2 μm.



If the angle of incidence differs from the Brewster angle, the thin film also contributes to the intensity of the *p*-component and the voltmeter pointer is set into oscillation by the intensity difference. Since the natural frequency of the pointer (about 1 Hz) is equal to the frequency at which the slide moves back and forth, there is a clear observable indication even at very small intensity differences. The method is most sensitive when the film thickness is $\frac{1}{4}\lambda$ [4].

The Brewster angle can usually be found in less than 30 seconds with our refractometer. The refractive index is read off directly from the scale *L*. This can be done because the distance *QR* is one side of the right-angled triangle *PQR*, in which *PQ* has a constant length (100 mm). The scale can be read accurately to 0.1 mm, giving an uncertainty of a few tenths of one per cent for the value of the refractive index. The reproducibility is equally good. This accuracy is amply sufficient, since the variations that can occur in the refractive index because of variation in the evaporation conditions are generally larger.

The rotary mount *F* with interference filter permits the refractive index to be measured at ten different wavelengths in succession. Sufficient data can be obtained in this way for the complete dispersion curve of an evaporated film. This is important in making design calculations for ADI (All-Dielectric Interference) reflectors and filters, which are produced by stacking transparent films.

J. van der Wal

J. van der Wal is with Philips Research Laboratories, Eindhoven.

Thick-film technology

W. Funk

The well established screen process, used in the printing industry for more than half a century, has been found to be particularly suitable for making the interconnection patterns and components for hybrid electronic circuits. In the application of this process to mass production in the electronics industry, there were however certain special problems. Successful solutions to these problems have been found at Philips Forschungslaboratorium Hamburg.

The optimum application of fast digital integrated circuits requires an interconnection technique that will enable transit times and impedances to be obtained that are compatible with the switching frequency. Further requirements, such as those necessary in a multiple ECL-RAM [*] with a switching time of 2 μs, are high reliability, low weight, small size, and adequate heat transfer. All these requirements can be met by applying thick-film technology.

In fast line-printers for input and output of computer data, mechanical moving parts cannot be used if full advantage is to be taken of the speed of the actual computer. Both non-eroding spark electrodes and thermal-contact high-speed printing heads can be made with thick-film technology, giving an up-to-date answer to the problem.

Electronic equipment for home use now contains disposable plug-in units, to simplify maintenance. With thick-film technology the manufacture of these modules can be fully automated, thus improving reliability and reducing costs.

What exactly is thick-film technology? It can perhaps be described as a process in which all the electrical components — including conductors, crossovers, resistors, capacitors and inductors — can be produced on the same substrate by screen printing and firing [1]. Active and passive elements that cannot be printed are mounted on the passive circuit thus obtained, making it a hybrid circuit. With its encapsulation and connecting leads it then forms an electronic functional unit. Even though the process may appear simple, its success or failure depends on a large number of details.

The process

The various stages of the process are summarized in the diagram of *fig. 1*. The starting point in the manufacture of a hybrid circuit by the thick-film technology is an electrical circuit that is fully specified, tested and proven in operation. A scaled-up layout for the circuit is next made, either by graphic techniques or perhaps by computer techniques. Masks are then made from these enlargements, as in the production of printed-circuit boards.

The screens used in the printing are usually of stainless steel or nylon, with between 85 and 400 strands per inch (mesh width between 180 and 36 μm), depending on the application. Nylon screens are also used. The screen material is stretched across a rigid metal frame of about 200 × 200 mm, and then evenly coated with a light-sensitive emulsion such as gelatine with potassium dichromate. The prepared screen is then placed underneath the mask and illuminated from above. All the holes in the parts of the screen to be printed, e.g. at the places where interconnections are required, are then cleared by washing. At all the other places the screen remains sealed by the hardened emulsion.

For very accurate printing or for printing very long production runs, metal stencils attached to a stainless-steel screen by solder or adhesive may be used. For printing conductors with widths between 0.2 and 0.05 mm we have developed directly etched screens. The limit of reproducibility of printing by the thick-film method is 0.05 mm, both for the width and the spacing of the conductors. A number of screens are necessary for printing the complete circuit, each suited to the paste required for that particular part of the circuit. The substrate is then printed with these pastes in the correct sequence, using a screen-printer. The correct mask must of course be fitted to the screen at each printing operation.

Sintered alumina (Al_2O_3) with an impurity content smaller than 4% has been found to be the most suitable substrate material for thick-film circuits. Beryllium oxide (BeO) and soda-lime glass have also been used for special applications. In the present state of the

Ing. W. Funk is with Philips Forschungslaboratorium Hamburg GmbH, Hamburg, West Germany.

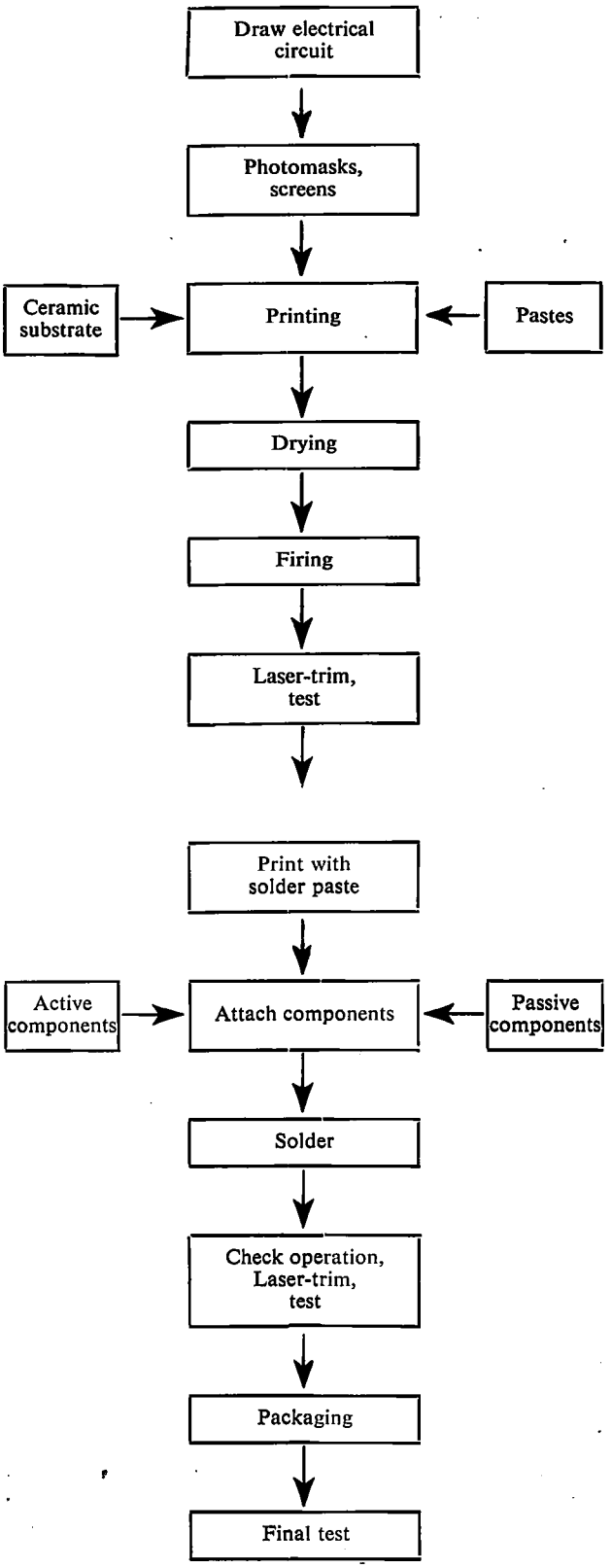


Fig. 1. Making a hybrid circuit by thick-film technology. This technology is a combination of screen printing, as used in the printing industry, and firing. The top part of the diagram refers to making the design for the circuits, making the masks and screens, printing and firing. The lower part refers to the attachment of the components that cannot be made by thick-film technology, further processing and testing.

technology substrates are available with dimensions of up to 50×50 mm, thicknesses of 0.5 to 1 mm and a surface roughness of 0.3 μm for the untreated ceramic. The first essential for success is to be found in the quality of the substrate material and the tolerances, particularly those relating to the flatness, thickness and other dimensions where deviations of as little as 1% can cause an appreciable reduction in the yield.

The pastes used in the printing consist of metals in powder form, usually noble metals, and a sintering agent, composed of glass frit, organic binders (resins) and organic solvents. The choice of metals and of the relative quantities depends on the application: e.g. conductors, capacitor electrodes, inductors or resistors. Gold is used for wiring if high conductivity is required, platinum/gold and palladium/gold if good adhesion and solderability are of greatest importance, and palladium/silver if costs are to be kept to a minimum. Materials used for resistor pastes include the oxides of ruthenium, iridium, thallium, palladium, indium and silver, mixed with other metal compounds and various percentages of glass. These resistor pastes are defined by the sheet resistance ('ohms per square') obtained at a nominal film thickness (25 μm after drying); the value of the resistance is determined by the length-to-width ratio. Designs are based on a typical dissipation limit of 40 mW/mm², but in fact this is a conservative value. The printed resistors are bonded to the conductors and the substrate when they are fired: they are called cermet resistors [**]. Various manufacturers can now supply pastes for resistors, with resistivity values in decade steps. Much has been said and written about these cermet resistors. Today the state of the art is such that resistors with resistance values from 1 Ω to 500 MΩ and temperature coefficients between 50 and 150 ppm/°C, depending on the composition and firing temperature, can be reproducibly produced from commercially available pastes.

Besides the pastes mentioned above there is also a range of glass/ceramic pastes for printing capacitor dielectrics and crossover insulation films.

For good reproducibility of the printing it is necessary to use a screen-printer that is highly stable, very accurate, and easy to operate. The speed of operation should be adjustable between 100 and 2000 prints per hour. All movements should be made by electronically controlled pneumatic-hydraulic systems. Such a machine operates in the following sequence:

[*] ECL = emitter-coupled logic.
RAM = random-access memory.

[1] A description of the screen-printing process as a printing technique and of its history is given in H. Ernst, Der Siebdruck, published by Lorenz Ellwanger, Bayreuth 1973.

[**] The term 'cermet' is made up from the words cer(amic) and met(al).

— The substrate is placed on a suction-operated holder;
— The screen is lowered on to the substrate;
— The paste is forced through the screen with a squeegee so that it makes contact with the substrate;
— The screen is raised;
— The substrate is transferred to the dryer.

This apparently simple process nevertheless has its problems; for example the screen only has to be raised through a distance of about a micron (the 'snap-off'). Moreover, the hardness of the squeegee, its angle of attack, the pressure and speed of the squeegee blade are all dictated by the viscosity of the paste, the method of preparing the screen, and the linewidth in the pattern to be printed. Even when all the parameters can readily be measured and pre-adjusted, the result still depends appreciably on the skill and experience of the operator.

The next process, drying for about 15 minutes at 125 °C, is not so critical; it can be carried out either in a drying box or a conveyor oven. After drying, the procedure can be repeated for another printed film. In this way the resistors can be printed and dried in successive decade steps on the substrate, and then fired simultaneously. The printed substrates are fired in a conveyor oven; the firing (temperature) profile is adjustable, with maxima between 600 and 1000 °C. The profile should of course always be selected to correspond to the recommendations for the resistor paste being used. The firing time is usually 1 hour. To obtain consistently reproducible results, it is necessary to stabilize the furnace temperature to within $\pm$ 2 °C at all temperatures. The production rate of the furnace is determined by the width and length of its conveyor belt; furnaces are available with belt widths of 100 to 1000 mm, lengths of 2 to 20 m, and three to fifteen hot zones.

After all the printing operations have been carried out — usually up to eight, each with the appropriate screen and paste and in the correct sequence — firing gives a passive circuit with conductors, resistors, capacitors and inductors.

Under ideal conditions the process will give resistance values that are accurate to within 10%. Since it is impossible to avoid small irregularities during ordinary production, and since tolerances of less than 10% are often required, the resistors must be 'trimmed' in practice. A resistance value can be increased by reducing the width of the resistor or by a local constriction of the current path. When the geometry is being designed the resistances are set at 75% of the desired value, so that in the worst case the value may have to be trimmed by 50%.

A few years ago the only method of trimming was to use a fine abrasive jet. Nowadays this method is gradually being replaced by laser-beam trimming. Computerized equipment is available for this, which can be programmed to work at a typical rate of 20 resistors per second and an accuracy of better than 1%, thus giving a large production of trimmed circuits. Printed capacitors can also be laser trimmed, although not so easily. The laser can also be used for scribing the ceramic substrate. A number of circuits can then be printed on a large substrate and later separated by breaking. Since the printed components are measured before trimming, test measurements and adjustment are combined, so that the values in the passive circuit are known before the active components are added.

To provide connections to other electrical components the circuit is printed at the appropriate locations with solder paste or conducting adhesive. Components can either be attached by hand, or by automatic preprogrammed machines, which can mount up to 20 components of different types on the substrate. Sometimes automated attachment is the more economic, sometimes attachment by hand; it depends on the rate of production and the availability of components suitable for attachment to hybrid circuits.

The multilayer capacitor with metallized ends is now an important passive component, since its suitability for automated component attachment — which is usually cheaper — outweighs the higher manufacturing costs. The small dimensions and the high reliability, compared with those of capacitors with wire leads, form another aspect of the economics of hybrid circuits; tantalum capacitors and fixed inductors in the form of chips and trimmable inductors are all available with minimum dimensions of 2 × 2 × 1 mm.

The situation for active components suitable for attachment to hybrid circuits is far less satisfactory. Since there is a shortage of small encapsulated components the usual practice, particularly in the United States, is to use unencapsulated chips, connected to the hybrid circuit by separate leads (wire bonding). The unencapsulated chips are cheap, which keeps the costs down, but there are problems with yield and reproducibility. This is why semiconductors in the conventional packages such as TO 5, TO 18 and TO 92, and integrated circuits in the TO 8 or DIL packages are still used in hybrid circuits. However, it does not seem consistent with the principles of microelectronics to attach a 1.2 × 2 mm component inside a 6 × 17 mm package merely for the convenience of using long-established bonding methods. This is why the discrete semiconductors in the SOT 23 and NO 149 packages are finding such success in Europe; they are available in increasing numbers and modifications. The same cannot be said of the SOT 43 package for integrated circuits. A method has recently been described for connecting an IC to its surroundings by means of a flexible plastic tape carrying a pattern of copper conductors [2]. In the United States 'flip chips' with solder bumps are being increasingly used as components for mounting in thick-film circuits. 'Beam-lead' devices (i.e. with short gold connector strips) are also available for special applications. The various versions mentioned meet the requirements for use in hybrid thick-film circuits: small dimensions, test measurement before mounting, and compatibility with soldering and automated attachment.

At a later stage of the manufacture the circuits are provided with the leads for the external connections, unless the circuit is a plug-in module, in which case the contacts will have already been printed. The circuits with leads then go through a soldering process in which time and temperature are accurately controlled. A continuous process is very useful at this stage. If the various components are attached together by a conductive epoxy-resin adhesive, the component-attachment stage is followed by a heat treatment ('baking').

Next, the circuit is tested for electrical operation. During the test faults may be rectified or a dynamic-trimming process applied, based on input and output parameters. The laser method is very suitable for dynamic trimming. Circuits at this stage of the process are shown in *figs. 2-5*.

The device is completed by encapsulating it. The range of methods extends from coating with epoxy resin (e.g. in a fluid bed) for consumer electronics, through mounting in nylon or moulded plastic packages for industrial applications, to hermetically sealed metal packages for 'professional' applications. These methods are well known in the electronics industry and are easily modified. The range of application of the circuits and the price determine the sort of encapsulation to be used.

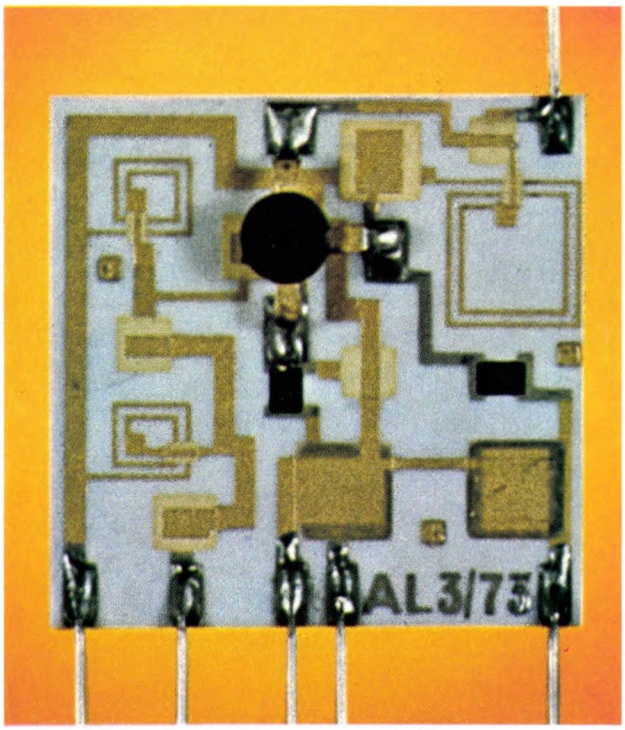


Fig. 2. Highpass UHF filter (Valvo GmbH). Two resistors (black), three inductors and seven capacitors have been printed on a ceramic substrate (Al_2O_3; 25 × 25 mm). The conductor leads cross at three places. A UHF transistor and six connecting leads can be seen attached to the substrate. The conductors and inductors are gold, the connector pads, the letters and numbers and the connections to the resistors are platinum/gold. Six printings and two firings were required in making this circuit.

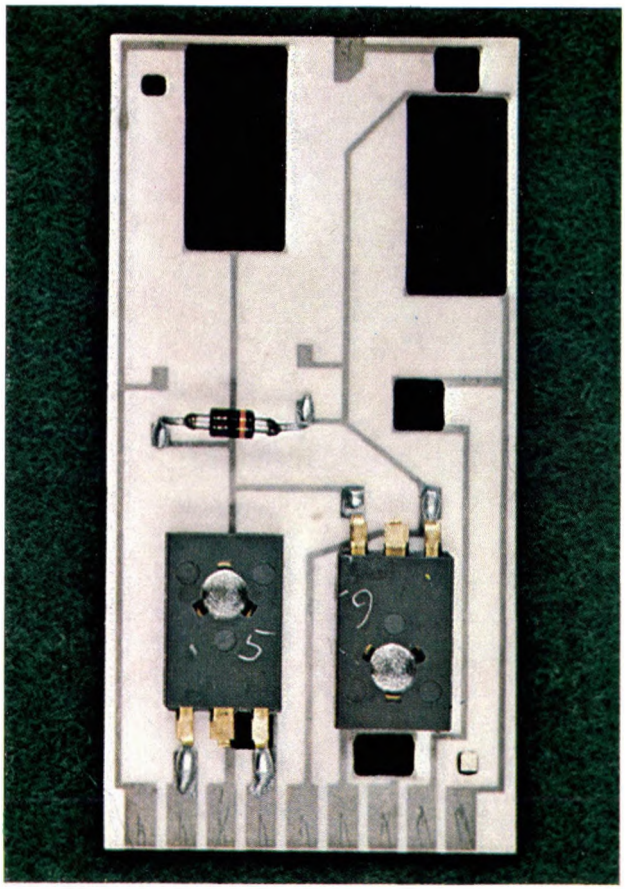


Fig. 3. Power amplifier (25 × 50 mm) for consumer electronics. The plug-in substrate carries six resistors (black), a diode and two power transistors. The two large resistors at the top are each rated at 3 watts.

Mass production for the electronics industry has its own special requirements

The processes briefly described here are the result of a more comprehensive study of film technology, carried out over a period of several years at Philips Forschungslaboratorium Hamburg. In seeking suitable technologies for interconnection in the electronics industry it was necessary to collect information about the thick-film process: this had in fact been known since the beginning of the thirties, but descriptions of the details of this process were rather vague. The information was necessary because an economic application of any production process requires an accurate knowledge of the physical, chemical and mechanical basis of the process. Much of the time was spent on a comparative study of the various ranges of pastes. In one part of these investigations alone, on equivalent resistor pastes made by three manufacturers, 64 000 resistors were printed and fired. The measurement programme

[2] A. van der Drift, W. G. Gelling and A. Rademakers, Integrated circuits with leads on flexible tape, Philips tech. Rev. **34**, 85-95, 1974 (No. 4).

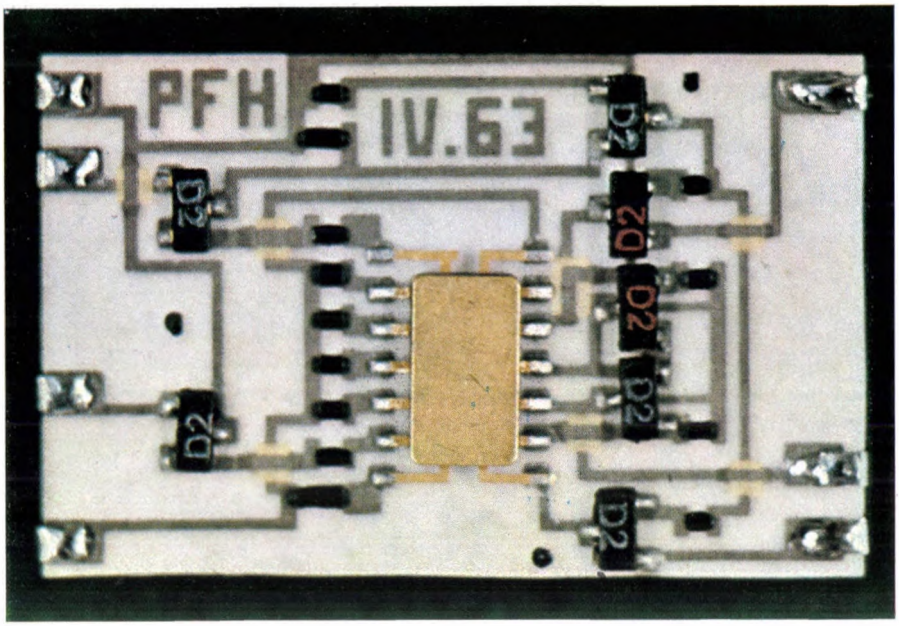


Fig. 4. Analog multiplier for industrial use. Besides the seven solder-bump connectors the substrate (18×28 mm) carries a monolithic IC in a flat package, seven SOT 23 transistors and fourteen resistors. The conductor leads cross at seven places.

and the evaluations of the results included the reproducibility, the size of the area, the temperature coefficient and the noise, and also the long-term stability. The correlation between all these quantities and the printing and firing conditions was closely investigated. Measurements were made with both trimmed and untrimmed resistors and with various conductor materials. The results were discussed with the manufacturers of material and equipment, to enable us to optimize our methods [3]. We found that the desired long-term stability could only be obtained for laser-trimmed resistors (which determine the norm) when these were fired with glass-frit pastes between 950 and 1000 °C. The same is true for the (long-term) adhesion and the conductivity of the materials for the conductors, and also for obtaining as large a dielectric constant as possible for the capacitor films. The important factor here is that the expansion coefficients of glass frit and that of the Al_2O_3 are almost equal.

Another question that has been the subject of a thorough investigation was the degree of compatibility of the pastes with the commercially available ceramic substrates. It was found that although high-purity Al_2O_3 ceramic had acceptable mechanical properties and was accurately dimensioned, its poor adhesion and high price made it unacceptable. In Al_2O_3 ceramic of purity less than 90% the additives were often found to be inhomogeneously distributed. Since these additives react with the paste material, the inhomogeneities introduce considerable spread in the resistance values. The material that most closely approaches the requirements for thick-film technology is Al_2O_3 ceramic of 96% purity, fired twice.

In many of the production methods used in the electronic industry the materials and equipment used were originally developed for other purposes. This is very much the case with the screens used in thick-film technology. The commercially available materials for preparing screens were in fact originally intended for use in the colour printing of posters and similar graphic

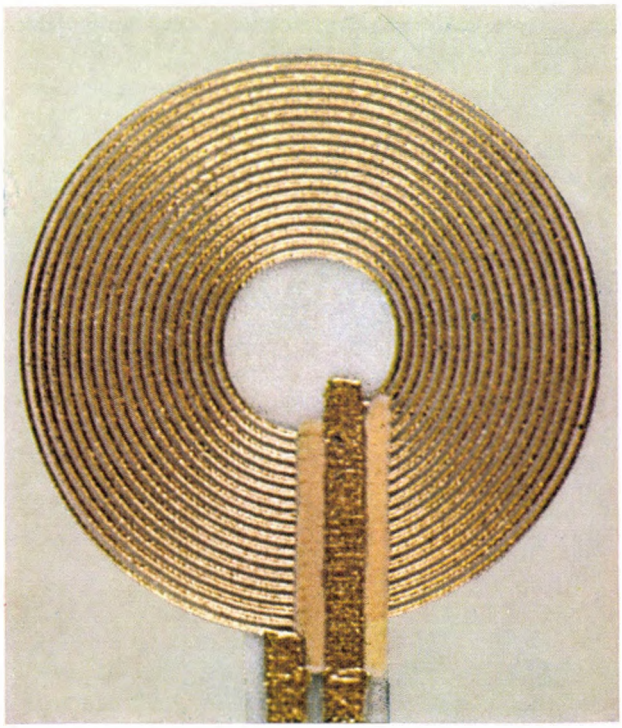


Fig. 5. FM modulation coil for a 2- to 3-GHz oscillator with YIG tuning. The gold spiral conductor is 60 μm thick and 100 μm wide, with a gap of 150 μm between the turns.

work. In that sort of application the numbers of copies required are small compared with the production required in the electronic industry. For our application the life of these screens was completely inadequate. It would be quite unacceptable to have to change a screen after only 2000 printings merely on account of the desired reproducibility. We therefore started by developing a screen for the production of fine lines and other details, which is suitable for the manufacture of prototypes. Screens of this type were necessary when thick-film technology was adopted for the manufacture of microwave integrated circuits [4]. A metal foil, tensioned on a screen frame, is etched by photochemical and electrochemical techniques in a fine grid pattern on one side and in the print pattern on the other. This produces a 'direct screen', which has the screen structure only on the parts to be printed (*fig. 6*). Care is taken to ensure a good wire-to-holes ratio and an even distribution of the holes, in particular at configurations that are curved or taper to a point. Linewidths and spacings of 50 μm can be obtained with such screens. We have also developed methods for the production of screens that can be used for 50 000 printings. This method is based on the combination of materials with suitable properties. Nylon screens, for example, have high elasticity and do not show fatigue effects as quickly as stainless-steel screens. However, the stainless-steel screens are more easily bonded with the normal process materials than the nylon screens. Calendered stainless-steel screens have the advantage that the structure does not change under tension, i.e. the holes remain of the same size everywhere. Metal foils have sharper edges and have a longer life than the ordinary screens prepared with gelatine and dichromate. In addition, they are not attacked by solvents and are more resistant to the abrasive effect of the pastes.

These various factors provided the basis for the design of a screen in which a metal foil of the desired thickness is firmly bonded to a stainless-steel screen and the print pattern is then formed on it by photochemical processes. A nylon screen is attached to the foil, which is a little larger than the substrate, and the compound screen thus produced is tensioned in the screen frame. These compound screens can be used to print very large production runs.

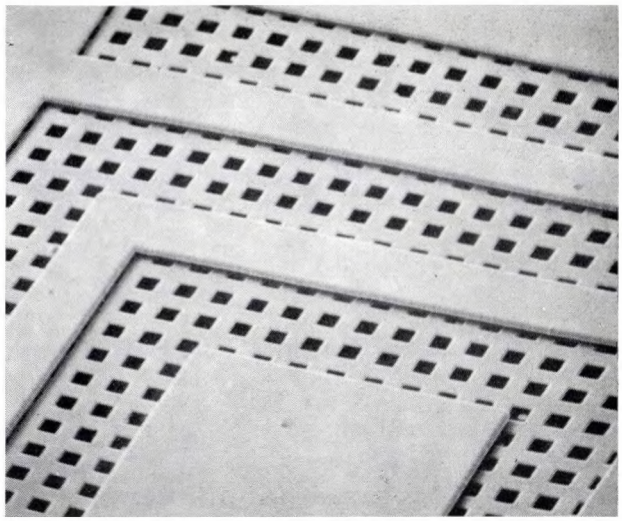


Fig. 6. Scanning-electron-microscope image of a directly etched metal screen. The holes are 30 × 30 μm.

Why thick-film technology?

No technology is used purely for its own sake, of course; it is the nature of the products to be made, and the economics of the process, that determine its use and possible wider acceptance. After one or two spasmodic attempts at the process in the years before 1950, hybrid circuits made by thick-film technology have been manufactured in numbers that have grown exponentially since 1960, when stable cermet resistor materials began to become available.

Starting with passive resistive circuits — to avoid having to connect individual resistors separately in printed circuits — the range of known applications grew apace, to include switching amplifiers in computers and electronic modules in communication equipment, and more recently, devices such as the heart pacemaker and modules for space vehicles. The applications are legion, and the future limits not yet in sight. Work on microwave circuits, disc triodes, travelling-wave tubes, large-scale displays and circuits for other optoelectronic equipment shows that thick-film technology is continuing to penetrate into other fields, and to their advantage.

If the incentives for applying thick-film technology are considered, the following general points arise:

— The reduction of the number of soldered joints, the sintering with a stable substrate material, and the use of noble metals make the circuit very reliable.

— The uncomplicated additive process, not tied to special requirements for purity, the wide range of resistance values that can be produced by printing, the ability to make capacitors, and the multilayer technique put the method in a very favourable position with regard to other methods for making hybrid circuits.

— Relatively little investment in machines is required and the equipment is adaptable.

---

[3] W. Funk, How to purchase thick film equipment and materials, Proc. Technical Programme Inter/Nepcon 71, Brighton.

[4] W. Funk and W. Schilz, Thick film techniques for hybrid integrated microwave circuits, Radio & electronic Engr. **44**, 504-508, 1974 (No. 9).

— The process can be almost completely automated. Reliable calculations indicate that labour charges amount to only 3 to 9% of the costs.
— The increasing scarcity of materials forces manufacturers towards the use of additive methods for making connections, i.e. only the strictly necessary amounts of material are used in making conductor tracks; they are not made by starting with much too large a quantity of material, most of which goes to waste. (The use of noble metals makes it essential to be economical with materials and to collect the waste.)
— The small volume and low weight also lead to savings, particularly in material.
— The process does not endanger the environment. No harmful waste products are produced.
— It is relatively inexpensive to hold stocks of modules, and for resistors the manufacturer is to some extent independent of the market. If the storage costs for 50 000 resistors from the E22 1/10 to 5.5 W ranges are compared with those for the 50 grams of resistor paste that could be used to produce 50 000 resistors, the cermet resistor is cheaper, even compared with carbon resistors.
— The good electrical properties of the ceramic material and the short connections, even for the active components used in hybrid circuits, give very good high-frequency properties, which permit thick-film technology to be used at frequencies up to 12 GHz.

Even with all these good features, thick-film technology is not yet as widely used in the manufacture of electronic circuits as printed-circuit boards. If screen-printing pastes based on non-noble metals can be perfected — and this seems likely in the near future — then thick-film applications will undoubtedly increase. This will coincide with coming changes that will compel the electronic industry to replace mechanical equipment and modules now in use by electronic versions. From the nature and number of the prototypes that have been produced by the laboratories for thick-film technology, it looks as though this will be the case in wired communication networks, and in the automobile and watchmaking industries. Changes like the move away from assembly-line working, the rediscovery of quality, the protection of the environment and the increasing scarcity of raw materials will force new production methods upon us. When that day comes, thick-film technology will be ready for use in the electronic industry as a connection method for monolithic integrated circuits.

Summary. Thick-film technology is a modification of the screen-printing process — used increasingly in the printing industry for some fifty years — combined with a ceramic firing process. The method meets the requirements placed on the interconnections in fast digital hybrid circuits. After a general description of the process the problems associated with application to mass production are discussed, with the solutions found. The most success have been obtained with ceramic Al_2O_3 substrates (96% purity), glass-frit pastes fired at 950-1000 °C, and specially developed screens. The many indications pointing to a rapid growth in the use of the method are summarized at the end of the article.

ERRATUM

In the article 'Electron resists for the manufacture of integrated circuits', by E. D. Roberts, Philips tech. Rev. **35**, 1975 (No. 2/3), fig. 6 (page 46) was unfortunately shown upside down.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or cooperate 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, 3 avenue Descartes, 94450 Limeil-Brévannes, 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*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.)	*N*

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, The Netherlands.

A. D. Annis & G. Simpson: Absorption of radiation in PLZT pyroelectric detectors.
Infrared Phys. **14**, 199-205, 1974 (No. 3). *M*

L. K. H. van Beek: Das PD-Verfahren. Fotografie für die Zukunft.
Schweiz. techn. Z. **71**, 553-559, 1974. *E*

H. Bernardet (SODERN, Limeil-Brévannes), **J.-F. Bonnal, J. Paulin & A. Pélissier:** Conception et expérimentation d'un propulseur ionique monofente.
Acta Electronica **17**, 421-433, 1974 (No. 4). *L*

H. Bernardet (SODERN, Limeil-Brévannes) **& A. Pélissier:** Emission électronique de filaments neutraliseurs sous bombardement ionique.
Acta Electronica **17**, 445-454, 1974 (No. 4). *L*

D. Bois & M. Boulou: Photocapacitance studies in high-purity GaAs.
Phys. Stat. sol. (a) **22**, 671-675, 1974 (No. 2). *L*

J.-F. Bonnal & A. Pélissier: Théorie de l'ionisation par contact.
Acta Electronica **17**, 375-382, 1974 (No. 4). *L*

L. Boonstra, C. W. Lambrechtse & R. H. W. Salters: Design considerations for a 4 k 1 MOS/bit RAM.
Onde électr. **54**, 175-180, 1974 (No. 4). *E*

H. Bouma (Institute for Perception Research, Eindhoven): Perception as a physical phenomenon.
Light and Sight, Anglo-Netherlands Symp. 1973, pp. 7-27; 1974.

J. Bouttes (ONERA, Châtillon) **& G. Eschard:** Les études sur le césium contact: un exemple de collaboration entre un organisme de recherches public et un centre de recherches industriel.
Acta Electronica **17**, 371-374, 1974 (No. 4). *L*

T. M. Bruton, J. C. Brice, O. F. Hill & P. A. C. Whiffin: The flux growth of some γ-Bi_2O_3 crystals by the top seeded technique.
J. Crystal Growth **23**, 21-24, 1974 (No. 1). *M*

P. Clément & J. Paulin: Technologie des ioniseurs de tungstène poreux.
Acta Electronica **17**, 405-414, 1974 (No. 4). *L*

R. D. Dawson, D. Elwell (both with Portsmouth Polytechnic) **& J. C. Brice:** Top seeded solution growth of sodium niobate.
J. Crystal Growth **23**, 65-70, 1974 (No. 1). *M*

H. Dötsch & H. J. Schmitt: Interaction of microwaves with ring domains in magnetic garnet films.
Appl. Phys. Letters **24**, 442-444, 1974 (No. 9). *H*

E. Fabre, R. N. Bhargava & W. K. Zwicker: Thermally stimulated current measurements in n-type LEC GaP.
J. electronic Mat. **3**, 409-430, 1974 (No. 2). *L, N*

E. Fischer & J. Fitzgerald: Heat and mass transport in cylindrical gas-filled incandescent lamps.
J. appl. Phys. **45**, 2895-2902, 1974 (No. 7). *A*

W. Guse (Universität Hamburg) **& D. Mateika:** Growth of mullite single crystals ($2Al_2O_3.SiO_2$) by the Czochralski method.
J. Crystal Growth **22**, 237-240, 1974 (No. 3). *H*

P. Hansen, P. Röschmann & W. Tolksdorf: Saturation magnetization of gallium-substituted yttrium iron garnet.
J. appl. Phys. **45**, 2728-2732, 1974 (No. 6). *H*

N. Hazewindus, J. M. van Nieuwland, J. Faber & L. Leistra: The magnetic analogue method as used in the study of a cyclotron central region.
Nucl. Instr. Meth. **118**, 125-134, 1974 (No. 1). *E*

L. Hollan & C. Schiller: Difference between the (001) facet and the vicinal planes in vapour phase epitaxial growth of GaAs.
J. Crystal Growth **22**, 175-180, 1974 (No. 3). *L*

W. H. de Jeu & Th. W. Lathouwers: Dielectric constants and molecular structure of nematic liquid crystals, I. Terminally substituted azobenzenes and azoxybenzenes.
Z. Naturf. **29a**, 905-911, 1974 (No. 6). *E*

Y. Kamp & J. P. Thiran: Maximally flat nonrecursive two-dimensional digital filters.
IEEE Trans. **CAS-21**, 437-449, 1974 (No. 3). *B*

A. W. C. van Kemenade & S. H. Hagen: Proof of the involvement of Ti in the low-temperature ABC luminescence spectrum of 6*H* SiC.
Solid State Comm. **14**, 1331-1333, 1974 (No. 12). *E*

A. J. R. de Kock, P. J. Roksnoer & P. G. T. Boonen: Effect of growth parameters on formation and elimination of vacancy clusters in dislocation-free silicon crystals.
J. Crystal Growth **22**, 311-320, 1974 (No. 4). *E*

C. Loty & J.-P. Roux: Détermination et optimisation des électrodes d'une optique ionique.
Acta Electronica **17**, 389-404, 1974 (No. 4). *L*

D. A. Lucas & R. P. Vincent: A portable air traffic control facility for MADGE.
SERT J. **8**, 135-136, 1974 (No. 6). *M*

F. Meyer: The ellipsometric method and its applications.
Silicates ind. **39**, 181-187, 1974 (No. 6). *E*

A. Molenaar, M. F. E. Holdrinet & L. K. H. van Beek: Kinetics of electroless copper plating with EDTA as the complexing agent for cupric ions.
Plating **61**, 238-242, 1974 (No. 3). *E*

A. Molenaar, J. E. A. M. van den Meerakker & J. Boven: The effect of polyoxyethylene and polyoxyethylene thioether compounds in electroless copper solutions.
Plating **61**, 649-653, 1974 (No. 7). *E*

A. E. Morgan: Ellipsometric studies of chemisorption on GaP(110) single crystals.
Surface Sci. **43**, 150-172, 1974 (No. 1). *E/H*

S. G. Nooteboom (Institute for Perception Research, Eindhoven): Eksperimentele bijdragen aan de fonologie.
Forum der Letteren **15**, 73-99, 1974 (No. 2).

L. G. Pittaway: The design and operation of a new extractor gauge for UHV pressure measurements.
Vacuum **24**, 301-305, 1974 (No. 7). *M*

T. J. A. Popma, A. M. van Diepen & J. M. Robertson: Lead substitution in yttrium iron garnet.
Mat. Res. Bull. **9**, 699-704, 1974 (No. 5). *E*

M. H. J. van Rijswick: Design of a computer program for off-line processing of gas-chromatographic data.
Thesis, Eindhoven 1974. (Philips Res. Repts. Suppl. 1974, No. 7.) *E*

J. M. Robertson, M. J. G. van Hout, J. C. Verplanke & J. C. Brice: Lead incorporation in thin iron garnet films produced by LPE and chemical analysis by the radioactive tracer technique.
Mat. Res. Bull. **9**, 555-567, 1974 (No. 5). *E, M*

B. Schneider: Magnetostatic wave drag on electrons in a hybrid sample of YIG and InSb.
Phys. Stat. sol. (a) **23**, 187-196, 1974 (No. 1). *H*

M. F. H. Schuurmans: Radiative decay of a pair of atoms.
Physics Letters **47A**, 493-494, 1974 (No. 6). *E*

G. B. Scott: Magnetic domain properties of $FeBO_3$.
J. Physics D **7**, 1574-1587, 1974 (No. 11). *M*

D. Sigurd, R. W. Bower (both with California Institute of Technology, Pasadena), **W. F. van der Weg** (Philips Research Labs., Amsterdam Division) **& J. W. Mayer** (Calif. Inst. Technol., Pasadena): Characterization of polycrystalline layers by channelling measurements.
Thin Solid Films **19**, 319-328, 1973 (No. 2).

M. Tasto: Motion extraction for left-ventricular volume measurement.
IEEE Trans. **BME-21**, 207-213, 1974 (No. 3). *H*

C. van Trigt: Comment on the paper 'Resonanzabsorption in Medien kleiner optischer Dichte'.
Z. Physik **268**, 265-266, 1974 (No. 2). *E*

C. H. F. Velzel: Inverse Fourier spectroscopy.
Thesis, Utrecht 1974. *E*

J. J. Vrakking & F. Meyer: Electron-impact-ionization cross sections of inner shells measured by Auger-electron spectroscopy.
Phys. Rev. A **9**, 1932-1937, 1974 (No. 5). *E*

K. Weiss: Phänomenologische Beschreibung des Materialtransports in einem elastisch verformten Kristall.
Z. phys. Chemie neue Folge **91**, 77-97, 1974 (No. 1-6). *E*

C. J. Wellekens: Simultaneous approximation of the attenuation and group delay of reciprocal lossless filters.
Proc. 1974 IEEE Int. Symp. on Circuits and systems, San Francisco, pp. 274-277. *B*

P. A. C. Whiffin & J. C. Brice: The kinetics of the growth of $KTa_xNb_{1-x}O_3$ crystals from solution in excess potassium oxide.
J. Crystal Growth **23**, 25-28, 1974 (No. 1). *M*

W. Widmer (ETH Zürich): Adaptiv kompandierende, differentielle Pulscode-Modulation zur Übertragung von Sprachsignalen.
Thesis, Zürich 1974. (Philips Res. Repts. Suppl. 1974, No. 8.)

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 6

The Philips PSEM 500 scanning electron microscope

W. Kuypers and J. C. Tiemeijer

The scanning electron microscope (or SEM) is based upon the same principle as the electron microprobe, which was described in the previous volume of this journal. Strictly speaking, the electron microprobe is a special version of the scanning electron microscope.

A scanning electron microscope for investigating the secondary electron emission of solids was described by M. Knoll [*] as early as 1935, and in 1938 M. von Ardenne [**] gave details of a scanning electron microscope of high resolving power that he had constructed. The present high quality of this type of microscope has been achieved through progress in electron optics and video techniques.

In designing the Philips PSEM 500 scanning electron microscope, described in this article, the main aim was to produce a compact instrument in which extensive automation and a number of appropriate safeguards provide the greatest possible ease of operation. A special feature that puts the PSEM 500 into the measuring-instrument class is the precision specimen stage for orientation and translation of the specimen in the microscope. The resulting equipment gives very satisfactory operation in the hands of a wide range of users, without the need for special training. A computer can also be connected, to give automatic operation and for processing the image information.

Scanning electron microscopy

Recent years have seen a steady increase in the numbers of photomicrographs, particularly in biological and metallurgical journals, that give a strikingly realistic picture of the specimen surface. This realism is the result of a large depth of field combined with very realistic shadow effects. These pictures, like the example in *fig. 1*, have been made with the scanning electron microscope, often referred to as the SEM.

The specimen in a scanning electron microscope is scanned with a narrow electron beam that traces out a line-raster pattern. One of the many types of radiation produced at the specimen as a result of the electron bombardment (the specimen responses) is then detected, and the intensity of the signal is shown on a monitor screen as a brightness modulation of a line raster swept in synchronism with the specimen raster. The ratio of the dimensions of the two rasters gives the magnification of the microscope. The magnification can

Fig. 1. Scanning-microscope picture of plastic foam, coated with an extremely thin gold film for this picture. The depth of field is clearly much greater than the resolving power. This is different from optical microscopy, where the two quantities are approximately equal. Magnification 75×.

Ir W. Kuypers and J. C. Tiemeijer are with the Philips Scientific and Industrial Equipment Division, Eindhoven.

[*] M. Knoll, Z. techn. Physik **16**, 467, 1935.

[**] M. von Ardenne, Z. Physik **108**, 338, 1938 and **109**, 553, 1938.

therefore be varied merely by varying the amplitude of the scanning movements of the beam over the specimen; the dimensions of the raster on the monitor screen are usually kept constant.

Images can be made not only by detecting electrons, either secondary, reflected or transmitted, but also with X-radiation or cathodoluminescence.

Fig. 2 is a schematic diagram of a scanning electron microscope. The specimen is shown at the bottom; it is usually mounted on a specimen stage that allows it to be rotated, tilted, or displaced in any direction. This specimen stage will be referred to as a 'goniometer' from now on, although it is not limited to the measurement of angles. Detectors for the specimen responses are located around the specimen. Above the specimen is the electron-optical column. This consists of an electron gun and a number of lenses, which produce a reduced image of the electron source on the specimen. Two sets of deflection coils make the beam scan the specimen surface in the desired raster pattern. The sawtooth currents in these coils are synchronous with the deflections that produce the raster on the monitor screen. The electron trajectories are of course described in a vacuum, which includes the specimen and the electron detectors.

The contrast in the images obtained with the scanning electron microscope can arise in various ways. The output of both secondary and reflected electrons depends on the angle from which the detector sees the surface (*fig. 3*). It also depends on the atomic number of the material encountered by the electrons and on the crystallographic orientation of the specimen surface (*fig. 4*). The absorbed electrons give a signal that in many cases is exactly the complement of the reflected electrons, at any rate for a thick specimen, as *fig. 5* shows.

The characteristic X-radiation from the different elements in the specimen gives a picture of the distribution of these elements, as in the electron microprobe [1]. Finally, it is also possible to detect the Auger electron peaks of the elements in the sample, by selecting the secondary electrons according to energy. This provides information about the chemical composition of the outer atomic layers of the specimen [2]. Energy selection also enables a picture to be obtained of the potential differences at the surface of the specimen, which can be very useful in investigations on integrated circuits (*fig. 6*).

It can therefore be seen that there are many possible ways of obtaining contrast in an image, so that special preparation techniques to improve contrast are usually unnecessary. However, it is necessary to deposit a thin conducting film on specimens that are electrical insulators, to prevent charge build-up on the surface. Charging effects can also be avoided by setting the energy of the primary electrons at a value for which the secondary-emission factor is exactly unity; the primary-electron energy necessary for this depends on the composition of the specimen; it is usually about 1.5 keV.

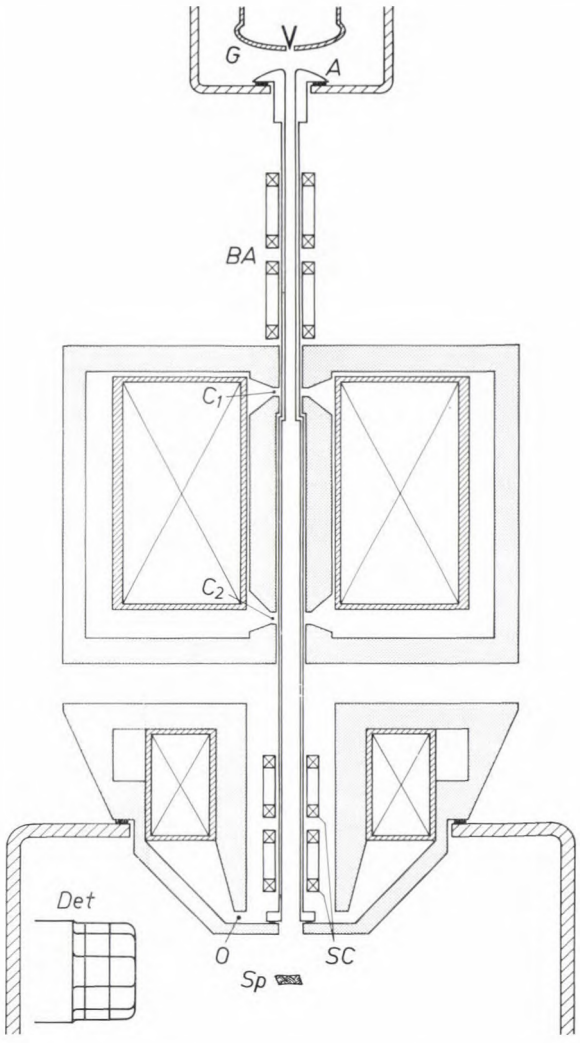


Fig. 2. Diagram of a scanning electron microscope (SEM). *G* electron gun. *A* anode. *BA* beam-alignment coils. C_1, C_2 double condenser lens. *O* objective. *SC* scanning coils that produce the raster. *Sp* specimen. *Det* detector for secondary electrons. The vacuum chamber is enclosed by the wall of the emission chamber (which also includes the gun), the demountable inner lining of the electron-optical column and the wall of the specimen chamber.

The images obtained with a scanning electron microscope reproduce the specimen as it would appear from the direction of the incident beam, while being illuminated from the detector. It is therefore the same view as that obtained through an optical microscope with

[1] A description of the electron microprobe is given in M. Klerk, Philips tech. Rev. **34**, 370, 1974 (No. 11/12).

[2] See for example H. H. Brongersma, F. Meijer and H. W. Werner, Philips tech. Rev. **34**, 357, 1974 (No. 11/12).

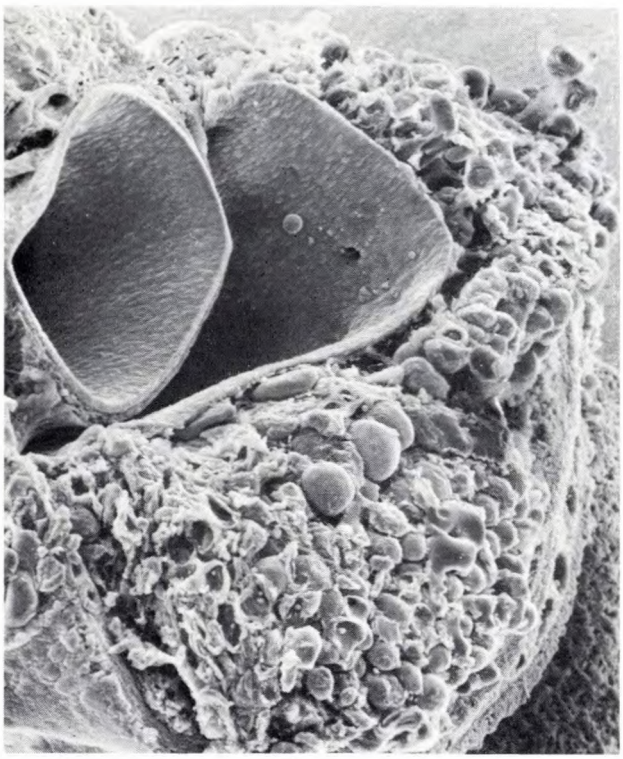


Fig. 3. SEM picture of an anatomical specimen that has been coated with a thin gold film. The contrast is entirely determined by the shape and structure of the surface. Magnification 95×.

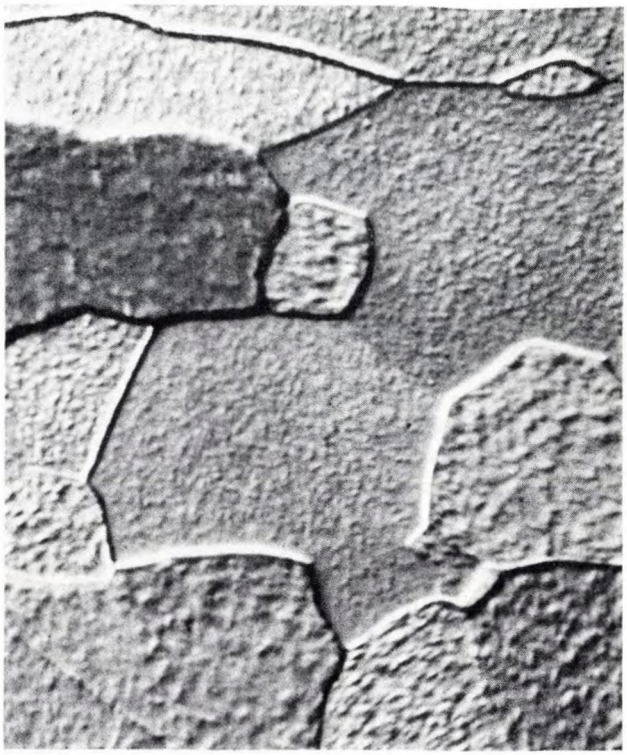
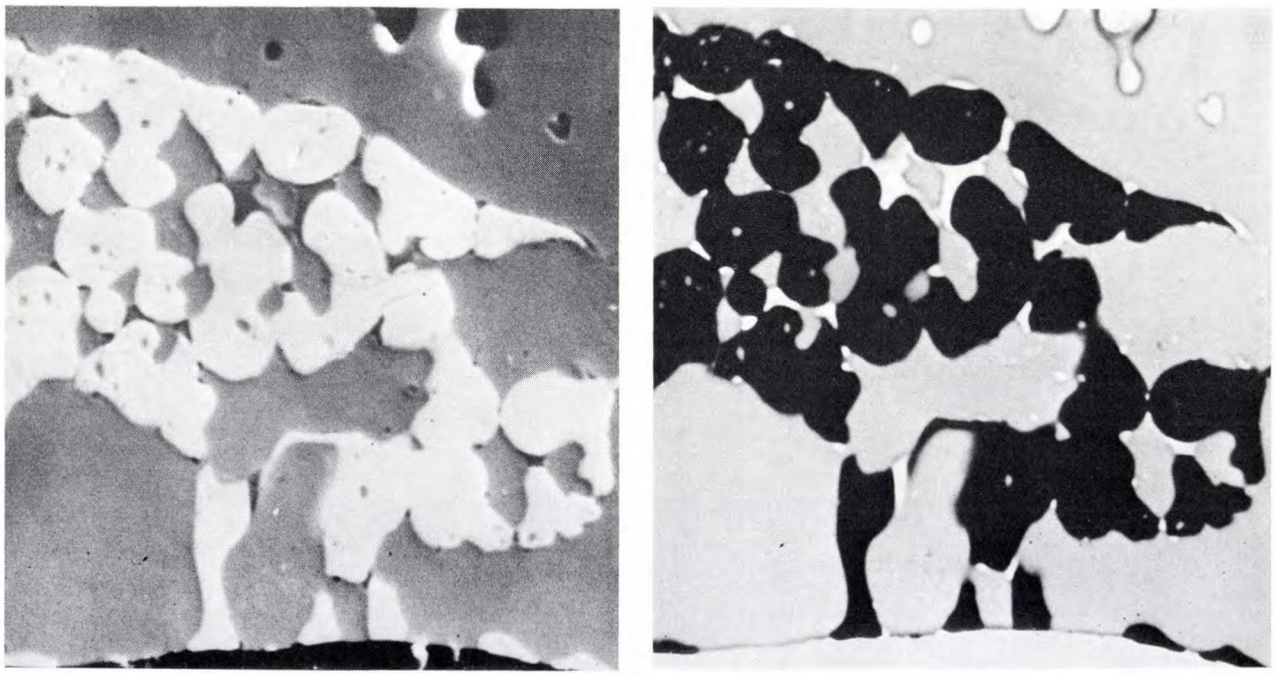


Fig. 4. SEM picture of a polished and etched tungsten surface. The grain boundaries are accentuated by shadows resulting from differences in height arising during the etching. The differences in the shade of grey are due to differences in crystal orientation between the grains. Magnification 575×.



a *b*

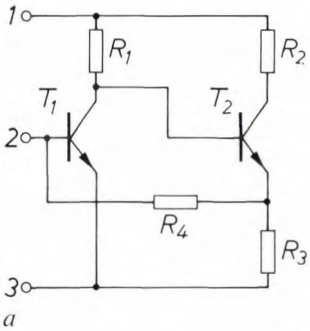
Fig. 5. Two SEM pictures of the same specimen. *a*) Picture obtained by detection of the back-scattered electrons. *b*) Picture obtained by detection of the absorbed electrons (specimen current). The two pictures in this case are almost exactly complementary. The specimen is the cross-section of a rod of a Co-Cr-Fe alloy that has been in contact for a long time with glass containing antimony. Antimony (the white areas on the left, the black ones on the right) has diffused into the metal. Magnification 830×.

light incident from one side. The production of this image is explained in more detail in *fig. 7*.

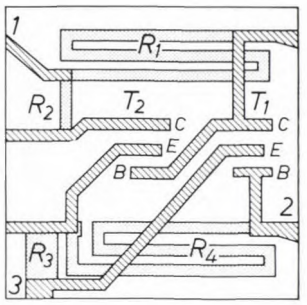
Since the aperture angle of the beam is small, the depth of field is large. This depth of field is a considerable advantage here, as in the examination of replicas with a transmission electron microscope: sharp images can be obtained for the complete depth of a relatively rough surface. However, in transmission microscopy this large depth of field can be a nuisance when fairly deep sections are examined. It is then necessary to resort to special techniques (tilting, stereo images) to establish the spatial relationship of the observed details.

The resolving power of a scanning electron microscope of the type to be described here is not entirely determined by the optics (lens defects, lens aperture and the smallest spot diameter). The diffusion length of the electrons in the specimen also has an effect; this length decreases with the energy of the incident electrons. No simple relation can therefore be given for the resolving power. A limit can however be quoted: at an electron energy of 25 keV and with a normal thermionic cathode the resolving power is about 7 nm in the most favourable case. For special transmission scanning microscopes much smaller values (about 0.3 nm) have been quoted in the literature [3].

Image quality is also affected by the signal-to-noise ratio. At high magnification a small spot is necessary to obtain a sufficiently high resolving power. The current density in the scanning beam is fixed, however, by the brightness of the cathode in the electron gun, the aperture angle of the beam, the accelerating voltage and the transmittance of the optical elements. At constant aperture angle and accelerating voltage the current in a small spot will be small, and the detector signal will also be small, so that the signal-to-noise ratio will be poor. There are three sources of noise: the current noise in the scanning beam, the current noise in the secondary-electron current and the inherent noise of the detector. The noise from all three of these sources can be reduced by integrating the electron signal over a time interval, which means that a low scanning rate should be used. *Fig. 8* gives an impression of the improvement



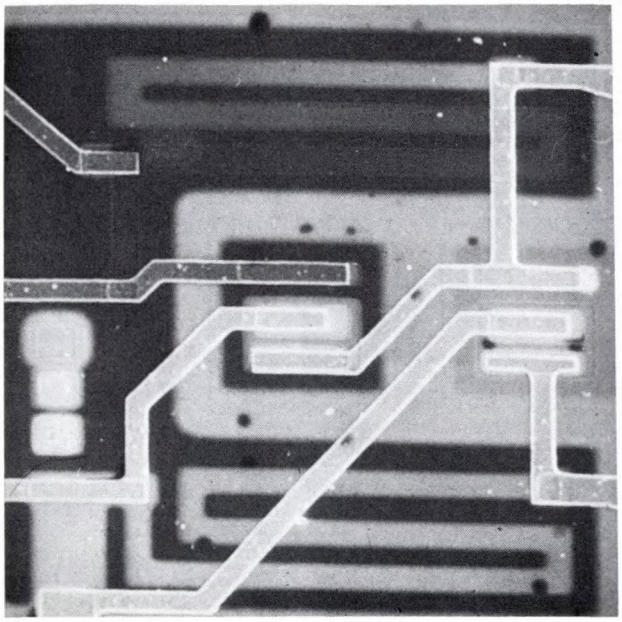
a



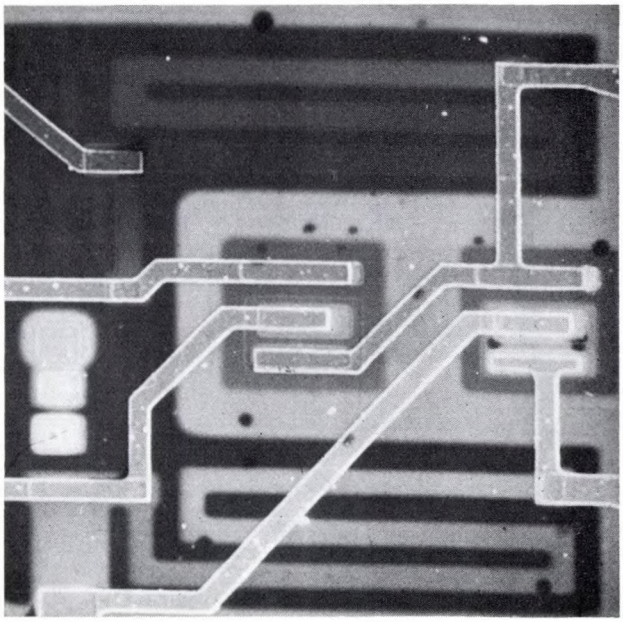
b

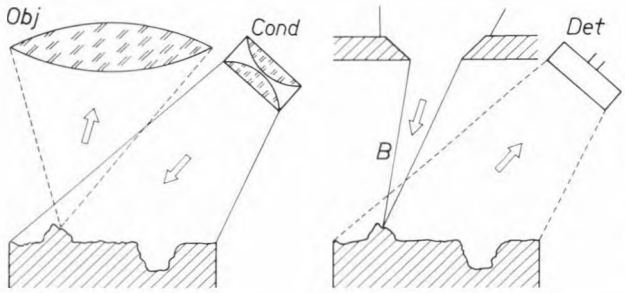
Fig. 6. SEM pictures of an integrated circuit. *a*) Electrical circuit diagram. *b*) Arrangement of the circuit elements. *c*) Picture of the circuit with voltage contrast. The terminals *1* and *2* are connected externally through a high resistance. T_1 is conducting, there are no potential differences in the transistor, and it is all at the same negative potential. The full supply voltage appears across R_1, as can be seen from the varying shade of grey. The base of T_2 is therefore connected to the negative side of the voltage source. T_2 does not conduct, so that potential differences can be seen in this transistor. R_2 carries no current, as can be seen from its even shade of grey. *d*) Here an external connection has been made between terminals *2* and *3*, so that T_1 does not conduct and potential differences appear across it. T_2 now conducts, but takes up a potential between those of the two supply terminals because of the voltage distribution across R_2 and R_3. The resistance R_2 now takes current; R_1 takes only a very small current via T_2 and R_3. A small voltage appears across R_3 and R_4. Magnification 150×.

c



d





The objective *Obj* produces an image of the specimen surface; each point of the surface corresponds uniquely to a point of the image.	The beam *B* scans the specimen surface; there is a unique correspondence between the scanned points on the surface and the points in the image.
The condenser *Cond* concentrates the light from a light source on to the specimen surface, illuminating the whole surface.	The detector *D* receives electrons leaving the surface.
The topography of the surface produces shadows.	The topography prevents electrons originating at some points of the surface from reaching the detector.
The brightness at which a point is reproduced in the image depends upon: —the illumination level at this point —the reflection characteristics of the surface at this point.	The brightness at which a point is reproduced in the image depends upon: —the number of electrons from this point that reach the detector —the electron-emission characteristics of the surface at this point.

Fig. 7. Production of an image in an optical microscope with oblique illumination (*left*) and with an SEM (*right*). Although the information for producing the image is collected in one case by the lens (*Obj*) and in the other by the detector (*Det*), the function performed by the condenser (*Cond*) and lens is closely similar to that performed by the beam (*B*) and the detector. Lens and beam are each responsible for a one-to-one correspondence between object and image, while condenser and detector each cover the complete specimen surface. This close correspondence accounts for the formation of analogous images in the two cases: a view of the specimen from the direction of the optical axis of the lens (axis of the beam), illuminated from the direction of the condenser (detector).

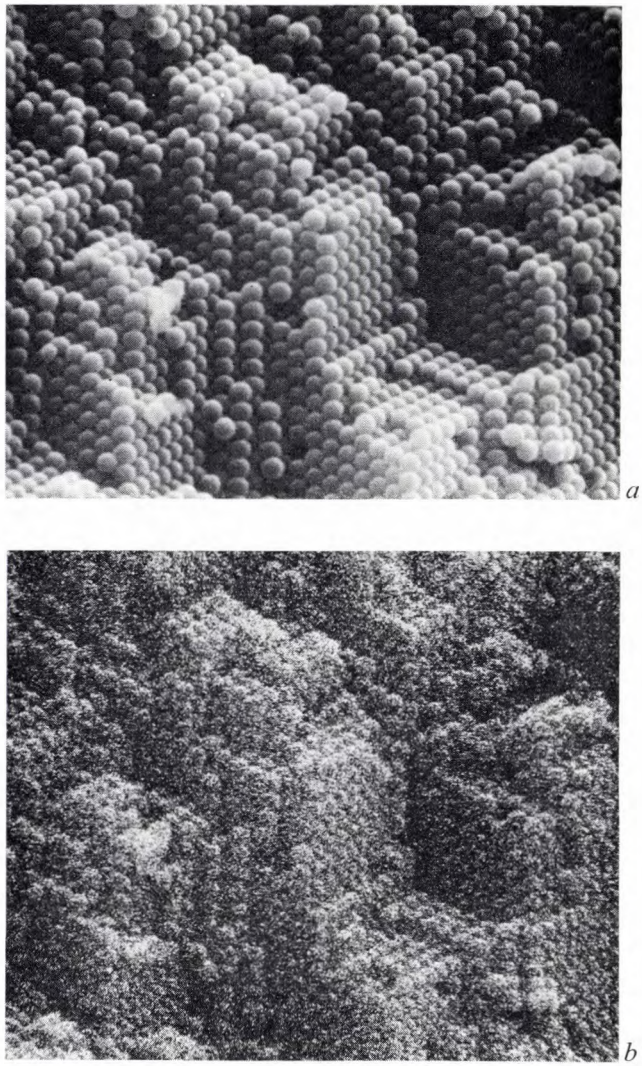


Fig. 8. Effect of the scanning rate on the image quality. *a*) Time for one line scan 125 ms. *b*) Time for one line scan 1 ms. The images were produced from a synthetic opal specimen, provided by Mr. B. Dawson, Monash University, Clayton, Melbourne, Victoria, Australia. Magnification 9000×.

in signal-to-noise ratio that can be obtained in this way.

At this point we can summarize by saying that a scanning electron microscope can give magnifications that may lie in the range from a few times to 100 000, that the depth of field is large and that only the simplest preparation techniques are required — and often none at all. Finally, there is another advantage that has not so far been mentioned: the image information is obtained as a time-variant electrical signal. This permits the image quality to be improved by signal processing. The image signal can also be stored, e.g. on magnetic tape, and the signal is then accessible for automatic data processing.

After this general introduction we shall now describe the PSEM 500, and we shall then look more closely at two features that have received particular attention during the development. These are the goniometer and the various circuits and accessories provided to increase both the ease of operation and the general applicability of the instrument.

Description of the instrument

The PSEM 500, whose control panel is shown in *fig. 9*, is a very compact instrument (*fig. 10*). At all stages of the design we made every effort to ensure the greatest ease of operation and the widest range of ap-

[3] P. S. Ong, 5th Int. Congress on X-ray optics and microanalysis, Tübingen 1968, p. 329.
A. V. Crewe, Proc. 5th Eur. Congress on electron microscopy, Manchester 1972, p. 640.

Fig. 9. The control panel of the PSEM 500. The standard range of control units has been extended by the addition of a number of optional units. The units at the extreme right of the panel are (from top to bottom) the control units for solid-state detectors (as used for the detection of back-scattered electrons), for the secondary-electron detector, which is part of the standard equipment, and for the detector for absorbed electrons (specimen current). A control unit for a cathodoluminescence detector and the circuit for electronic contrast enhancement are included as separate units on top of the panel, at the left. Finally, an extra cathode-ray tube has been included, used here to display a picture with 'y modulation'. In this kind of presentation the detector signal gives a vertical deflection of the electron beam producing the picture, rather than an intensity modulation.

plications. The microscope as shown in fig. 9 includes a number of optical control panels for special applications; these units are available separately. Automation and special safeguards ensure that the number of operations required in use is as small as possible and that it is impossible to carry out incorrect operations that might have disastrous consequences. This means that the technical operation of the equipment takes up very little of the microscopist's attention and that very little technical training is required. Various numerical data for the PSEM 500 are summarized in *Table I*.

The specimen is introduced into the specimen chamber via the airlock in a predetermined sequence of operations; the goniometer takes up the correct position automatically and the airlock doors cannot be opened in the wrong order.

A point that requires special attention in the design of an electron microscope is the contamination of the

Table I. Some numerical data relating to performance and characteristics of the PSEM 500

Magnification	10-80 000 ×
Resolving power	< 10 nm
Electron energy	1.5-50 keV
Beam current	10^{-12}-10^{-6} A
Electrically operated eucentric goniometer	
X- and Y-translation	25 mm
Fine z-movement	4.6 mm
Rotation	370°
Tilt	−7 to +46°
Max. specimen dimensions	25 mm thick 45 mm dia.
Manual goniometer	
X-translation	80 mm
Y-translation	60 mm
Tilt	−30° to +70°
Max. specimen dimensions (with full X and Y)	146 × 154 × 84 mm
Amplitude of scan	max. 12 × 9 mm min. 1.5 × 1.1 μm
Time for 1 line scan	0.001-1 s
Number of lines per picture	125-1000

specimen and instrument by hydrocarbons present in the residual gas. All the surfaces gradually become coated with a film of these hydrocarbons, which are then 'cracked' by electron bombardment and converted into a film of carbon. This gives a local reduction in the secondary emission from the specimen surface and a corresponding black raster in the secondary-electron image. The diaphragm of the objective lens can also be affected by the carbon deposit, which may collect in such a way that the aperture is then no longer perfectly circular and the astigmatism increases to a level at which it cannot be corrected.

The residual gas in the PSEM 500 is almost completely free of hydrocarbons: the partial pressure of hydrocarbons is less than 10^{-9} Pa (10^{-11} torr). This low level has been achieved by using only high-grade plastics, and even then only sparingly. The number of seals has also been kept to a minimum and sliding or rotating components that have to be sealed in the wall of the system are avoided where possible. Because of these precautions, specimen contamination is negligible The demountable inner lining of the electron-optical column never has to be cleaned more than once in six months, and usually much less frequently.

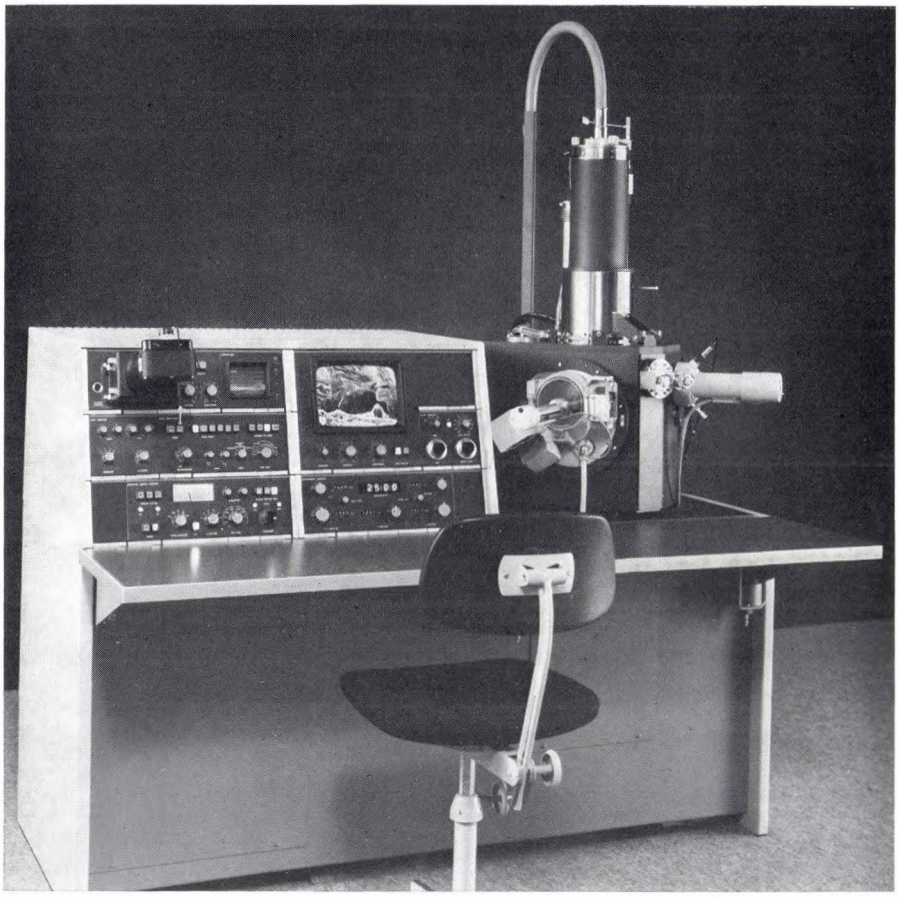


Fig. 10. The standard version of the PSEM 500. The electron-optical column is on the right, the various control units on the left. The upper left-hand part of the control panel carries a cathode-ray tube and camera for recording images, controls for brightness and contrast, and an oscillograph screen for checking the settings. Below this are the controls for the electron gun and for selecting the type of scan. The rest of the electron-optics control units are located at the lower left: high-voltage, beam diameter and focus. On the upper right-hand part of the panel there is a viewing screen, and the amplifiers for the various detectors. The controls for setting and reading the goniometer position are at the lower right, and the five numerical indicator tubes show the magnification. Below the optical column, from left to right: the specimen airlock, the selector control and the alignment control for the final-lens diaphragm and the secondary-electron detector.

A number of detectors can be connected simultaneously to the various ports of the specimen chamber (*fig. 11*). Reflected and transmitted electrons are detected by solid-state detectors placed above and below the specimen. The detector for transmitted electrons is made in two parts: a disc-shaped detector with an annular detector around it. With this arrangement images of the specimen seen in transmitted 'light' can be ob-

tained, for either bright-field or dark-field illumination. The amplifier associated with these detectors can give either negative or positive images (black-white reversal).

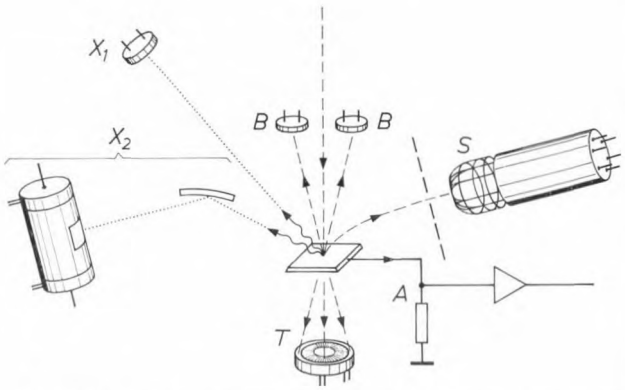


Fig. 11. The location of the various detectors around a specimen. *B* detectors for back-scattered electrons. *S* scintillator crystal and photomultiplier tube for detecting secondary electrons. *A* circuit for measuring the specimen current (absorbed electrons). *T* detectors for transmitted electrons; the central disc detects the directly transmitted electrons, giving a 'bright-field' image, the annulus detects electrons scattered at details in the specimen, giving 'dark-field illumination'. X_2 crystal spectrometer for detecting the long-wave X-radiation from light elements in the specimen. X_1 energy-dispersive solid-state detector for X-radiation from the other elements in the specimen. If required, the detector *S* fitted to the flange in the specimen-chamber wall can be replaced by a photomultiplier tube, for the detection of cathodoluminescence.

The secondary electrons are detected by a scintillator and a photomultiplier. A specially doped yttrium silicate is used here as the scintillator material. This material, known in the literature as P-47, has a much better sensitivity and longer life than the conventional plastic scintillators [4]. The surface of the scintillator is enclosed by a wire-mesh grid. A voltage applied between the grid and the specimen gives an energy selection of the secondary electrons; a voltage between the grid and the metallized scintillator surface ensures that the electrons strike the scintillator at a sufficiently high velocity.

The characteristic X-radiation emitted by the specimen when the electrons in the incident primary beam have a sufficiently high energy can be analysed by an energy-dispersive solid-state detector. This detector covers the greater part of the X-ray spectrum. For the lighter elements (such as C, N and O), a crystal spectrometer can be used to obtain a better resolution in this range.

The signals from the various detectors are processed in separate amplifiers and can readily be displayed in turn on the monitor. Auxiliary monitors can also be used for a good comparison of the various signals.

Fig. 12 gives an impression of the different images that can be obtained in this way from the same specimen. Signals directly suitable for computer input are

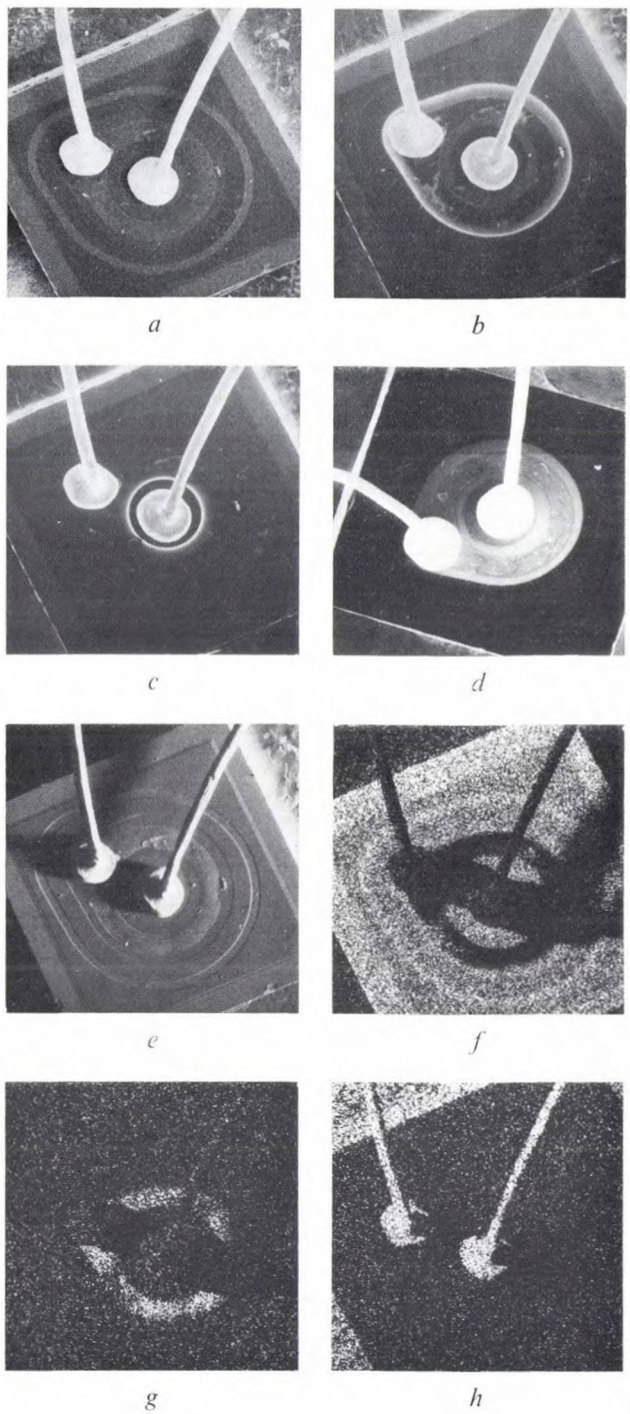


Fig. 12. A number of pictures of an *N-P-N* planar transistor (2N929), obtained with signals from different detectors. The emitter is at the centre, with the base around it; both have connector leads. The collector, formed by the remainder of the silicon chip, is electrically connected to the metal base-plate on which the transistor is mounted. *a*) The secondary-electron picture. *b*) Base and emitter earthed: the current induced by the electron beam in the base-collector junction has been used to produce an image (a secondary-electron picture has been superimposed to indicate relative positions). *c*) As the previous picture, but now collector and base are earthed and the current in the emitter-base junction has been used to produce an image. *d*) Potential contrast. Emitter and base are earthed and the collector is at a potential of 4 V with respect to earth. *e*) Picture obtained with back-scattered electrons. *f*) Picture obtained with the characteristic X-radiation from silicon. *g*) As (*f*) but for aluminium. *h*) As (*f*) but for gold. Magnification 40×.

also available from the various amplifiers, to facilitate automatic data processing.

The PSEM 500 is provided with two monitors, one of which has a long-persistence screen for easy visual observation. The other monitor, for photographic recording, has a short-persistence screen with a high resolution.

A horizontal dashed line can be written in the image as a measurement aid. The lengths of the dashes and the spaces between them are 100, 10 or 1 μm, depending on the magnification selected. The position of the first dash in relation to the others indicates the length of the dashes in any particular case (*fig. 13*). This method of coding has the advantage that it is not affected by enlargement or reduction of photographs produced from the image.

In selecting the raster to be used for scanning a specimen, it is not only necessary to choose the scanning amplitudes (which determine the magnification), it is also necessary to select the line time and the number of lines per picture. If another line time is selected, the brightness on the photo-monitor screen is altered automatically to keep the photographic conditions the same.

Two aids are available for setting to the optimum photographic conditions. A built-in generator provides a scale of shades of grey, either varying linearly from black to white, or in ten steps. This can be used to find the best brightness and contrast settings for the photographic material being used. The second aid is an oscilloscope screen that can display the video signal. It is used in setting the gain and black level of the detector in such a way that the limits of the output signal exactly match the limits in the grey scale. For reference the black and white levels are indicated by two horizontal lines on the screen (*fig. 14*). This oscilloscope screen will also indicate video-chain overloading.

Adjustment of the electron optics (focusing and astigmatism correction) can be carried out by switching in a raster that only occupies 1/16 of the normal area. The speed of the scanning beam at the surface and hence the signal-to-noise ratio remain unchanged. The reduced raster is scanned in 1/16 of the time and at a correspondingly higher repetition rate. This means that variations in the setting of the electron optics can be followed closely, even at low line-scan rates.

Other kinds of special scan that can be used are: scanning along an arbitrary horizontal or vertical line, or just at a particular point. Such scans are used primarily for an elemental analysis of the specimen from its X-radiation. The positions of the point and

[4] J. B. Pawley, in O. Johari and I. Corvin (ed.), Scanning electron microscopy/1974, ITT Res. Inst., Chicago, p. 27.

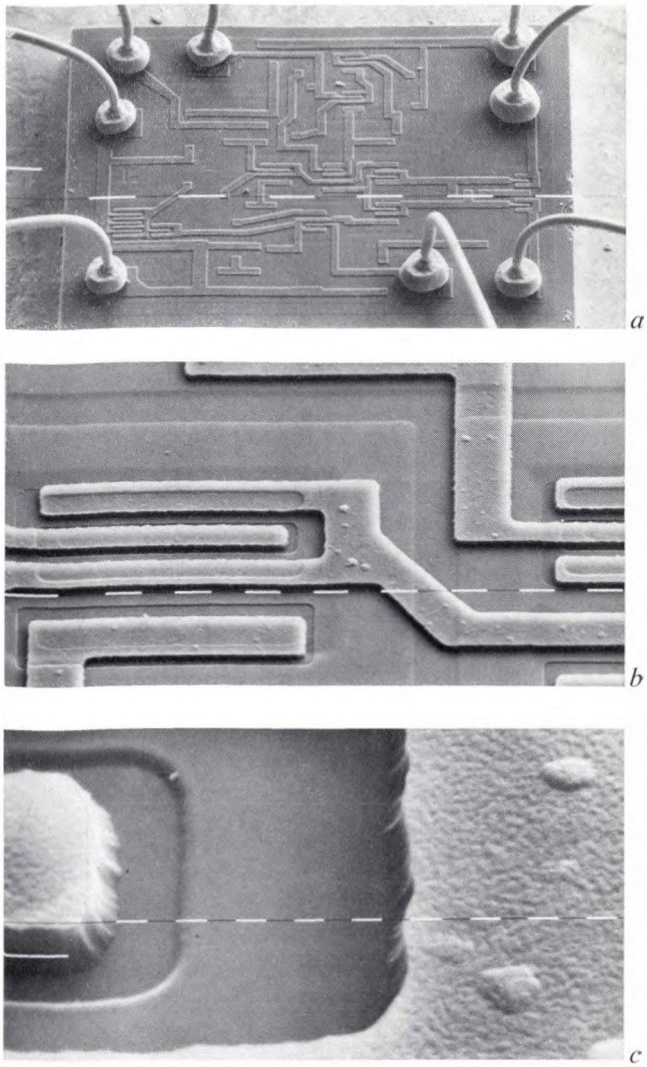


Fig. 13. Dashed line in the image to indicate the magnification. The relative position of the first dash indicates the length of the dashes. *a*) First dash above the line: dashes and spaces 100 μm. *b*) First dash on the line: dashes and spaces 10 μm. *c*) First dash below the line: dashes and spaces 1 μm.

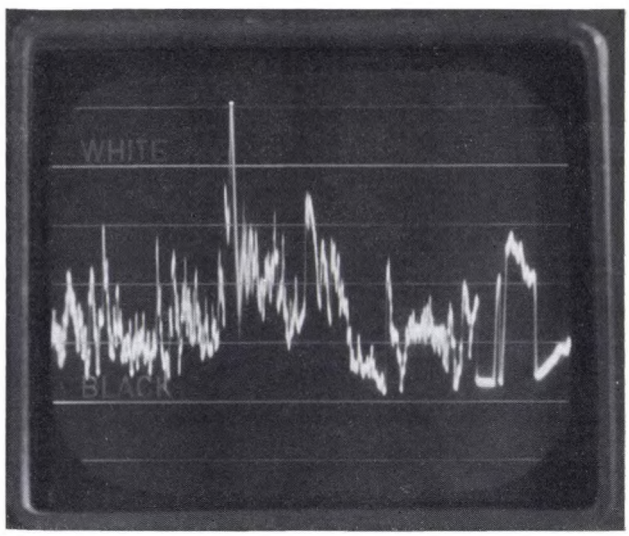


Fig. 14. The oscilloscope screen displaying the video signal as an aid in setting the contrast range on the photo screen. Black and white levels are indicated by the horizontal lines.

the two lines can be chosen anywhere in the image field. A cross can be recorded with the image to indicate the location where X-ray analysis was made in the spot mode (*fig. 15*).

Finally, it is also possible to scan the specimen surface at television rates; appropriate synchronizing signals are then added to the video signals. The image information can then be recorded on a video recorder and reproduced on a television monitor.

The goniometer

The goniometer of a scanning electron microscope should permit the greatest possible free movement of the specimen in all directions. This can be achieved from three translation movements in combination with two rotations. For the PSEM 500 we have designed a 'eucentric' goniometer — i.e. a goniometer in which the point of intersection of the two axes of rotation, the 'eucentric point', always remains on the optical axis of the instrument. If the surface of the specimen coincides with this point the centre of the image will always portray the same point of the specimen surface whatever tilt or rotation is applied.

The sequence of the various rotation and translation movements is as follows (*fig. 16*). Starting from the specimen stage we first have two translations parallel to its surface. These are the *X*- and *Y-movements*. Next there is a *rotation* about an axis perpendicular to the *XY*-plane, followed by a translation in the direction of the rotation axis, the *fine z-movement*. The complete mechanism so far described is supported on a *tilt movement* that rotates about a horizontal axis passing through the intersection point of the rotation axis and the optical axis. Finally, there is the *coarse Z-movement*, a translation in the direction of the optical axis. The fine *z*-movement allows the specimen surface to be brought to the best possible coincidence with the intersection point of the tilting axis, rotation axis and optical axis of the microscope.

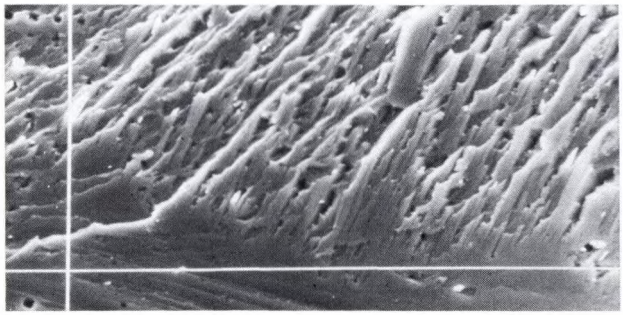


Fig. 15. Scanning-microscope picture of a tungsten cleavage plane, on which a cross has been superimposed to indicate the locations where the electron-microprobe analysis was made. Magnification 4000×.

All the goniometer movements except the coarse *Z*-movement are driven by stepping motors, which are mounted in such a way that they can drive the various goniometer mechanisms with the minimum of gearing. With this arrangement there is hardly any backlash in the drive systems, and no external forces are applied to the goniometer. The moving parts cannot be lubricated with oil or other organic substances, since these would contaminate the vacuum. To avoid cold welding at metal-to-metal contacts, and to ensure smooth operation of the various rotating and sliding movements, optimum combinations of different materials are used where parts are in sliding contact. In the very few cases where such combinations could not be used, the surfaces are lightly coated with 'dry' lubricants [5].

The use of stepping motors for the goniometer drives facilitates digital control and read-out of the goniometer position. This is important in measurement

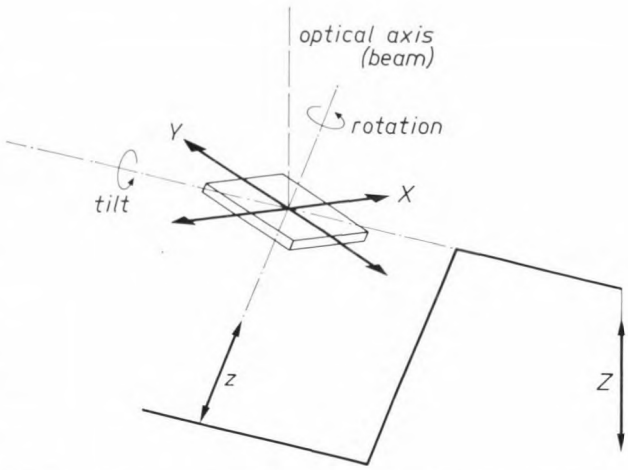


Fig. 16. Diagram indicating the combination of the various goniometer movements. The *X*- and *Y*-translation mechanisms must be mounted on the rotation and tilt movement, since the rotation and tilt axes must always intersect. There would be no essential difficulty in exchanging the fine *z*-movement, which operates along the rotation axis, with the rotation movement. If the rotation movement is required to operate about an axis perpendicular to the specimen surface, this movement must be mounted on the tilt movement.

work, particularly when automated routine measurements are made in conjunction with a directly connected computer.

Fig. 17 is a photograph of the precision goniometer described here. The maximum specimen height it will accept for eucentric motion is 25 mm; the maximum diameter should be less than 45 mm. A manually operated goniometer that will take much larger specimens (see Table I) is also available. The manual goniometer can also be used for investigations made inside the microscope — such as tensile tests — where some of the apparatus has to be included in the specimen chamber.

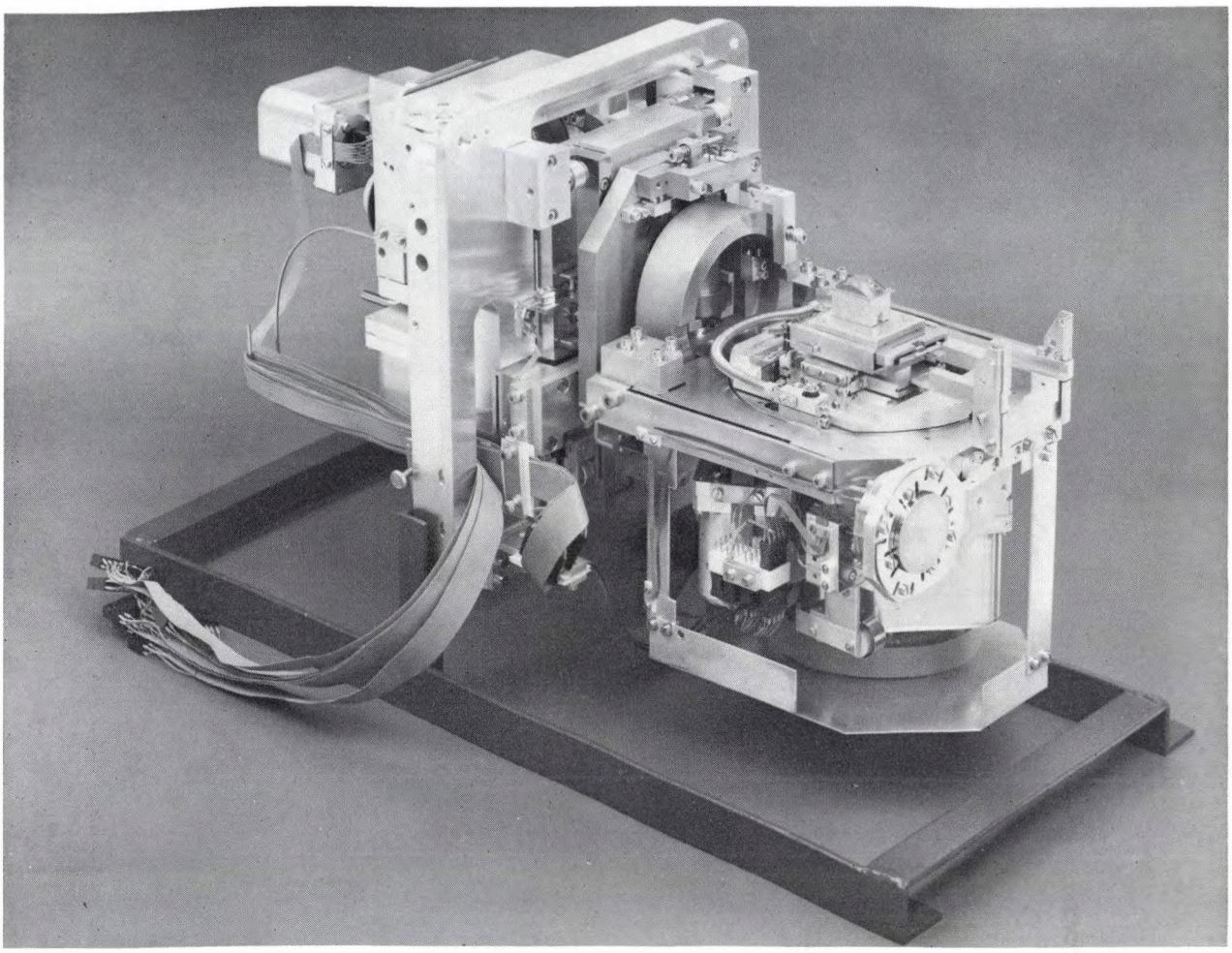


Fig. 17. The eucentric goniometer of the PSEM 500. The motor for driving the tilt movement can be seen at the upper left. The motors for the other rotation and translation movements are mounted in the lower right-hand part of the instrument. On top of this right-hand part is the specimen-holder with the specimen, in this case an integrated circuit. The supply leads for the various motors, and those for the specimen stage, are made in the form of flexible strip.

Special circuits and accessories

As we said at the beginning, the electron-optical system consists of an electron source and three electromagnetic lenses that produce an image of this source on the specimen surface. The control that sets the current in the first two lenses is calibrated in terms of 'spot size' at the specimen. Coupling between the circuit and the high-voltage supply that determines the electron energy ensures that the spot-size setting is independent of the electron energy.

The focus (i.e. the current in the coil of the third lens) is adjusted by means of a single knob, with three push-buttons that can be used to select the steps in focal-spot size for a particular rotation of the knob. In normal use this focusing knob will be rotated many times. Therefore, to ensure a long life, the knob is arranged as a pulse source, with the aid of a light-emitting diode and a light-sensitive diode. These control a digital circuit that sets the current in the final lens via a digital-to-analog converter and a control amplifier (see *fig. 18*).

The focus control includes a reset circuit, which will return the setting to a particular position. The reset setting corresponds to the eucentric point for a spot of 125 nm and an accelerating voltage of 25 kV. The lens current is always switched in such a way that the set point is approached in the same direction, so that hysteresis in the pole pieces always produces the same effect. In this way the current in the lens coil and also the corresponding magnetic field-strength both arrive at the same final values on reset.

The filament forming the cathode of the electron gun has a limited life because of the high temperature necessary to produce a sufficiently high beam-current density. This means that the filament will have to be replaced

[5] The method used is described in J. Crucq, Philips tech. Rev. **34**, 106, 1974 (No. 4).

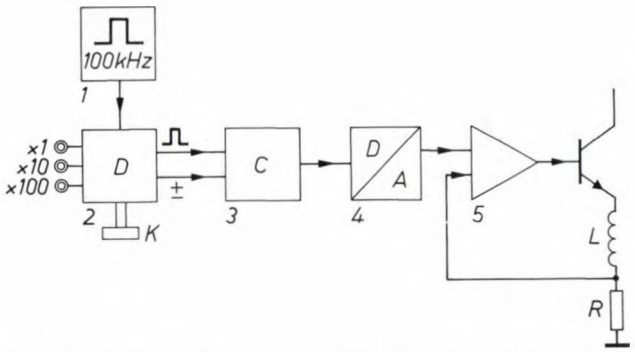


Fig. 18. Block diagram of the circuit for focusing with one knob and three push-buttons. The 100-kHz pulse generator *1* supplies the clock signals for the digital circuit. The decoder *2* transmits a number of clock pulses corresponding to the rotation of the knob *K* and to the size of step selected with the push-buttons. The counter *3*, which counts the number of pulses transmitted by the decoder, can count both forwards and backwards, depending on the information provided by the decoder about the direction of rotation of the knob *K*. Finally, the state of the counter controls the current in the objective-lens coil *L*, via the digital-to-analog converter *4* and the control amplifier *5*. A resistance *R* is included in series with the coil to provide feedback from the coil current to the control amplifier.

occasionally and the gun readjusted. Accurate realignment of the gun, after changing the filament, is not however necessary if instead the beam from the gun is deflected so that it travels along the optical axis. A built-in electromagnetic deflection has been included for this purpose. This system consists of two sets of coils, one compensating for a tilt of the electron beam, the other for a sideways displacement (*fig. 19*). For accurate alignment the two corrections are applied successively to obtain a maximum of the specimen current.

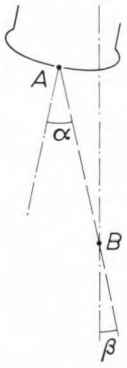


Fig. 19. Principle of the deflection system for aligning the electron beam. One coil set has to give a rotation of the beam through an angle α about the point *A*, the other then has to rotate the beam through an angle β about the point *B* to make the beam coincide with the axis of the electron-optical column.

The raster pattern on the specimen surface is obtained by applying a sawtooth voltage to the beam-deflection coils. A d.c. bias current can also be passed through these coils, to give an electrically produced displacement of the image field over a small distance. This displacement cannot be too large, since it is in fact obtained by deflection of the optical axis of the system so that this axis no longer passes through the eucentric point and the eucentric operation of the goniometer is impaired.

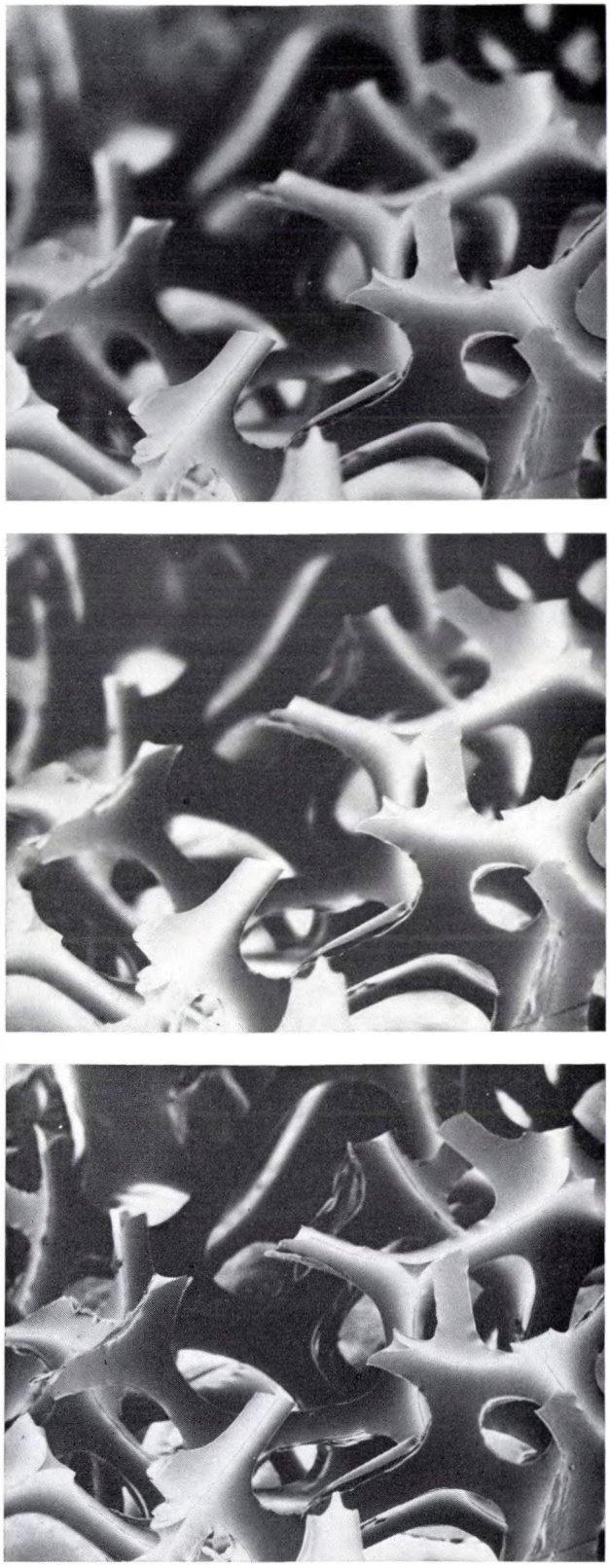


Fig. 20. Demonstration of the effect of the aperture angle of the scanning electron beam on the depth of field. As in fig. 1, the object is a plastic foam coated with a gold film. The three pictures were made with a distance of 23 mm between the diaphragm of the objective and the specimen. The aperture diameters were, downwards, 400 μm, 200 μm and 100 μm. An aperture of 100 μm was also used in making the photograph for fig. 1, but the distance between diaphragm and specimen was 43.5 mm. The depth of field was also correspondingly greater. Magnification 75×.

The amplitude of the 'raster deflection' determines the magnification of the microscope, given the specimen distance. The value of the magnification is indicated, for both the upper and lower positions of the coarse Z-control of the goniometer, on a set of numerical-indicator tubes on the control panel.

The aperture of the objective lens can easily be changed. Any of the three available final-lens diaphragms can be aligned with the optical axis of the column via a bellows attachment in the wall of the evacuated enclosure. The effect of this choice on the depth of field is shown in *fig. 20*. If a small diaphragm is chosen, to obtain a large depth of field, the beam current becomes smaller, and the scanning rate therefore has to be made correspondingly slower to prevent a degradation of the signal-to-noise ratio.

As mentioned earlier, electronic processing of the detector signals can be used to improve the quality of the image. An example of this is electronic contrast enhancement. *Fig. 21* gives an impression of the improvement that this will give. The contrast enhancement is obtained by separately amplifying the deviations of the detector signal with respect to a continuously averaged signal, and then adding them to the original signal. Low contrast can be improved in this way and large differences in brightness between the different parts of the image can be reduced.

It is also possible to modify the sawtooth deflection currents as well as the detector signals. This can for example be done to correct the distortion of the rectangular raster that occurs when the specimen surface is not perpendicular to the optical axis of the microscope.

Many options are open to the user. Those relating to the detectors and the signal processing have already been described, but there are others, as yet unmentioned, more concerned with the treatment of the specimen. For example, the specimen can be heated or cooled during the observations. Many voltages and currents can also be applied to the specimen; this is a particularly valuable facility in studying semiconductor devices and integrated circuits. Much of the equipment required for this can be supplied as separate units, which means that the PSEM 500 can readily be adapted to the individual needs of a wide range of users.

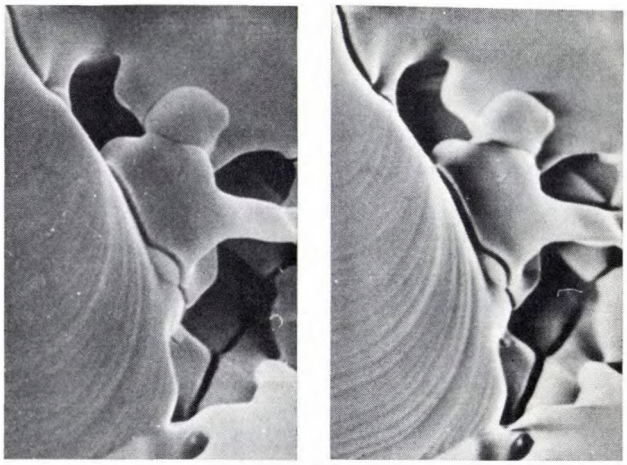


Fig. 21. Electronic contrast enhancement. Two pictures of a ZnS surface, with contrast enhancement of the right-hand picture. Magnification 1300×.

Summary. In a scanning electron microscope the specimen surface is scanned in a raster pattern by a fine electron beam. One of the responses of the specimen to the incident beam is detected, and an image is produced by displaying the detector signal on a cathode-ray tube scanned in synchronism with the microscope beam. An image can be obtained from reflected, secondary, absorbed or transmitted electrons, or from cathodoluminescence or characteristic X-radiation, as desired. Contrast in the image can be obtained from surface geometry, variation in crystal orientation, differences in elemental composition, or from potential differences.

In designing the PSEM 500 particular attention was paid to ease of operation, and provision was made for automatic operation and for processing the measured data. One such facility is the precision digitally controlled goniometer. Operation of the electron optics and the vacuum equipment has been automated as far as possible, and operational safeguards have been included where necessary. The detector signals are processed and presented on the monitor screen in such a way as to give reproducible and trouble-free photographic recording of results.

An experimental endoscope with miniature television camera

J. Périlhou

Very fast synchronous motors with a diameter of only a few millimetres have been developed at L.E.P., Limeil-Brévannes. These motors have been used in a mechanically scanning miniature television camera. The simplicity and high resolution make the camera an excellent head piece for an experimental endoscope that outperforms the conventional types. A significant advantage which present endoscopes do not have is that the image appears in colour on a television monitor.

Endoscopy: internal viewing

An endoscope is an instrument used mainly in medicine, but also in technical applications, for viewing the interior surface of small cavities. It consists of a fairly flexible tube with built-in optics for illumination and imaging. In older types the optical system consists of a series of tiny lenses — there may be as many as forty or fifty — one behind the other; in the newer types, however, the image is transferred by a bundle of thousands of extremely thin optical fibres. The use of a bundle of fibres instead of lenses makes the tube more flexible, but a drawback is that the resolution of the image is not then as good. The endoscope is pushed into the body, possibly turning corners on the way, until the distal end (the far end) reaches the place to be examined. The length of the tube is about one metre and the diameter 8 to 10 mm.

In medicine endoscopy is a valuable method for internal investigation of the patient. It can be used in examining internal organs such as the larynx and gullet, the intestines, and in certain gynaecological examinations. Endoscopy also has its uses in technology and science, where it is particularly useful in investigations, maintenance and repairs in industry. A typical example is the *in situ* inspection of turbine blades in certain types of aircraft engine. The endoscope could also be a useful tool in many processing techniques, for monitoring purposes and even for process control. In most applications the introduction or presence of an endoscope does not significantly affect an existing situation. Endoscopy is often the means of obtaining a maximum of internal information with a minimum of discomfort and trouble.

The new endoscope (*fig. 1*) contains an extremely small television camera in the distal end of the flexible tube. The camera is 25 mm long and its diameter is less than 5 mm. The actual camera is therefore located at the place to be examined; it views the major part (200 °) of a strip of the cavity wall situated around the end of the tube. The image transmitted back appears on a colour-television monitor. Images can thus be viewed by several people at the same time, which is very useful in instructional applications. Existing endoscopes do not have this feature.

The flexible tube contains only three optical fibres. Two of these supply light to the camera and provide the local illumination. The returning light is transmitted by the third fibre to three photomultiplier tubes in the camera-control unit. Each photomultiplier tube has a filter — one green, one red and one blue. The control unit processes the three video signals to produce an approximately rectangular colour display on the monitor screen (*fig. 2*).

Even with the first experimental camera the resolution is more than 160 lines over the complete field of view, i.e. 80 white lines alternated with 80 black ones. With this resolution objects no bigger than about 0.1 mm can be distinguished on the monitor screen, provided they are no further than 1 cm from the centre-line of the camera.

The new endoscope has been used for experimental studies of blood-vessel walls in a live dog. The tube was inserted through a front-paw vein until the camera almost reached the auricle of the heart. The camera was contained in an extremely thin and transparent rubber sheath. This sheath was inflated to press against the blood-vessel wall and force the blood away from the area under examination, giving a good view of the blood-vessel wall. The inspection proceeded successfully almost to the auricle, where the heart beats made further observation difficult.

In a later phase the endoscope has been used in certain neurosurgical operations, for example during brain

J. Périlhou is with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

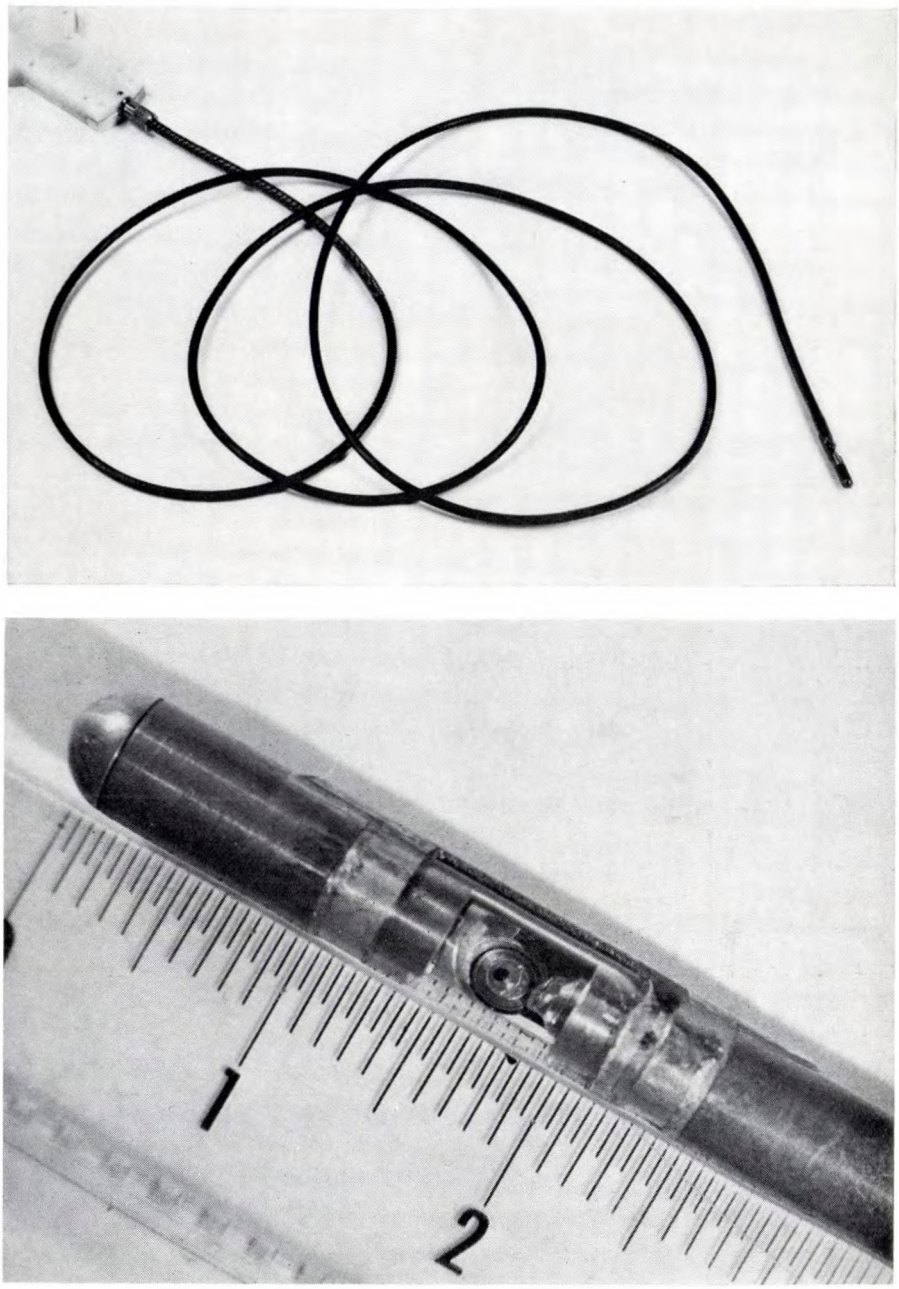


Fig. 1. The endoscope. It consists of a flexible tube (length 2 m, diameter 4 mm) with an extremely small television camera mounted at the end. A rotating optical system in the camera scans a cylindrical observation surface around the longitudinal axis. The multiway connector at the other end of the tube connects the camera to a colour-television control unit and monitor. The enlarged view of the actual camera shows the cylindrical window and some of the components of the optical system.

operations to monitor the brain cavities or 'ventricles' (fig. 2).

In examinations of pregnant women the endoscope has provided what is believed to be the first *direct* observation of cases in which the umbilical cord was wound around the neck of the foetus.

These particular cases [1] are only a small sample of the many in which the new endoscope could prove of value to the medical practitioner.

The prospects of the instrument in industry seem even better, not only because the range of applications is so much wider but also because increasing familiarity with the instrument will undoubtedly lead to new applications.

In the next section the construction and operation of the miniature camera will be described. Details of the mechanical, electronic and optical components of the instrument are given in the final section.

Construction and operation of the camera

To make a television camera that is no bigger than a few millimetres it seems impossible to proceed from the principle used in modern studio cameras. We there-

[1] These applications of the endoscope were made as part of the work at three medical centres in Paris: the Marie Lannelong Hospital (Prof. Binet), the Salpétrière Hospital (Prof. Lebeau) and the Marignan Clinic (Dr Henry).

fore chose a mechanical scanning system, but one that differed completely from the Nipkow system [2]. This was made possible by the use of synchronous motors with a diameter of only 3.5 mm, developed in our laboratories.

Fig. 3 shows a cross-section of the miniature camera, and illustrates schematically the way in which it is connected to the control unit and the monitor.

gradually displaced. The camera thus scans the entire surrounding wall; the field of view in the longitudinal axis of the tube is approximately 55°. Each complete rotation of *PL* therefore corresponds to a complete scan, which will take about 25 ms. The permissible distance between the wall surface and the centre of the camera is between 10 and 30 mm. At distances greater than 30 mm the signal-to-noise ratio in the image is

Fig. 2. Example of an image on the monitor screen during a brain operation. The camera was inserted into one of the ventricles (brain cavities). The approximately rectangular image is a plane projection of the cylindrical field of view.

The control unit contains the electronic circuits required for camera and monitor, such as power-supply circuits, video amplifiers, and also the power supply for the motor, the three photomultiplier tubes, and a light source with a condenser lens. This lens transmits an image of the light source — a high-pressure mercury-vapour lamp — to the two fibres F_{III}, which transmit the light to the immediate surroundings of the camera.

The camera consists essentially of a piece of transparent plastic tube *L*, with a synchronous motor at each end. Fixed to the rotor of M_1 is a mirror *R*, inclined at 45°, which rotates at a speed of 3037.5 revolutions per second. Fixed to the rotor of M_2 is a prism-lens system *PL*, which rotates at 37.5 revolutions per second, i.e. about 80 times slower. For every combination of the positions of *R* and *PL* the end of the read fibre F_r receives light from a small element of the wall *W* of the cavity under examination. If *PL* were fixed in position the rotation of *R* would cause this element to describe a trochoid-like line on the wall — because the front surface of *PL* is oblique to the axis. Owing to the rotation of *PL*, however, the position of this trochoid is unacceptable, since the illumination is then too weak.

The image on the monitor is produced electronically in the conventional way from two interlaced fields, one with the 'odd' lines, one with the 'even' ones. The interlacing reduces flicker, as in television.

The resolution has been determined by means of line patterns applied to the inside surface of a tube that can be fitted around the end of the endoscope.

Motors, electronics and optical system

The motors

In motor M_2 (fig. 3) the shaft is stationary. The rotor turns in two jewelled bearings. The great advantage of a stationary shaft is that it enables the read fibre to be firmly fixed. Attached to the rotor is a piece of heavy metal (gold); this keeps the motor running very smoothly, so that a good steady picture appears on the

[2] See for example B. Grob, Basic television — principles and servicing, McGraw-Hill, New York 1949, chapter 2, or W. A. Holm, How television works, Philips Technical Library, Eindhoven 1963, p. 21.

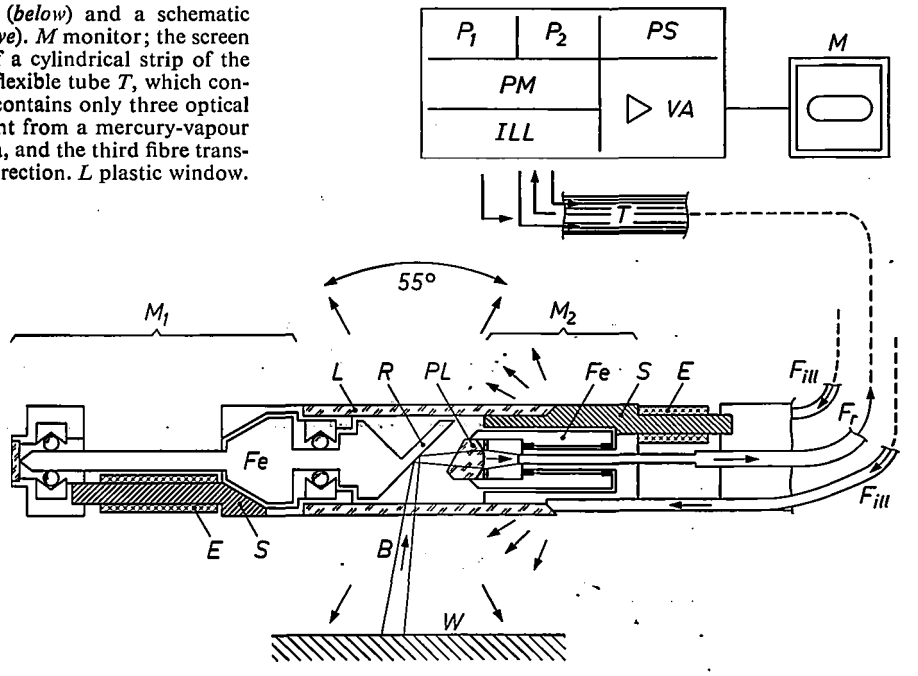


Fig. 3. Cross-section of the camera (*below*) and a schematic representation of the control unit (*above*). *M* monitor; the screen shows a rectangular colour picture of a cylindrical strip of the wall *W* surrounding the camera. The flexible tube *T*, which connects the camera to the control unit, contains only three optical fibres. Two of them, F_{ill}, transmit light from a mercury-vapour lamp *ILL* to the location of the camera, and the third fibre transmits light from the wall in the other direction. *L* plastic window. *R* fast-rotating mirror (3037.5 rev/s). *PL* slowly rotating prism-lens system (37.5 rev/s). In each position of *R* and *PL* the end of the read fibre F_r can only be reached by light originating from a small element of the wall (beam *B*). $M_{1,2}$ synchronous motors (length about 10 mm, diameter 3.5 mm) with ferroxdure rotor *Fe*, stator *S* and windings *E*. The control unit contains the power-supply circuits $P_{1,2}$ for the two motors, three photomultiplier tubes *PM* with colour filters and high-voltage source, and power-supply circuits *PS*. *VA* video amplifiers. The aperture angle 55° of the field of view corresponds to the height of the image on the screen.

monitor. The motor dissipates 70 mW, which is easily removed by natural convection.

The other motor (M_1 in fig. 3), which runs much faster, was not so easy to develop. The rotor shaft is solid; it runs in two miniature oil-lubricated ball bearings and has a jewelled bearing at one end. At the other end sufficient axial thrust is provided magnetically by the oblique air gap between the rotor and the stator. The mirror is cemented to the rotor, and consists of a piece of glass 0.5 mm thick; in shape it is a perfect ellipse. The motor runs at a very uniform speed and can be operated continuously for hundreds of hours. Very little heat is generated (60 mW) and cooling presents no problems. The motor delivers a power of 0.8 W from a two-phase a.c. supply at 3037.5 Hz.

The electronic circuits

The a.c. signal for the fast motor is generated by an electronic clock (966 kHz) connected to a divider circuit. The scaling factor, 2×159, is obtained in the usual way with a number of decade scalers and a binary scaler; the binary scaler also gives the signal the correct shape. The supply for the slow motor is obtained as a beat signal by combining the 3037.5-Hz signal with a signal whose frequency is 3000 Hz, obtained by division from the clock signal (scaling factor 2×161). The supply system adopted has the advantage of keeping the motors automatically synchronized.

The a.c. currents that drive the motors are also used for energizing the deflection coils of the television tube in the monitor, in combination with a sawtooth voltage, which functions in the usual way as a trigger signal. In this way fixed picture synchronization is ensured.

The optical system

The light source (100 W) has an area of about 0.1 mm² and gives a luminance of about 1.9×10^9 cd/m². Its spectrum is such that the light can be considered as white. The light-transmission fibres (F_{ill} in fig. 3) consist of a plastic material ('Crofon', DuPont), and are therefore easy to mount. The intensity in such a fibre decreases by a factor of two per metre. The read fibre (F_r in fig. 3) is made of glass, and the loss of intensity is therefore smaller: a factor of two per *two* metres. The lens of the combination *PL* (fig. 3) has a focal length of 1.5 mm; a diaphragm with an aperture of 0.7 mm ensures that the image spot produced on the read fibre is sufficiently sharp.

The use of better optical fibres, which are already becoming available, will no doubt make it possible to make tubes several tens of metres long. The applications of the endoscope could then easily assume the nature of closed-circuit television.

Summary. A type of endoscope is described that picks up internal images for observation in colour on a television monitor. The inserted end contains a television camera (diameter 4.5 mm, length 25 mm); the flexible tube (length 2 m, diameter 4 mm) contains two fibres for illuminating the cylindrical field of observation and one 'read fibre' for transmitting the light to the camera-control unit. Scanning is performed sequentially and mechanically with the aid of an oblique rotating mirror (3037.5 rev/s) and a rotating combination prism lens (37.5 rev/s). The resolution is 80 line pairs. Two miniature electric motors (130 mW total) rotate the mirror and prism lens; the power is supplied by a clock circuit, which also supplies the signal for the deflection coils of the television tube, thus ensuring synchronization. The new endoscope has applications in medicine and in industry.

The automatic measurement of medical X-ray photographs

W. Spiesberger and M. Tasto

The medical profession would readily agree that the preventive sector of public health care does not always work in the most efficient way. The checks required are comparatively straightforward, but unfortunately the scale on which they should be carried out annually is not usually compatible with available resources in staff and time. In X-ray diagnostics, which plays an increasingly important part in preventive medicine, the introduction of the computer has resulted in the solution of some of the problems.

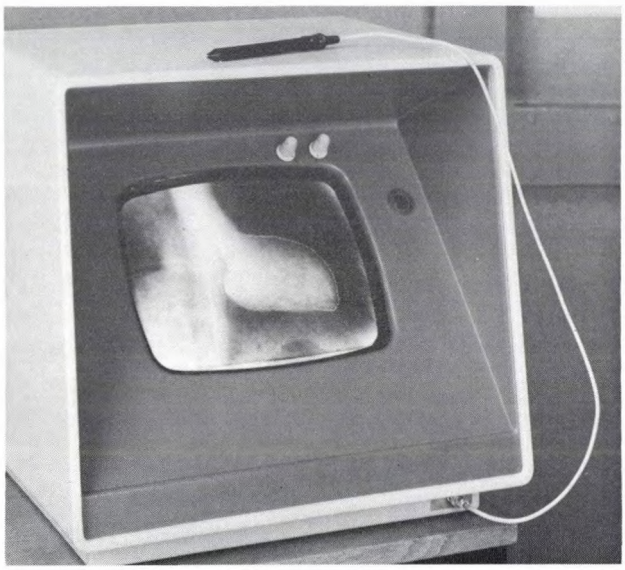
The quantitative element

Between 60 000 and 500 000 X-ray records are examined annually in a modern hospital, depending on its size. These figures indicate that doctors must spend very many working hours in the interpretation of all this information. Nowadays, moreover, the records are usually not only visually examined, but attempts are made to derive more information by measuring particular objects on the image [1]. The number of working hours required for the measurement and processing of such numerical results is even greater, which rather limits the popularity of the measurement procedures. This explains the need for a partially or completely automatic data-processing package, which could be used both for the measurement of the X-ray photographs and for the processing of the resultant data.

Some examples of interesting problems include the measurement of the cardiothoracic ratio, the heart volume and other cardiac quantities [1] [2]; the determination of the volumes of the heart ventricles as a function of time, from film or television pictures of the X-ray image — by the technique of cineangiography, in which the ventricles are filled with a contrast medium—; the determination of the thickness of the cardiac wall and the motion of the wall [3]; the measurement of the flow in blood vessels containing a contrast medium, using film or television recording methods [3]; the measurement of the shape and size of the carpus (wrist bones) in children in order to ascertain the extent of skeletal development [4]. In particular, the measurement of dynamic quantities by cine techniques is extremely time-consuming, because between fifty and one hundred separate images (frames) often have to be measured for each patient.

Investigation of all these kinds of measurement has demonstrated that the viewing and 'normal' analysis of the images — such as marking the contours of an object to be measured — are *not* the most labour-intensive

Dipl.-Phys. W. Spiesberger and Dr M. Tasto are with Philips Forschungslaboratorium Hamburg GmbH, Hamburg, West Germany.



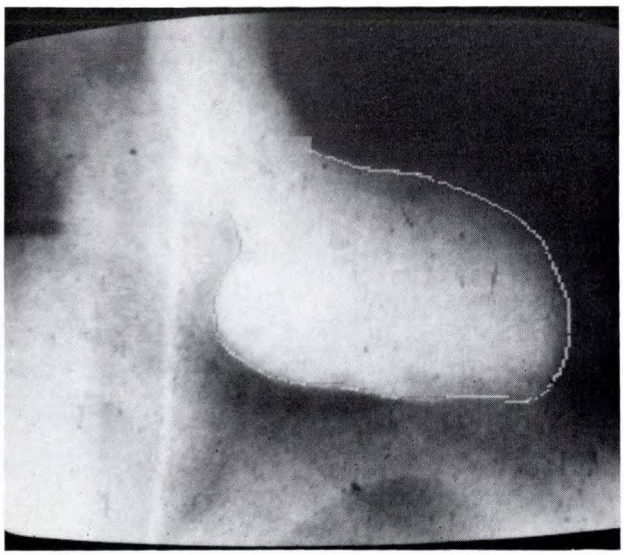


Fig. 1. Television monitor with light pen, constituting part of the equipment for semi-automatic image processing. The doctor only has to trace the outline of the object (a left cardiac ventricle in the case shown) on the image screen with the light pen. The resulting coordinates appear on the screen as illuminated dots (see lower photograph) and are processed directly by a computer.

steps and do not produce the largest number of errors. On the contrary, the computation of surfaces and volumes, the processing of various kinds of data and the accurate numerical weighting of measured values for the determination of diagnostic index values are examples of operations that tend to be the most time-consuming and most frequently contain errors. Various groups, including our own, have therefore developed a number of image-processing methods that depend on the use of computers, but leave to the doctor the task of locating and marking objects of interest. A typical example is the measurement of an object from a picture on a television monitor (*fig. 1*). The specialist traces the outline of the object with a light pen, and the coordinates of the contour traced are simultaneously and automatically fed into the computer, which then performs all further calculations — extremely accurately and without errors. Similar processing of results can often lead to a reduction of working time by a factor of ten or more.

A further step in the direction of automation, and therefore in the reduction in man-hours, is taken by utilizing the brightness information contained in the electronic video signal. If for example the selected object is uniformly brighter than its surroundings, it is easy to provide an electronic threshold circuit to ensure that only the parts in the image above threshold are shown. Quantities such as the area can then be measured directly [5]. A similar threshold can easily be adjusted by means of a single knob for the individual patient. *Fig. 2* demonstrates this method of operation.

The method is not applicable where the information comprising the image is rather more complex, and the chosen object can no longer be distinguished from its surroundings by the adjustment of a single knob, or perhaps two or three at the very most. A less straightforward procedure is then necessary, and the use of a computer with a larger memory is an obvious requirement. The program is arranged so that the brightness of the image, whose shape is irrelevant, is defined at a large number of points forming a square array. The 'brightness matrix' thus obtained is stored in the computer memory for further processing.

In the following sections two types of procedures are described that enable the X-ray photograph to be processed completely automatically for certain measurement applications. One type is used to determine the shape of the heart and to measure the cardiothoracic ratio, both of which are of primary importance in preventive medicine. The other procedure is used for the recognition of contours of cardiac ventricles, which are of value in a refined diagnosis of the cardiac function. In conclusion we shall look at the implementation on the computer, and at the prospects for the future.

Cardiothoracic measurements

The early detection of heart and vascular diseases has an important place in preventive medicine. In general X-ray photographs, electrocardiograms (ECG) and cardiophonograms are used [6] [7]. The serious nature of these diseases has led to an increase in the work required for preliminary investigations. Computer methods for the acquisition and processing of data obtained from the diagnostic procedures mentioned above are therefore of particular interest for mass screening of the general public.

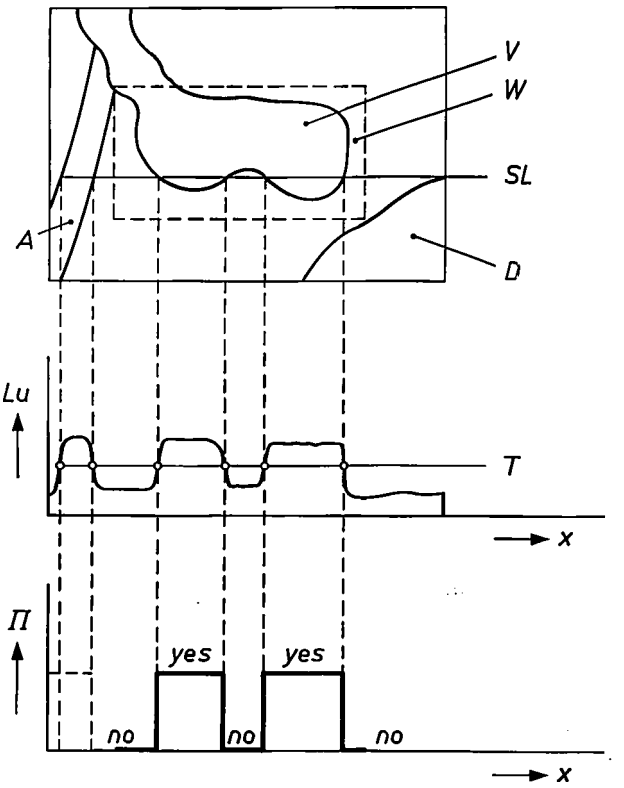


Fig. 2. Area measurement of a left ventricle (*V*) from a video signal (*Lu*). The threshold *T* and the window *W* can be adjusted manually. *SL* scan line. *D* diaphragm. *A* aorta. The logic signal *Π* indicates whether the object sought is present (yes) or not (no). *x* space coordinate. The signal from the aorta does not cause an error since it appears outside the window.

[1] L. B. Lusted and T. E. Keats, Atlas of röntgenographic measurement, Year Book Medical Publishers Inc., 1972.
[2] H. Reindell, Funktionsdiagnostik des gesunden und kranken Herzens, Thieme Verlag, Stuttgart 1967.
[3] P. H. Heintzen (ed.), Röntgen-, cine- and videodensitometry, Thieme Verlag, Stuttgart 1971.
[4] W. W. Greulich and S. I. Pyles, Radiographic atlas of skeletal development of the hand and wrist, 2nd edition, Stanford 1959.
[5] H. R. Schelbert, H. Kreuzer, J. Dittrich, H. Reitsma and P. Spiller, Videometrische Ventrikelflächenbestimmung mit halbautomatischer Korrektur des Bildhintergrundes, Res. exp. Med. **158**, 66-74, 1972.
[6] L. M. Henikoff, W. A. Stevens Jr. and L. W. Perry, Detection of heart disease in children 1919-1967, Circulation **38**, 375-385, 1968.
[7] W. Marton, M. S. Hoffman, R. L. Cleere and H. J. Dodge, Comparison of three methods of screening for pediatric heart-disease, J. Amer. Med. Ass. **169**, 1169-1172, 1959.

The aim of computer measurements on X-ray images of the heart is to provide an objective, reproducible and quantitative description of the image. In mass X-ray screening the images obtained are in the AP projection (anterio-posterior, i.e. from the front and perpendicular to the chest). The earlier work in our laboratories and elsewhere has been on automating the determination of the cardiothoracic ratio (i.e. of the sizes of the heart and thorax) and on the determination of the shape of the heart — both comparatively straightforward measurements. In clinical practice the cardiothoracic ratio is taken as a basic quantity; deviations from the 'normal' ratio indicate changes in the heart. Congenital heart defects in children, for example, can be inferred from the shape of the heart [1].

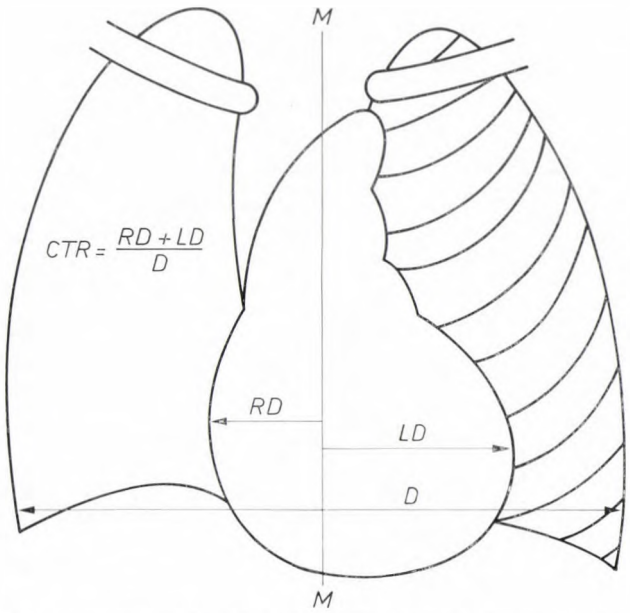


Fig. 3. The cardiothoracic ratio (*CTR*). This is a basic quantity in clinical practice. *MM* artificial axis of symmetry, determined by the computer. The width of the heart is defined, as usual, as the sum of the maximum distances between the axis of symmetry and the two half-contours. The width of the thorax is the value just above the diaphragm.

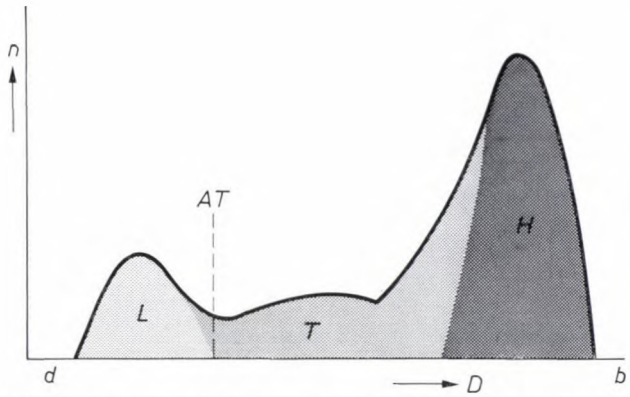


Fig. 4. The choice of an adaptive brightness threshold (*AT*) using the individual brightness distribution determined by the computer. The computer processes an X-ray image of the thorax and calculates a histogram of the resulting grey values (*D*, with *d* = dark and *b* = bright). The frequency of appearance *n* is plotted vertically. *H* heart. *T* tissue. *L* lungs.

The cardiothoracic ratio

Several methods for the determination of the cardiothoracic ratio, as defined in *fig. 3*, are known, but the equipment and procedures required differ somewhat.

H. C. Becker *et al.*[8] determine the width of the heart and the thorax from a 'signature' curve. The computer determines the signature curve by an integration of the brightness values along a strip, without considering other details of the image. The signature curve of an $n \times m$ image is represented by the function

$$S(k) = \sum_{l=1}^{m} I(k, l), \quad \text{with} \quad 1 \leqslant k \leqslant n,$$

where $I(x,y)$ is the brightness at the point (x,y). The width of the heart and thorax can be determined from the profile of the signature curve. The method does not of course yield exceptionally accurate values but the inherent error does not amount to more than a few per cent.

The methods that we have developed for the measurement of the cardiothoracic ratio and for the description of the shape of the heart originate in principle from a *contour image*. Processing this contour informa-

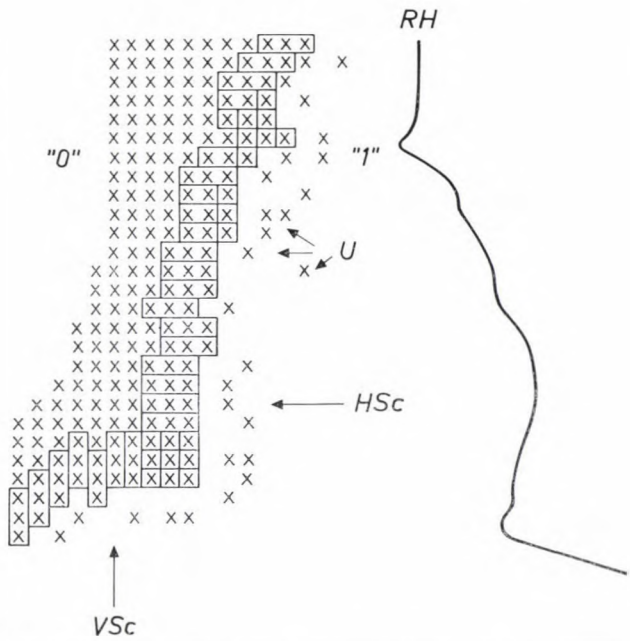


Fig. 5. The determination of the cardiac contour from a binary image field by searching for at least three successive '0' points (crosses). *U* '0' points caused by noise. *HSc* horizontal scan. *VSc* vertical scan. *RH* right half of the cardiac contour. The computer produces the binary field by application of a brightness threshold (fig. 4).

tion yields sufficient information to give the maximum extent of the heart and thorax in the transverse direction. The left and right halves of the cardiac contour are scanned in the transverse direction with the object of finding the maximum width. The width of the thorax is determined just above the diaphragm. The ratio is the required cardiothoracic ratio.

Images containing only a contour represent about a hundredth of the amount of information in the original complete image, and can therefore be processed very much faster.

tion of the patients, together with the exposure time and the development of the film, have a considerable influence on the content of the image and its contrast. Consequently a *fixed* brightness threshold cannot be used. We have therefore developed a method [9] in which an individual threshold is derived afresh for each image from a histogram of the brightness (*fig. 4*). Using known methods the computer then determines a binary image field from the derived threshold ('0' for points of brightness below the level, '1' for the other points). A boundary contour between the heart and

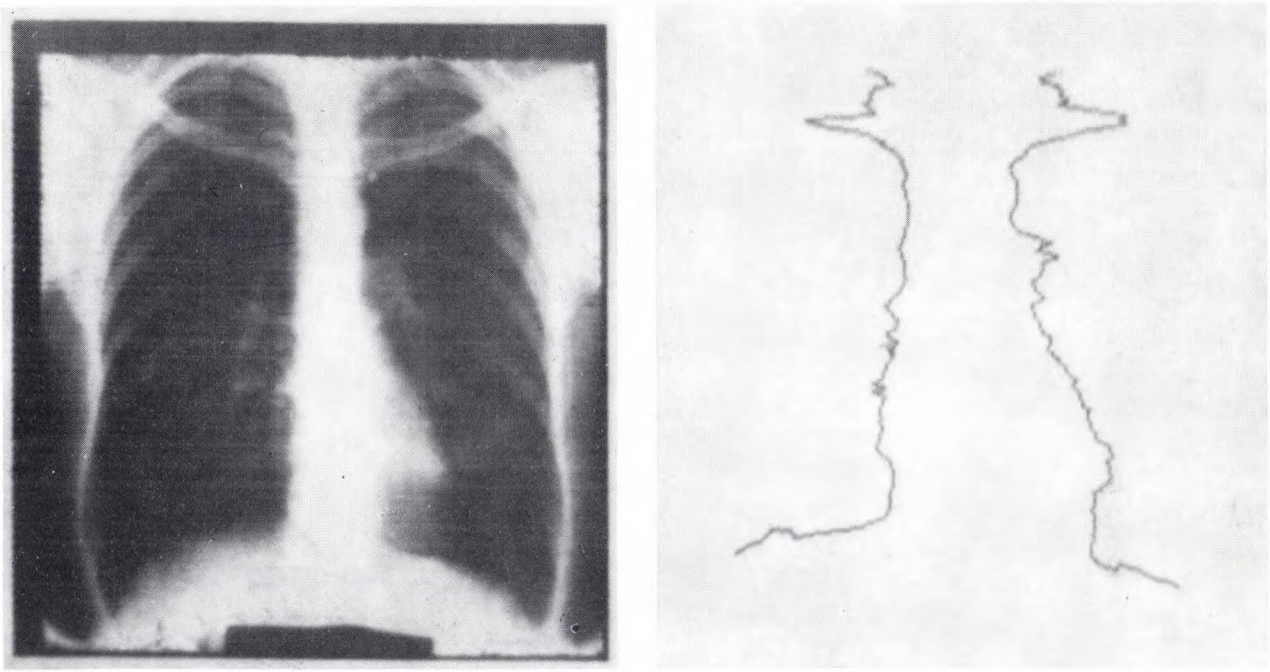


Fig. 6. Example of the search for a cardiac contour. *Left:* original X-ray photograph; *right:* the contour found by using a variable threshold and a binary field (figs. 4 and 5), which separates the heart and lung areas from each other.

For investigations on a large scale, the ability to determine the cardiac contour completely automatically is a good starting point. The classification of the measured contour is also an important point for diagnosis; however, a description of the parameters of the contours cannot then be avoided. Both points are discussed briefly below.

The cardiac contour

The universal image-processing methods described in the literature are not suitable for the determination of the cardiac contour, because they require a great deal of computer time. Simplified methods, such as those using a fixed brightness threshold, again have the disadvantage that they can only be used if the X-ray photographs are standardized. The size and the position of the patients, together with the exposure time and the development of the film, have a considerable influence on the content of the image and its contrast. Consequently a *fixed* brightness threshold cannot be used. We have therefore developed a method [9] in which an individual threshold is derived afresh for each image from a histogram of the brightness (*fig. 4*). Using known methods the computer then determines a binary image field from the derived threshold ('0' for points of brightness below the level, '1' for the other points). A boundary contour between the heart and lungs is now considered to have been found if at least three successive '0' points are found when scanning from inside to outside in the coordinate direction most perpendicular to the estimated contour. With this criterion it is always certain that a connected '0' region will be found as a boundary (*figs. 5* and *6*). The boundaries of the heart on the upper and lower sides cannot be derived from the X-ray image because the appropriate contours are not shown. It is therefore necessary to try to obtain them artificially. The lower boundary of the heart is defined by the line connecting the left- and

[8] H. C. Becker, W. J. Nettleton Jr., P. H. Meyers, J. W. Sweeney and C. M. Nice Jr., Digital computer determination of a medical diagnostic index directly from chest X-ray images, IEEE Trans. **BME-11**, 67-72, 1964.

[9] W. Spiesberger, Klassifizierung von Herzfehlern durch digitale Röntgenbildverarbeitung, Biomed. Technik **20**, Ergänzungsband Kongreß Medizin-Technik 1975, Stuttgart.

right-hand corner points where the heart, lungs and diaphragm meet. The computer finds these two corner points by measuring the distances between the appropriate part of the contour and a fixed reference point (*fig. 7*). This distance is a minimum at the desired corner point. Our current program uses the smallest horizontal distance to define the boundary on the upper side. This method enables a closed cardiac contour to be determined in a simple way (*fig. 8*). The contours obtained in this way can be used for the measurement of the cardiothoracic ratio, and also for classification purposes.

The parametric description of the cardiac contours

R. P. Kruger *et al.*[10] report that in heart diseases associated with a change in shape, it is possible to make an automatic diagnosis that in some respects will give more accurate results than the evaluation of X-ray photographs by a number of doctors. They have used a number of very different lengths, distances and angles as criteria.

C. T. Zahn and R. Z. Roskies [11] have described a method for the coding and classification of plane closed curves, that is independent of the position, orientation and size of the curves. If this method is applied to cardiac contours it should be possible to correlate automatically certain heart diseases or defects with the shape of the heart. Such a diagnostic technique does of course require that the classification programs should be applied to a very wide range of material.

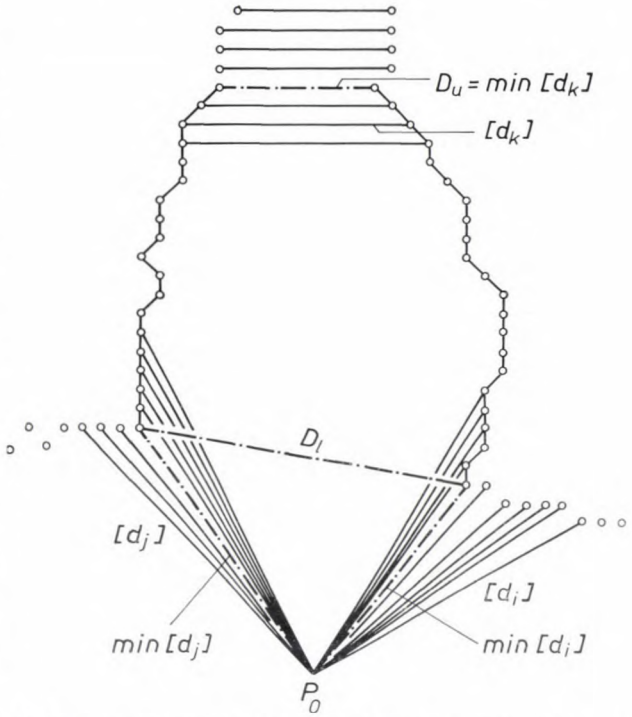


Fig. 7. Determination of the lower and upper boundaries D_l and D_u of the heart from the minima of d_i or d_j and d_k. P_0 fixed reference point. The method of contour determination is shown in figs. 4 and 5.

In the method of coding that we have developed, the contour is traced out in a clockwise direction and the change in direction is continually evaluated as a function of the distance travelled. The change in direction is equal to the angle between the local tangent and the

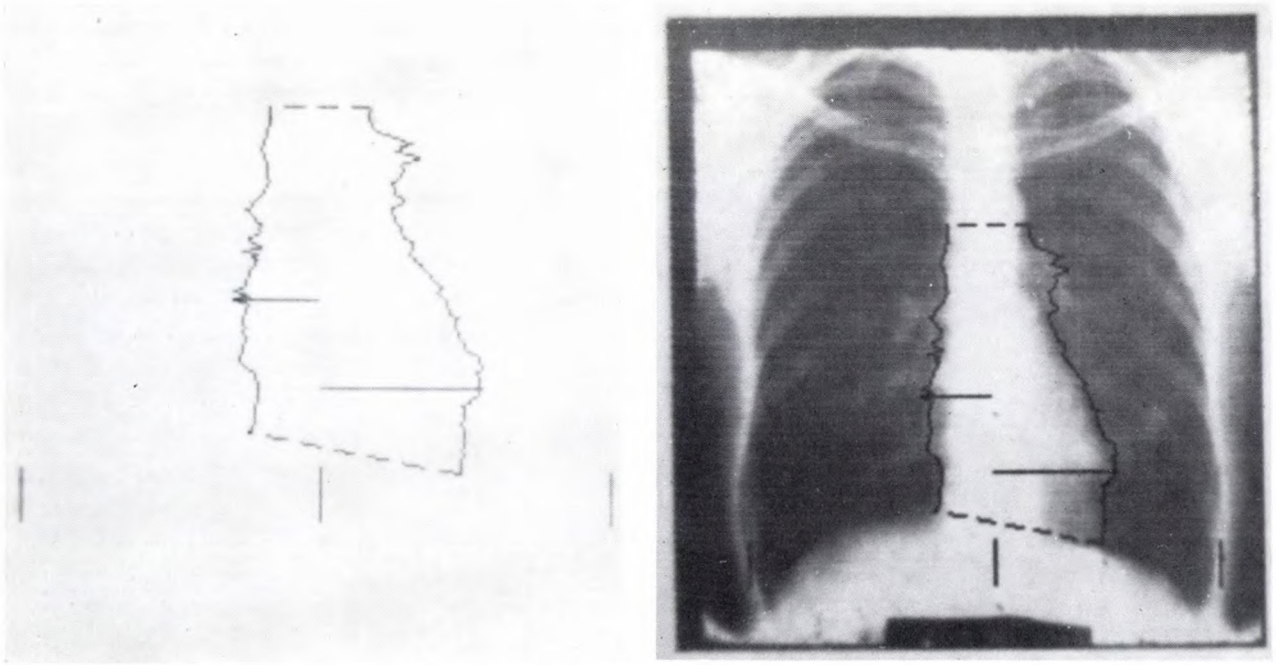


Fig. 8. *Left:* cardiac contour determined completely by the computer. The two horizontal lines establish the width of the heart. The width of the thorax is the distance between the vertical lines. *Right:* original photograph with the derived contour and distance lines drawn in.

tangent at the chosen starting point; the total contour length corresponds to an angle of 2π radians. The periodic shape function derived from this calculation is expanded in a Fourier series:

$$\Phi^*(t) = \mu_0 + \sum_{k=1}^{n} A_k \cos(kt - \alpha_k).$$

The function has been defined in the interval $(0,2\pi)$ and is equal to zero at the end points. The pair A_k,α_k are known as Fourier descriptors, μ_0 is a constant term which is related to the chosen starting point. The required details relating to the shape are supplied by the Fourier descriptors. As is always the case in Fourier analysis, the first terms of the series contain the 'coarse' information whereas the fine detail is contained in the terms of higher order. As the heart has no sharp edges or protuberances, a small number of terms will usually be sufficient. The Fourier descriptors that are necessary to describe the cardiac contour constitute a much smaller quantity of data compared with the contour itself. *Fig. 9* shows three contours that have been completely automatically constructed from 5, 10 and 100 pairs of descriptors.

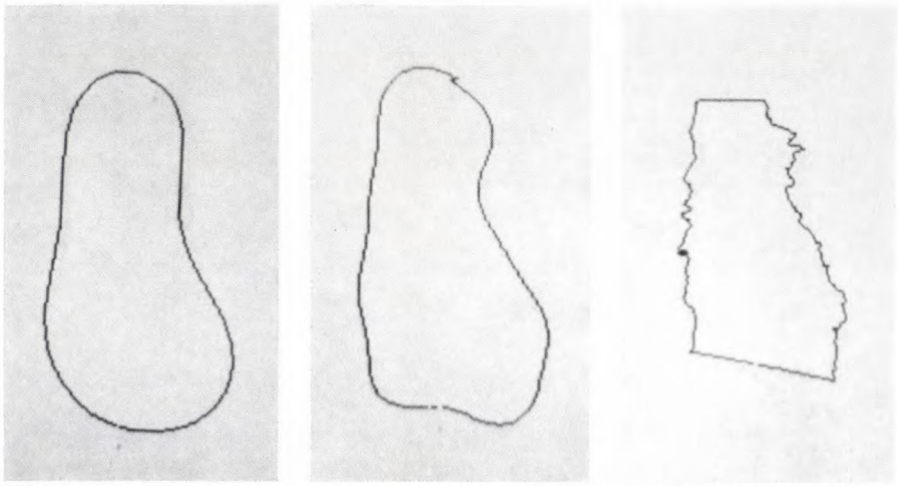


Fig. 9. The three displayed cardiac contours have been constructed with the help of 5, 10 and 100 pairs of Fourier descriptors. Small numbers such as 5 and 10 are essential if the information is to be used for classification purposes. In this way an efficient data reduction can be achieved.

A great deal of importance is attached to the reduction in the amount of data, because the amount of computer work in classification programs increases more rapidly than linearly with the number of parameters. The result of a computer analysis must of course be produced in a reasonable time, i.e. a time comparable with that which a doctor would need for the diagnosis. This consideration makes the parameterization of the cardiac contour a virtually essential aid in classification.

Computer determination of ventricle contours

The determination of the left-ventricle cardiac contour has three objectives: the determination of the volume as a function of time, the measurement of the wall movement at a number of places and the observation of wall thickness as a function of time. The results obtained are used for the refined examination of the cardiac function, in which more quantities such as the blood pressure and its variations are combined with the volume data. The ventricle is made visible by the injection of a contrast medium via a catheter. The image is usually recorded at a speed of 50-100 frames per second; the normal procedure is to record a single complete cardiac cycle, so that a filming time of one second is adequate. In *fig. 10* two photographs of a ventricle taken in rapid succession are shown. In general, other objects such as the diaphragm, ribs and aorta will be visible in the recorded image. In this case it will be necessary to apply a more complicated data-processing method than the simple threshold method.

The block diagram of our data-processing method is given in *fig. 11*. The first stage of the procedure is a coarse search for the contour, characterized by the use of only 32×32 image points during the scan, i.e. a low resolution. In the second stage the contour is more accurately located. Starting from the coarse contour already found, the computer searches the first frame for new position coordinates with a higher resolution. Then the complete film strip is scanned, one frame at

[10] R. P. Kruger, J. R. Townes, D. L. Hall, S. J. Dwyer III and G. S. Lodwick, Automated radiographic diagnosis via feature extraction and classification of cardiac size and shape descriptors, IEEE Trans. **BME-19**, 174-186, 1972.

[11] C. T. Zahn and R. Z. Roskies, Fourier descriptors for plane closed curves, IEEE Trans. **C-21**, 269-281, 1972.

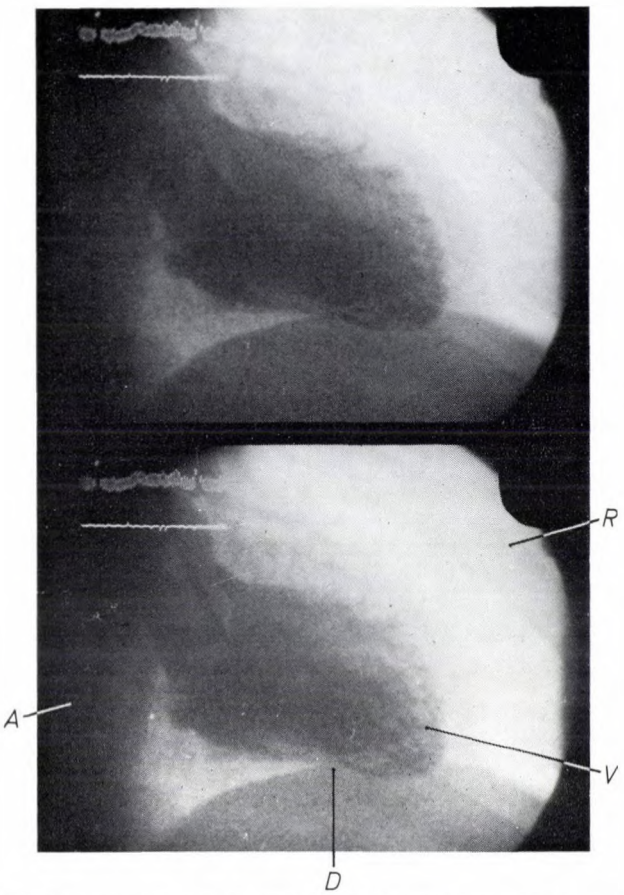


Fig. 10. Two X-ray photographs of a left cardiac ventricle. that has been filled with a contrast medium. The time that has elapsed between each exposure is about 20 ms. *V* ventricle. *R* rib. *D* diaphragm. *A* aorta.

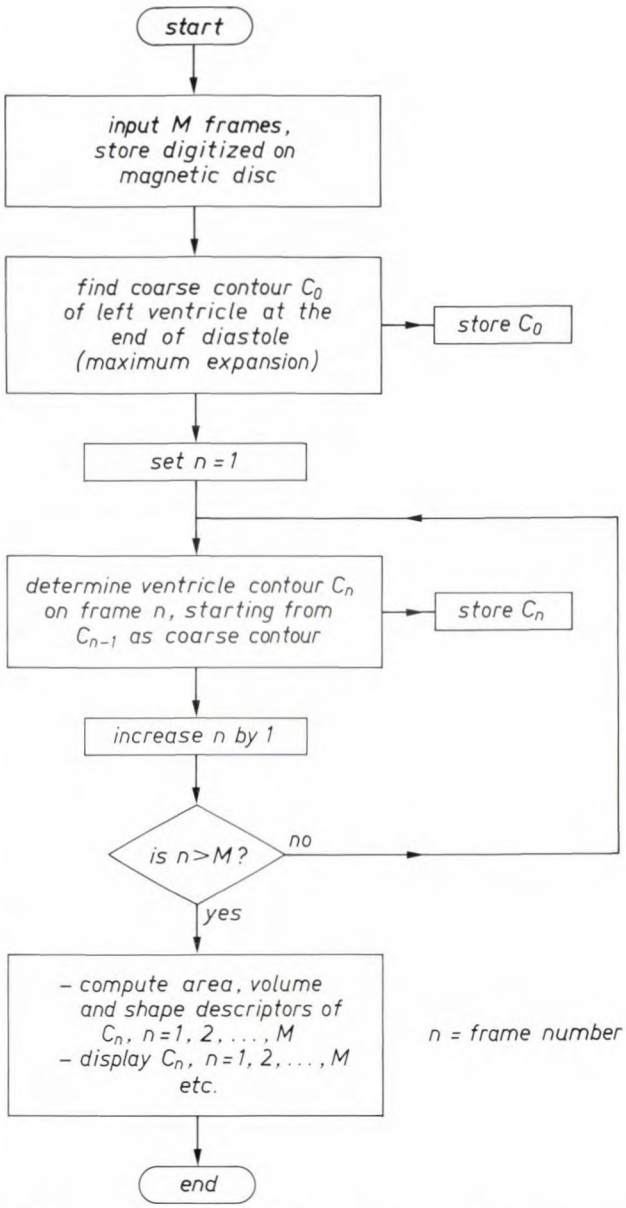


Fig. 11. Block diagram of the method for automatic determination of a left cardiac ventricle. The input information is on *M* X-ray frames; *n* is the frame number. The first step in the procedure is the coarse search; the high-resolution search occurs in the second, iterative step.

a time; at each new frame the computer makes use of the position coordinates from the previous frame in calculating the position coordinates of the new contour.

We shall now briefly describe two methods for the low-resolution search. In the next section we shall consider a method for the high-resolution search.

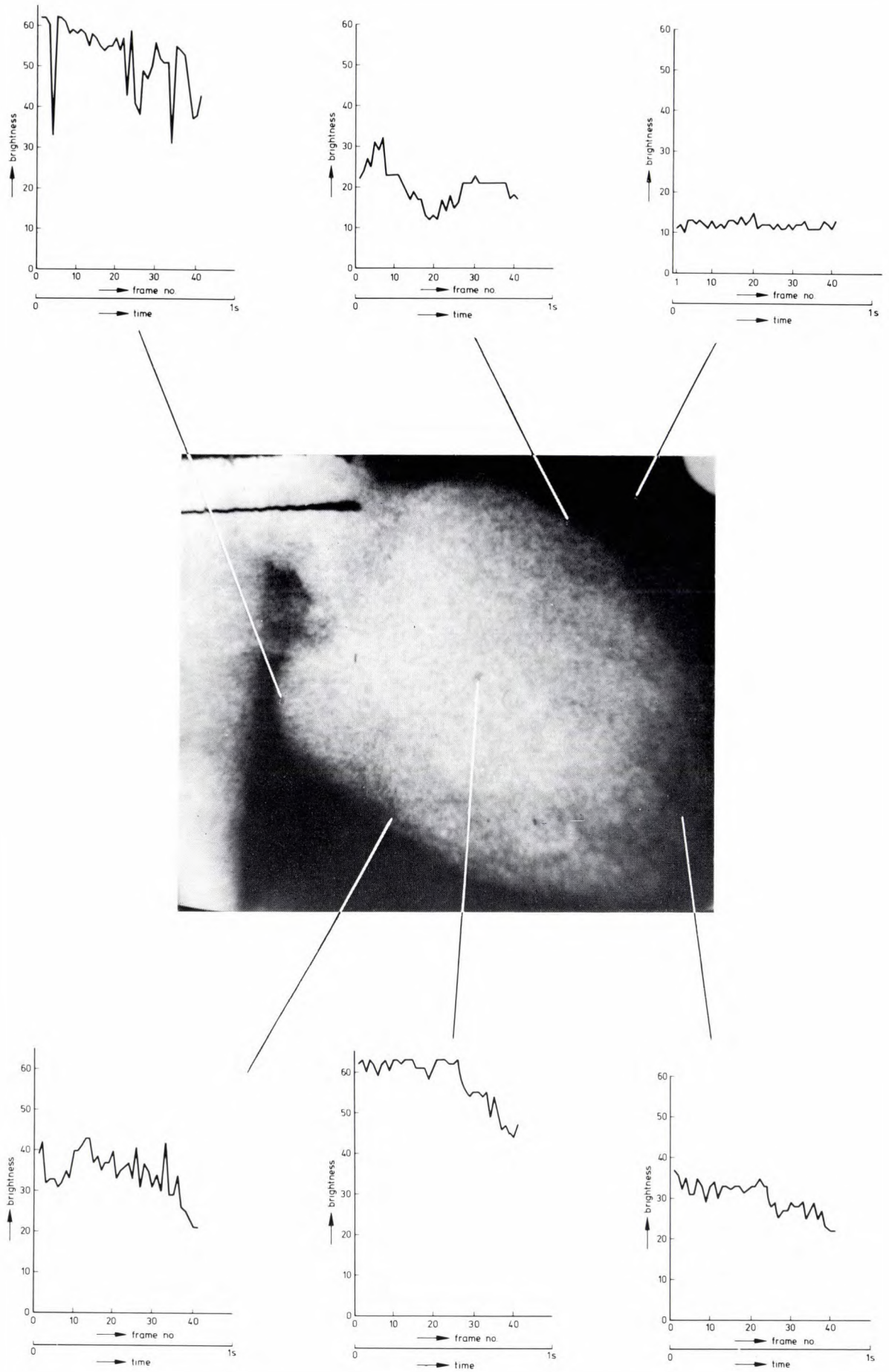
Our first method for the coarse search depends on the strong brightness variations in the region of the ventricle caused by the pulsation of the heart and the movement of the contrast medium. Outside the ventricle the brightness variations are much less, since they are determined there by noise or by respiratory movements, which are slow. Examples of the change in brightness as a function of time for a number of different regions are given in *fig. 12* [12]. If the change in the brightness rather than its absolute value is expressed quantitatively in a suitable way, e.g. as the mean-square deviation, then it is certainly possible to distinguish the area of the ventricle from its surroundings with a single threshold. The value of the threshold can itself be derived automatically by the computer from the variations; it does this by computing the mean of the average error, which is then very useful as the threshold.

Fig. 13 shows a binary field obtained when the threshold is applied to the change in brightness. The values '1' indicate points where the threshold has been exceeded, and all the others are equal to '0'. There are also points that have been labelled with a '1' because of

[12] M. Tasto, Motion extraction for left-ventricular volume measurement, IEEE Trans. **BME-21**, 207-213, 1974 (No. 3).

→

Fig. 12. The graphs give the brightness variation on approximately 50 successive photographs, in the indicated positions on the X-ray image of a left cardiac ventricle.


brightness
frame no.
time
1s

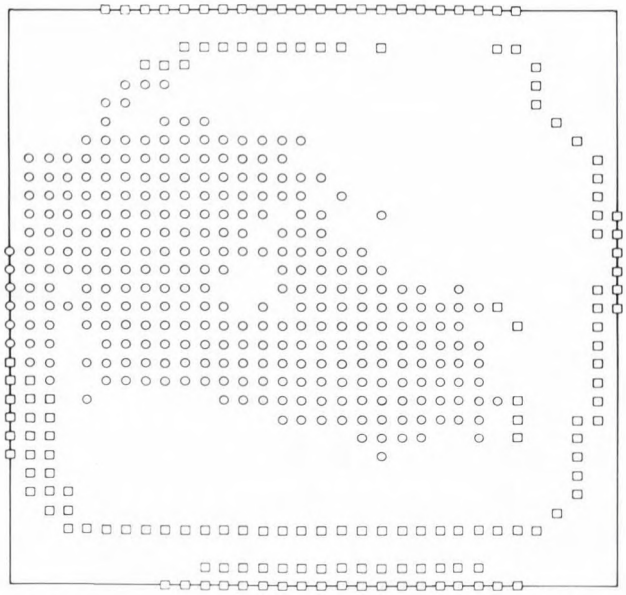


Fig. 13. The first method of coarse searching yields a binary field. The circles are the points where the threshold in the brightness change has been exceeded (level '1'). The squares are points to which a level '1' has been erroneously ascribed (e.g. by noise) and they could be eliminated later by a 'cleaning' process.

noise or the intrusion of the edge of the picture; these can be eliminated by using special noise-suppression methods ('noise cleaning'). A contour can be derived from the binary field, since the computer takes every '1'-'0' transition it encounters while it scans radially outwards as a point of the contour. The computer starts the scan somewhere close to the 'centre of gravity' of the field of '1' values.

The second method for low-resolution search is the 'template' method [13]. In this case the frame to be processed is first transformed into a 'dash-point figure', which implies that all the contours (including those of objects not required) are determined and reproduced initially in the form of dashed lines. The computer then moves a template of the required object over the dash-point figure, and the object is considered to have been found when the deviations between template and object are as small as possible. The distances between different points on the template and those points that lie closest on the contour of the given figure are taken as a measure of the deviation. An example of such a dash-point figure with a ventricle template mask that has been moved across it is depicted in *fig. 14*. The distance lines are also shown. In this idealized example the template is of exactly the same shape as the desired object; this never occurs in practice, of course. The distances decrease to zero when the template is accurately located over the object. *Fig. 15* shows a situation in which there is still a large error in the position of the template, as can clearly be seen from the distance lines.

High-resolution search

The contour that has been found with the help of one of the two coarse methods is used as starting material for the fine search. Three types of prior data are derived from the coarse contour: the direction and the magnitude of the interval to be searched and the sign of the

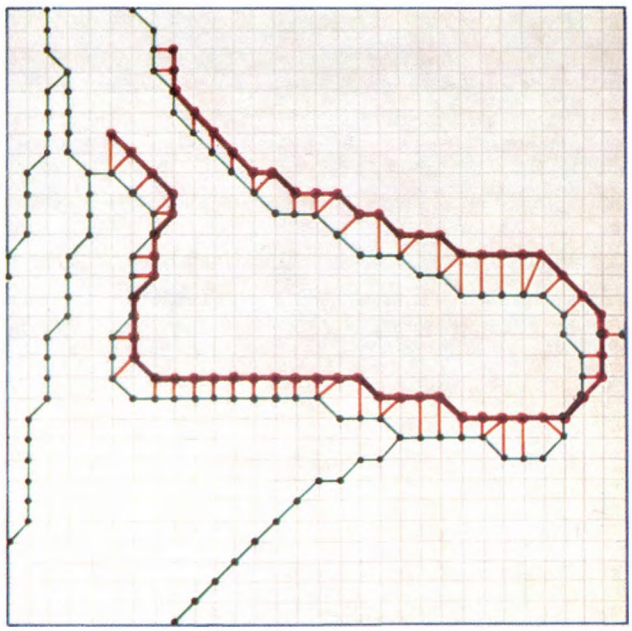


Fig. 14. The second method of coarse searching uses a template. A cardiac-ventricle template (the thick brown line) is shown as an example. This is 'moved' close to the object under investigation by the computer. The distance lines are therefore short. The case shown has been idealized; in practice the template and object will not have precisely the same shape.

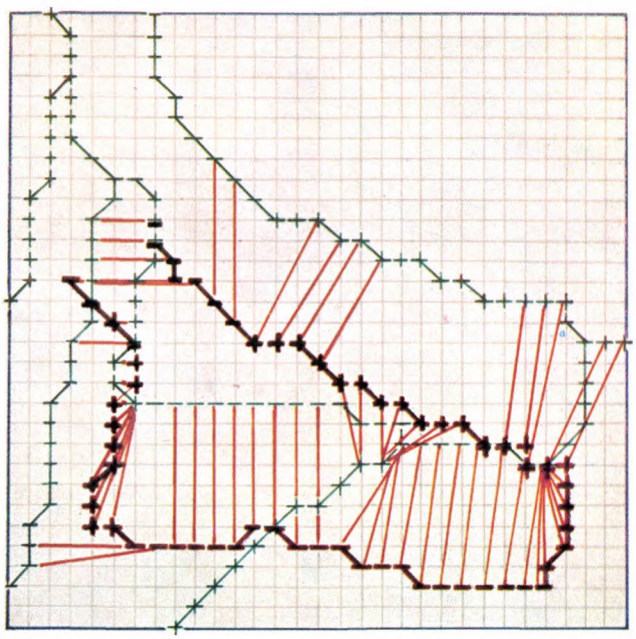


Fig. 15. The template is still some way away from the object. The distance lines are much longer than in the situation shown in fig. 14.

brightness gradient to be determined [14]. In *fig. 16* a contour found by the coarse method is drawn, with a normal constructed by the computer; the normal establishes the direction of search for a new contour point. The magnitude of the interval to be searched is also known to the computer, as the maximum uncertainty in the position of the coarse contour.

When a series of frames are processed, the computer always takes the contour from the preceding frame as the coarse contour for the ensuing fine search, as we saw above. The maximum uncertainty follows, in this case, from the maximum rate of travel of the ventricle wall and the frame rate.

The search for a contour point in a fixed interval occurs as follows. The computer determines the brightness gradient, which is scanned radially from inside to outside in the search direction. The point where the gradient has the greatest positive value is the point that the computer is looking for. The *sign* of the gradient obtained is the means by which the computer distinguishes the required contour from undesirable overlapping objects, such as the diaphragm. The brightness must always *decrease* at the contour being searched for when the scan goes from inside to outside.

In *fig. 17* two heart-ventricle contours have been drawn on the original X-ray photograph. These contours were determined fully automatically by the computer. The boundary line between aorta and ventricle can be deduced approximately by searching at the side of the ventricle directed towards the aorta for the place where two contour halves that lie opposite each other have the smallest spacing (the short dashed line). Finally, for computing the heart volume, the computer constructs the axis of symmetry of the contour (the long dashed line). In the computer program it is assumed that the cross-section across the long axis of the ventricle is elliptical or circular, depending on whether the computer has access to images in two projection directions or only in one direction (as is most usual).

Current implementation and future prospects

The methods that have been described so far can be executed on medium-sized computers with a memory capacity of 32 kbyte or more. The programs should ideally be formulated in 'machine-independent' languages; the methods may then be implemented with few compatibility problems in most medical centres. The digitizing of images, as well as the presentation of processed images and results, can be separated from the pure data-processing activities. Standard television equipment can be used for digitizing and presentation. In all our ventricle methods only relevant information from each image is digitized and this results in processing times of only a few seconds. The cardiothoracic procedure with its associated parameter description always utilizes the same number of image elements. The rather computation-intensive algorithm for the classification of the contour information results in a computing time of several minutes for each image.

The determination of the ventricle contour and the cardiothoracic measurements dealt with in the earlier sections are only two examples from a multitude of possible medical applications. A number of the image-processing operations used are in many cases identical and are employed for different purposes in different applications. It seems desirable to give a modular structure to a general data-processing program for X-ray photographs, to avoid having to re-program each image-processing routine at every changeover to a new field of application. With a suitable range of appropriately designed image-processing programs it would then be possible to test the solution of new problems with the existing modules. The time required, compared with the current situation (writing a complete program, testing, modification, etc.) would be much shorter. If the way to a solution was found by experimenting with

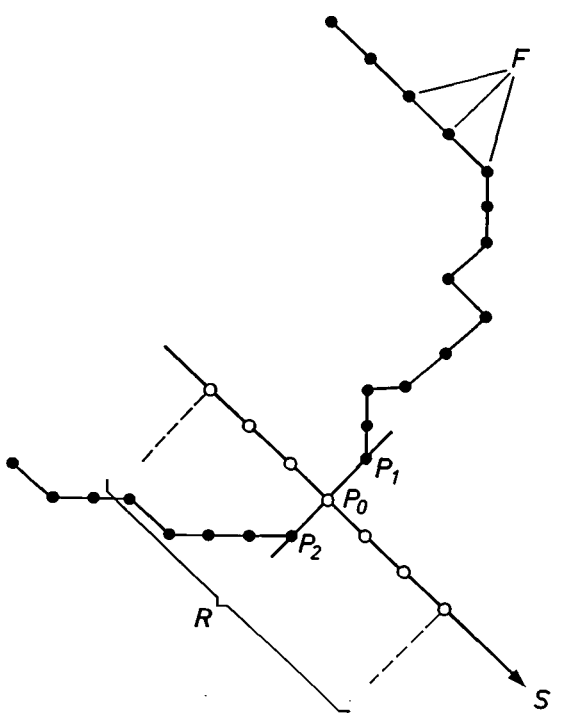


Fig. 16. Principle of fine searching for a cardiac contour. *F* coarse contour; this provides the initial input data. *S* the new search direction, constructed by the computer as a normal to the coarse contour. ($P_{0,1,2}$ are additional construction points). *R* search interval; its size is determined by the maximum error in *F*.

[13] M. Tasto and U. Block, Locating objects in complex scenes using a spatial distance measure, Proc. 2nd Int. Joint Conf. on Pattern Recognition, Copenhagen 1974, pp. 336-340.

[14] M. Tasto, Guided boundary detection for left ventricular volume measurement, Proc. 1st Int. Joint Conf. on Pattern Recognition, Washington D. C. 1973, pp. 119-124.

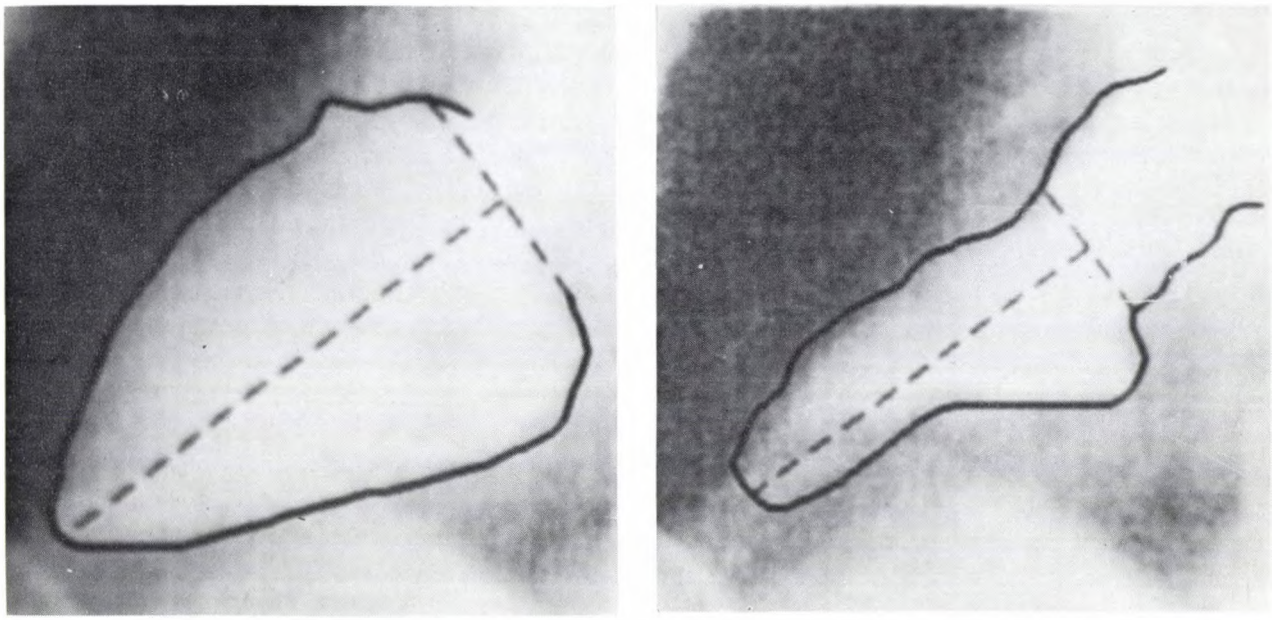


Fig. 17. Two examples of cardiac contours that have been located automatically, both displayed on the original X-ray photograph. The short dashed line determines approximately the boundary between the aorta and the cardiac ventricle. The long dashed line is an artificial axis of symmetry, which the computer uses for the continuous, completely automatic computation of the ventricle volume. In the right-hand photograph, the ventricle is in its contracted state (systole) and in the left-hand photograph it is in its expanded state (diastole)

the data-processing programs, a separate program for the particular case could always be written later — which would not take long.

The time and effort required, in the short term, for the preparation of such a useful general image-processing program package, would be completely recovered in the long term by the resulting possibilities for rapid and flexible image-processing in many medical (and non-medical) situations. We hope that this more general solution to the problem will become reality in the not-too-distant future.

Summary. X-ray diagnostic methods are in need of automated image-processing procedures (= measurement of photographs + processing of the results), especially for preventive health care. Program packages, both existing and under development, for selection, recognition and digital processing are discussed; the emphasis is on cardiothoracic measurements and the determination of cardiac-ventricle contours. The cardiothoracic measurements concentrate on the automatic determination of the width ratio from the contour measurements in AP projection. Further data reduction, necessary for shape-classification investigations, is accomplished by parameter representation with Fourier descriptors. The contour recognition — by a gradient method — is performed iteratively on a series of film images of one cardiac cycle; the computer finds the required initial contour with a threshold or a template method. Noise errors can be eliminated.

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 7/8

Crystal defects and their transport in solids; industrial applications

H. J. Vink

Introduction

The manufacture of materials that meet electrical, magnetic, optical and mechanical specifications is sometimes described as 'molecular engineering'. It is the art of placing the desired atoms with the desired valency at the desired crystallographic sites of the desired crystallographic structure. The materials obtained in this way more often than not contain a number of different kinds of atoms in wide ranges of concentrations. They are applied as bulk materials, and frequently as layers of various compositions, with thicknesses as small as 1 μm in devices such as transistors, and down to 1 nm or less in some ceramic components.

In this article special attention will be drawn to the transport effects in the solid state involving vacancies or dopant atoms. Such effects are important in the manufacture of all kinds of materials. Only two types of transport will be dealt with. The first to be discussed will be the diffusion in a semiconductor like silicon. The driving force for this material transport is a consequence of the existence of a concentration gradient. The second, which will be discussed more briefly, is the sintering process. Here the driving force is a consequence of the fact that the surface energy of a compacted powder decreases on sintering.

Both mechanisms will be treated in a similar way. First of all some account will be given of the general background associated with these effects. Some of the difficulties encountered in the actual processes of manufacture will then be dealt with in the light of these principles. Three types of material will be chosen as illustrative examples: one electronic, one optical and one magnetic. For the electronic material silicon is an obvious choice — it is after all the material from which such complicated structures as integrated circuits are made, with the help of the diffusion process. The optical and magnetic examples have been chosen to illustrate the various important points in the sintering process. Two optical materials will be treated briefly: dense Al_2O_3, a translucent ceramic suitable for envelopes for high-pressure sodium lamps operating at high temperature, and dense Y_2O_3, a material characterized by an extremely low porosity and which can be used as the matrix material in solid-state lasers. Finally, Mn-Zn ferrite has been chosen as an example of a magnetic material. This ferrite is used after appropriate molecular-engineering processing a) for applications in transformers, b) as a core material in radio-frequency *LC* filters and c) — because of its low porosity and its small pores — in recording heads for video-tape equipment.

The treatment will show how important scientific knowledge is for the understanding and control of industrial processes, and also that the actual process of manufacture invariably leads to new avenues of scientific research and a deeper understanding of fundamental phenomena. It is this interaction between research and industry that makes industrial research such an interesting world.

Diffusion of donors and acceptors in silicon

As stated earlier, the discussion here will be limited to an account of diffusion in the semiconductor silicon. It is well known that this is a process of great technological importance, since it has been used for the last 25 years to fabricate *P-N* junctions in diodes, transistors and integrated circuits. In view of this technological importance it is hardly surprising that much work has been done on elucidating the atomic mechanisms of diffusion in semiconductors. In a recent review of atomic diffusion in semiconductors [1] [*] the chapter on Si and Ge, although by no means complete, contains about 400 references. It is therefore surprising that, notwithstanding all this, the diffusion of donors

Dr H. J. Vink is a Director of Philips Research Laboratories, Eindhoven. *The text of this article is an expanded version of a lecture given on the 10th of May 1975 in Vienna for the annual meeting of the Deutsche Bunsen-Gesellschaft. The original English text of the lecture can be found in Berichte der Bunsen-Gesellschaft für physikalische Chemie 79, No. 10, 1975.*

[*] The references are listed at the end of the article.

and acceptors in Si is still not well understood. It is found that under the circumstances of the actual manufacture of *P-N* junctions, the concentrations used are often so high that a number of secondary effects become important enough to dominate the result. Again, even for diffusion in conditions where high concentrations do not occur, and these secondary effects are therefore of no importance, it is only recently that experiments have provided any understanding of the basic phenomena.

The primary diffusion process

Experimental results

A few words should first of all be said about this primary diffusion process. Until recently there was considerable scatter in the reported diffusion parameters in silicon. The activation energy E_D (eq. 1) for boron in silicon varied from 1.67 eV to 4.25 eV and the pre-exponential factor D_0 (eq. 1) varied from 10^{-3} to about 10^2 cm^2/s. A similar scatter was found for phosphorus in silicon [2]. R. N. Ghoshtagore has suggested that, although the effects due to high concentration might perhaps have played a role in these determinations, the scatter of these results can be explained by the fact that most experiments used an oxide-diffusion source on the silicon surface. Effects such as 'out-diffusion' and evaporation, surface strain, boundary motion and reactions between the source and the underlying silicon — the latter leading to deviations in the concentration of intrinsic defects — may therefore have influenced the results [3–5]. In Ghoshtagore's experimental arrangement all these effects were excluded and the concentrations of the dopants were made sufficiently low for the silicon always to be considered as intrinsic at the diffusion temperatures (concentrations $< 2\times10^{19}$ cm^{-3}).

Under these conditions (a temperature between 1130 and 1450 °C) the diffusion of B and P in silicon via a (111) plane can be described by Fick's laws. *Fig. 1* makes this clear. The line drawn through the experimental points corresponds to the solution of Fick's laws (for constant diffusivity) for the boundary conditions of Ghoshtagore's experiments.

Similar results were obtained for the diffusion of the donors As, Sb and Bi in Si and also for the acceptors Al, Ga and In. In all cases the diffusivity D can be described by

$$D = D_0 \exp(-E_D/kT). \qquad (1)$$

D is expressed in cm^2/s. The values of D_0 and E_D for the various donors and acceptors are listed in *Table I.*

From fig. 1 and *fig. 2* it can be seen that the group-III and group-V elements diffuse rather slowly in silicon. Most of them have similar activation energies and pre-exponential constants.

Brief survey of the various mechanisms proposed for the primary diffusion process

Various mechanisms have been proposed for the 'intrinsic' diffusion of the group-III and group-V elements in silicon. These elements are present almost entirely in the substitutional form. A diffusion mechanism based on vacancies is therefore often assumed; however, there have been suggestions that a kind of interstitial mechanism is predominant. D. L. Kendall and D. B. de Vries [7] have suggested in a comprehensive review article that the mechanism for the group-V elements depends on monovacancies or divacancies, whereas A. Seeger and K. P. Chik [8] have argued in favour of a kind of interstitial mechanism. In a more recent article, S. M. Hu [9] has proposed a dual mechanism, with the interstitial component predominating for boron — and phosphorus, however not to the same extent — but with arsenic diffusing mainly by a vacancy mechanism.

Ghoshtagore accepts the principle of a vacancy mechanism for the group-V elements. However, in his opinion, the equilibrium vacancy concentration is too low to explain the diffusivities. He therefore assumes the *impurity-vacancy pair* to be the main agent in the

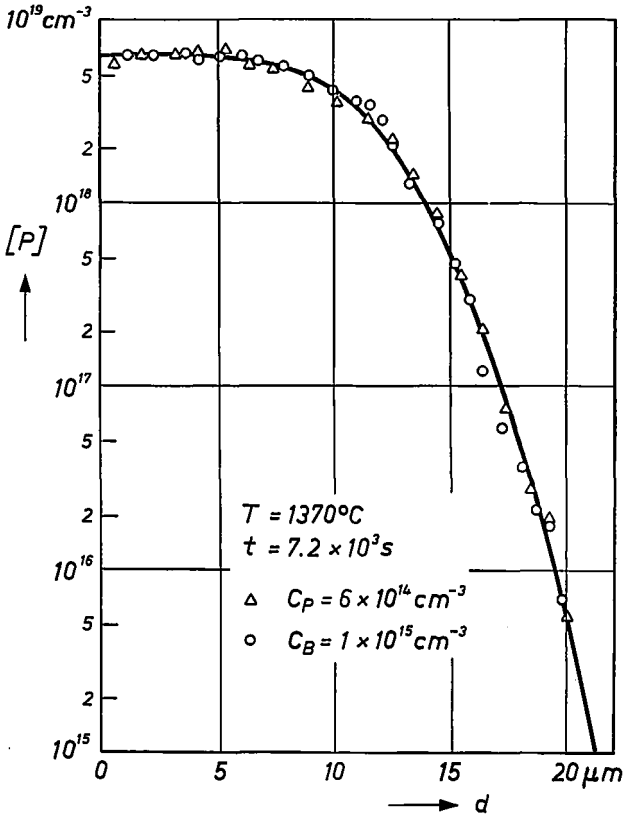


Fig. 1. A concentration profile for phosphorus in silicon, for both *P*-type material (the triangles) and *N*-type material (the circles). The horizontal scale indicates the distance d travelled by the diffusion, which started at the surface. The heat treatment was the same in both cases. The diffusion time and temperature are shown in the figure, and also the concentrations of the charge carriers in the starting material. (After R. N. Ghoshtagore [2].)

Table I. The values of the pre-exponential factor D_0 and the activation energy E_D from the equation for the diffusivity (1), for various donors and acceptors in silicon. (After Ghoshtagore [6].)

Dopant	D_0 (cm²/s)	E_D (eV)
P	7.4×10^{-2}	3.30
As	6.55×10^{-2}	3.44
Sb	2.14×10^{-1}	3.65
Bi	1.08	3.85
B	2.1×10^{-3}	2.85
Al	1.39	3.41
Ga	0.37	3.39
In	0.79	3.63

diffusion of these impurities in silicon. This mechanism will now be treated in somewhat more detail. The argument will be restricted to the donor-vacancy pair. It will be seen that a fairly coherent picture can be obtained by considering enthalpies alone. Strictly speaking, entropy should also come into the picture, but for brevity it will not be considered here. At this point a word of warning is perhaps appropriate: although it seems to hold together logically the mechanism of donor diffusion by means of donor-vacancy pairs cannot yet be considered as definitely established.

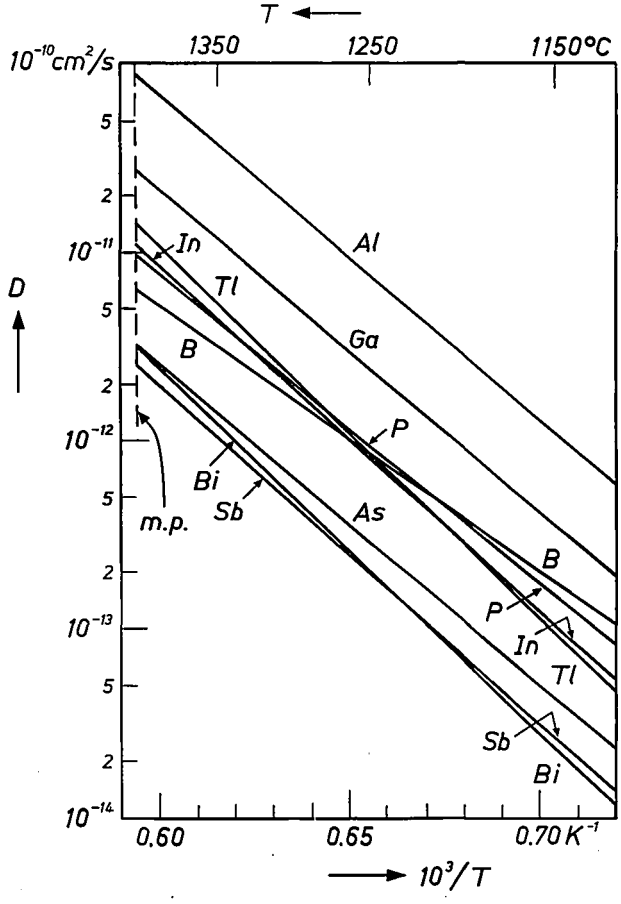


Fig. 2. The diffusivities D for a number of elements from groups III and V in intrinsic silicon as a function of the temperature T. *m.p.* melting point of silicon. (After Ghoshtagore [6].)

Quantitative discussion of the activation energy of diffusion on the basis of the donor-vacancy-pair mechanism

The arguments of M. Yoshida and R. R. Hasiguti [10] suggest that the donor-vacancy mechanism is in fact the most promising. As in Yoshida and Hasiguti's article the activation energy (E_D) of diffusion of the group-V elements in silicon can be put equal to

$$E_D = E_V^f - E_{I\text{-}V}^b + E_{I\text{-}V}^m, \qquad (2)$$

where E_V^f is the energy of formation of a monovacancy, $E_{I\text{-}V}^b$ the binding energy of the formation of an impurity-vacancy (I-V) pair (an 'E centre') and $E_{I\text{-}V}^m$ the activation energy for the migration of these E centres. From Ghoshtagore's values of E_D, and if values of E_V^f and $E_{I\text{-}V}^m$ are known, the binding energy $E_{I\text{-}V}^b$ can be calculated for various donors. If the values of $E_{I\text{-}V}^b$ arrived at in this way appear to be acceptable, this argument can be regarded as evidence for the validity of the idea of predominant impurity-vacancy-pair migration.

No reliable direct experimental determination of E_V^f exists. This could be because the equilibrium concentration is very low — which is indeed probable — and because the vacancies have a high mobility, which implies that higher concentrations can only exist for a very short time. Therefore E_V^f can be determined only by indirect means or by theoretical calculation. The theoretical calculations [11–14] lead to values of E_V^f that lie close to 2.3 eV. Indirect experimental determinations have been discussed by L. Elstner and W. Kamprath [15], who carried out Hall measurements on heat-treated and quenched P-type silicon. They tentatively proposed a value of 2.5 eV for E_V^f, one of their assumptions being that vacancies were 'frozen in' by the quenching.

E_V^f can also be determined from self-diffusion experiments. The activation energy E_{Ds} for self-diffusion can be put equal to

$$E_{Ds} = E_V^f + E_V^m, \qquad (3)$$

assuming a mechanism of self-diffusion dominated by the migration of monovacancies. E_V^m is the activation energy of migration of such a monovacancy.

One method of determining E_{Ds} is to carry out diffusion experiments by the tracer method. The following values have been found in this way for E_{Ds}: 5.13 eV [16] [17], 4.78 eV [18] and 4.73 eV [19]. Another, more indirect, way to obtain E_{Ds} is to study the diffusion of Ni and Au in Si. In this case substitutional and interstitial Ni or Au atoms are in local equilibrium with monovacancies [20] [21]. A value of $E_{Ds} = 4.5 \pm 0.5$ eV has been deduced from these experiments. Taking all these results into account a value for E_{Ds} of 4.8 ± 0.2 eV seems reasonable. It is

now necessary to know E_V^m to calculate the energy of formation E_V^f from eq. (3).

Again, there is uncertainty with regard to the value of this activation energy. Two values of E_V^m have in fact been found. At low temperature (160 K) a value of $E_V^m = 0.33$ eV has been determined [22]. Electron-paramagnetic-resonance and infrared-absorption experiments [23] [24] have clearly confirmed that this value can be attributed to the presence of monovacancies. At high temperatures ($\approx$ 1700 K) the activation energy has been found to be as high as 1.2 ± 0.2 eV. This high value was determined from quenching experiments [15] [25] [26] and their interpretation in terms of what kind of defect actually migrates is uncertain.

Divacancies have been proposed and also 'extended defects' [8]. Recently ideas have been put forward that indicate that these models might be debatable thermodynamically and that the high-temperature value should likewise be attributed to a monovacancy [27]. The higher value of E_V^m would then be caused by another transition state, in association with the higher temperature. The diffusion experiments (with $E_{Ds} = 4.8 \pm 0.2$ eV) therefore enable a value of $E_V^f = 4.5 \pm 0.2$ eV for $E_V^m = 0.33$ eV or $E_V^f = 3.6 \pm 0.4$ eV for $E_V^m = 1.2 \pm 0.2$ eV to be calculated. On examining the values for E_V^f that have been obtained from different arguments: 2.3, 2.5, 3.6 ± 0.4 eV and 4.5 ± 0.2 eV it can be seen that there is indeed little certainty in the value of the energy of formation of a monovacancy.

For $E_{I\text{-}V}^m$ in eq. (2) the certainty is greater. These impurity-vacancy pairs can be produced by neutron irradiation and have been extensively studied by G. D. Watkins and his associates [22–24] [28] and also by a group in Japan [29]. The value of $E_{I\text{-}V}^m$ can be evaluated from thermal annealing (M. Hirata *et al.*) and also from the kinetics of the reorientation of these pairs on recovery from stress-induced alignment (Watkins *et al.*). Hirata and associates did this by means of life-time measurements, whereas Watkins and associates used the much more specific electron-spin-resonance method. If it is assumed that the final step — the direct exchange of the impurity with the adjacent vacancy, see *fig. 3* — does *not* determine the rate, then the process that limits the migration of the pair must be that of the reorientation. Both methods give values of $E_{I\text{-}V}^m$ that are in striking agreement (see the third column of *Table II*).

However, it should be noted that the temperatures used in these experiments never exceeded 200 °C. It is therefore not impossible that for the temperatures normally used in diffusion the values of $E_{I\text{-}V}^m$ will in fact be different.

Table II gives a list of the values of $E_{I\text{-}V}^b$ for the various impurities, calculated from eq. (2) using the experimental values of E_D determined by Ghoshtagore, the (low-temperature) values of $E_{I\text{-}V}^m$ found by Watkins and Hirata and taking $E_V^f = 3.6 \pm 0.4$ eV as a reasonable guess. The binding energy $E_{I\text{-}V}^b$ found in this way for the different impurities appears to be constant to within about 0.1 eV. Although unexpected, this is perhaps not entirely unreasonable and could therefore be taken as an indication that the arguments used above have some inner (quantitative) coherence. It is perhaps also reasonable to assume, from fig. 3, that the binding energy $E_{I\text{-}V}^b$ must always be greater than the activation energy of migration $E_{I\text{-}V}^m$. This would signify that $E_{I\text{-}V}^b$ has a value of 1.6 ± 0.1 eV, and that the energy of formation E_V^f of a monovacancy in Si has a value of about 4.0 eV, thus excluding the values of 2.3 and 2.5 eV found by theoretical calculations or deduced from the experiments of Elstner and Kamprath [15].

Table II. Calculated binding energies $E_{I\text{-}V}^b$ for various vacancy-donor pairs in silicon. The values were derived from the equation $E_D = E_V^f - E_{I\text{-}V}^b + E_{I\text{-}V}^m$, using Ghoshtagore's [6] results for E_D and those of Watkins [22–24] [28] and M. Hirata [29] for $E_{I\text{-}V}^m$ and a value of 3.6 ± 0.4 eV for E_V^f (all in eV).

Donor	E_D (exp.)	$E_{I\text{-}V}^m$ (exp.)	$E_{I\text{-}V}^b$ (calc.)
P	3.30	0.93	1.2 ± 0.4
As	3.44	1.07	1.3 ± 0.4
Sb	3.65	1.29	1.2 ± 0.4
Bi	3.85	1.46	1.2 ± 0.4

Some secondary diffusion processes

The effects of the Fermi level and of the built-in electric field

In making *P-N* junctions the concentrations of the impurities from groups III and V are often much higher than the concentration of intrinsic carriers at the

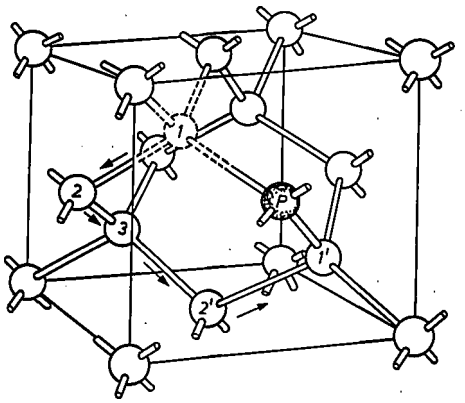


Fig. 3. The crystal lattice of silicon with a vacancy (*1*) and an impurity atom (*P*). It is shown how the direct exchange of an impurity atom with a nearest-neighbour vacancy can take place in this lattice (G. D. Watkins and J. W. Corbett [23]). The vacancy in the first nearest-neighbour position (No. *1*) must make two jumps away from the *P* atom and then arrives at position No. *3*; after this the vacancy returns to another nearest-neighbour position (No. *1'*).

diffusion temperatures. This leads to a number of effects that determine diffusion rates and profiles in the manufacturing processes. For instance, by keeping the concentration of diffusing donors higher than the intrinsic concentration of free electrons, the concentration of free electrons will be determined mainly by the concentration of these donors, which is time- and location-dependent. This situation gives rise to two effects. First, the higher concentration of free electrons will give a larger total concentration of vacancies. This is a consequence of the fact that a (neutral) monovacancy in the silicon lattice ($V_{Si}^{\times}$) acts as an acceptor [22–24] with an energy level 0.5 eV below the conduction band. On accepting an electron from the conduction band the silicon vacancy then becomes effectively negatively charged (V'_{Si}). The capture of an electron (e′) by a neutral silicon vacancy can be described by the reaction

$$V_{Si}^{\times} + e' \rightarrow V'_{Si}. \qquad (4)$$

In equilibrium, by the law of mass action, this leads to:

$$K_1 = [V_{Si}^{\times}]n/[V'_{Si}], \qquad (5)$$

or

$$[V'_{Si}] = K_2 n. \qquad (6)$$

In these equations $[V_{Si}^{\times}]$ and $[V'_{Si}]$ are the concentrations of neutral and singly negative monovacancies in the silicon lattice and n is the concentration of the free electrons. The concentration $[V_{Si}^{\times}]$, and hence the reaction constant K_2, is dependent on temperature only. From eq. (6) it follows that for a concentration of donors higher than the concentration n_i of free electrons in intrinsic silicon at the diffusion temperature, the concentration $[V'_{Si}]$ will increase with the donor concentration. For donor diffusion $[V'_{Si}]$ will be much larger than $[V_{Si}^{\times}]$. Since donors are effectively positively charged, their diffusion will be increased as effectively by a negatively charged silicon vacancy as by a neutral vacancy. Either a vacancy mechanism or a donor-vacancy-pair mechanism will therefore give an enhanced donor diffusion for increasing donor concentrations ($> n_i$). This is the 'Fermi-level effect'. This effect is in fact further enhanced, first because the singly negative vacancy V'_{Si} has a second acceptor level at 0.18 eV below the conduction band, and secondly because the donor-vacancy pair is also an acceptor, with a level at about 0.4 eV below the conduction band.

A second effect (due to the built-in electric field) is a consequence of the change, as a function of distance, of the position of the Fermi level — with respect to the conduction band — owing to the varying concentration of the diffusing dopant. Because of this, electric fields are created in the semiconductor, which in general give an enhancement of the diffusion of the impurity.

Both effects, of course, are interrelated as they are both influenced by the position of the Fermi level. The two effects cannot therefore be treated separately, but only in a single coherent mathematical treatment [1] [30]. The (isothermal) diffusivity $D(A)$ of a donor-type impurity A can be described, after making certain simplifications, by:

$$D(A) = D^{i}(A)\gamma_V^{-1}\left(1 + \frac{\partial \ln \gamma_A}{\partial \ln [A]}\right), \qquad (7)$$

where $D^i(A)$ is the diffusivity of the donor A in intrinsic material, [A] the donor concentration and γ_V and γ_A the activity coefficients for vacancies and donors respectively. The coefficients are defined by:

$$\mu_A = kT \ln\left(\frac{N_A}{N_S}\gamma_A\right) \qquad (8)$$

and

$$\mu_V = kT \ln\left(\frac{N_V}{N_V^0}\gamma_V\right), \qquad (9)$$

where μ_A and μ_V are the chemical potentials for donor and vacancy respectively; N_S denotes the total number of lattice sites and N_A and N_V denote the numbers of lattice sites occupied respectively by an impurity atom and a vacancy. N_V^0 is the equilibrium-vacancy concentration at infinite dilution of A.

If Fermi-Dirac statistics are used and it is assumed that the donor and vacancy levels and the energy-band structure are not affected by the impurity concentrations, then γ_A and γ_V are the functions of [A] shown in *fig. 4* and *fig. 5*. It can be seen that with increasing donor concentration the diffusivity (eq. 7) increases because of the two factors γ_V^{-1} and $(1 + \partial \ln \gamma_A/\partial \ln [A])$. The first factor describes the Fermi-level effect: γ_V decreases with increasing donor concentration to values smaller than unity, indicating an increase in the concentration of charged vacancies. The second factor can be interpreted as describing the electric-field effect. The coefficient γ_A increases with increasing donor concentration to values larger than unity. The difference in chemical potential (μ_A) of the donor at two different distances is therefore larger than that calculated from the difference in concentration only. This leads to a greater driving force for the diffusion. It can be seen from the expression that this second factor has the greatest effect when the concentration of impurity atoms approaches that of the intrinsic carriers at the temperature of diffusion.

The relation between the two effects is more complicated than would appear from this simplified equation.

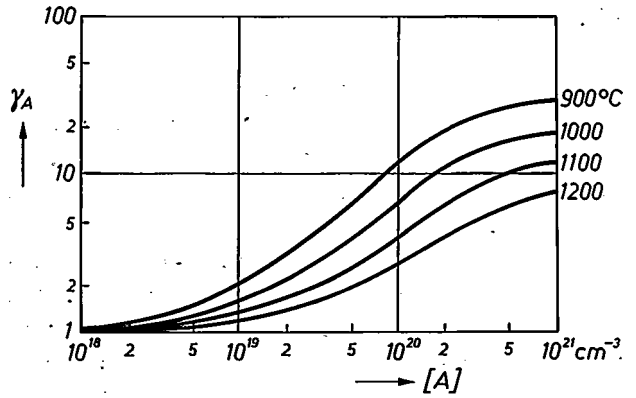


Fig. 4. The donor activity coefficient γ_A for silicon (S. M. Hu [30]), as a function of the donor concentration at four temperatures.

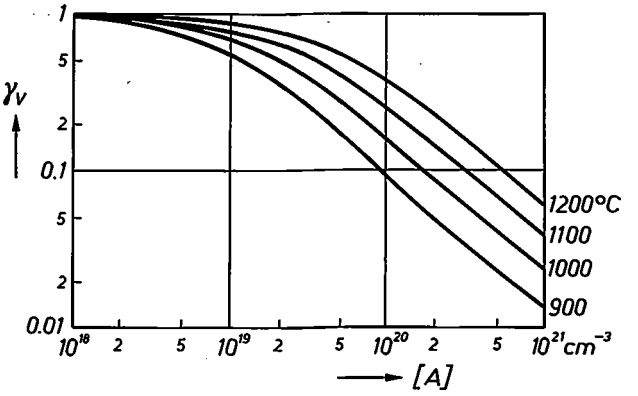


Fig. 5. As fig. 4, but now for the activity coefficient γ_V of a vacancy in silicon. (After Hu [30].)

Fig. 6 gives impurity profiles produced by diffusion from a constant surface concentration [A(0)], as described by eq. (7). The parameter ϱ, equal to $[A(0)]/2n_i$, is a measure of the extent to which the material is non-intrinsic (n_i again denotes the concentration of free electrons in pure silicon at the diffusion temperature). Profiles like these have been observed, and appear to describe the diffusion of As in Si quite well up to a concentration of about 3×10^{20} cm^{-3}. It can be seen that these profiles are not at all related to the one often assumed to be present and calculated on the assumption that these two effects can be neglected ($\varrho = 0$).

It will be clear that these large and concentration-dependent deviations from the 'normal' functions (complementary Gaussian error functions) introduce difficulties in calculating the exact concentrations and distances in *N-P-N* junctions.

The push-out effect

Another effect is often even more important in transistor manufacture in determining the actual structure of the transistor. This is the push-out or emitter-dip effect.

In the well known process for making *N-P-N* junction transistors boron, a *P*-type dopant, is first diffused into selected areas of an *N*-type epitaxial silicon layer. This produces the *P*-type base layer. Next, phosphorus is diffused into a smaller area of the *P*-type layer, to form the *N*-type emitter region. The push-out effect, schematically represented in *fig. 7*, is a local increase in the depth of the base layer (boron-doped) under the emitter layer. The effect is of great importance in device fabrication because the electrical characteristics of the device, such as its high-frequency limit, depend upon the distance between the *N*-type phosphorus-diffusion layer (the emitter) and the original *N*-type layer (the collector) beneath it. This distance (the base width), which is proportional to $c - a$ in fig. 7, can be greatly affected by the magnitude of the push-out. In an analysis of this effect D. B. Lee [32] has made the following observations:

— The phosphorus-diffusion profile has a distinct 'tail', as indicated in *fig. 8*. This tail and the associated kink in the profile become smaller as the diffusion temperature increases, and have almost vanished at 1200 °C.

— The push-out of the base during the phosphorus

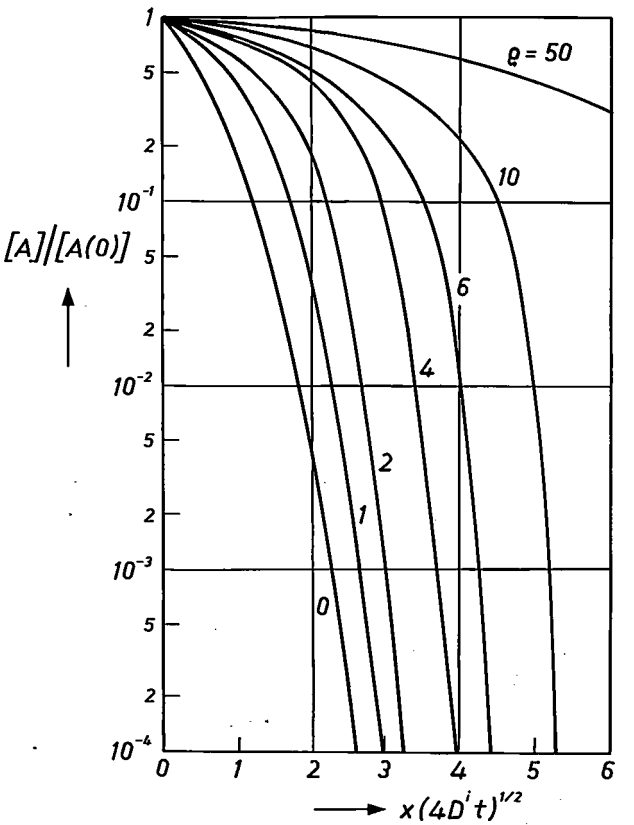


Fig. 6. Calculated concentration profiles for the diffusion of a donor A in silicon. The diffusion source is assumed to be a constant surface concentration [A(0)]; the apparent diffusivity is given by eq. (7) in the text. The parameter ϱ indicates the extent to which the material is non-intrinsic; $\varrho = [A(0)]/2n_i$. (After Hu and S. Schmidt [31].)

diffusion at 950 °C is characterized by a local enhancement of the boron diffusivity by a factor of about 20, provided that the concentration of phosphorus atoms near the surface is greater than 9.5×10^{20} atoms cm^{-3}. This enhancement factor can be much higher (about 700× max.) at lower temperatures.

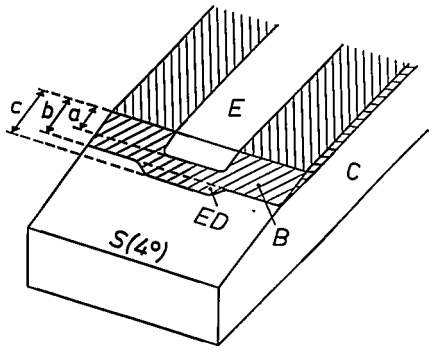


Fig. 7. Diagram explaining the push-out effect in a diffused-junction transistor of *N-P-N* type. Beneath the emitter region *E* the base layer *B* is more deeply diffused into the collector region *C*; the magnitude of the dip *ED* affects the distance between the emitter and collector, and hence also the important electrical characteristics of the transistor. The depth of the emitter is proportional to *a*, and the original depth of the base to *b*. The increased depth of the base, due to the push-out effect, is proportional to *c*. These dimensions are shown here in a plane *S* at an angle of 4° to the upper surface of the transistor.

— The phosphorus atoms at the lower edge of the emitter, where the boron-doped base begins, i.e. in the tail of the phosphorus profile, give a similar enhancement in their diffusivity.

Push-out also occurs in *P-N-P* junctions with emitters containing boron. Both P and B atoms are smaller than Si atoms. The lattice constant of a silicon lattice containing large concentrations of these impurities is in fact reduced and extra planes — dislocations — therefore have to be formed. The front of this network of dislocations moves into the crystal and coincides approximately with the front of the diffusion profile. The density of the excess vacancies generated by the part of the movement of the dislocations that can be described as climb is a function of time and of the distance from the surface. The excess vacancies will be annihilated at the surface and by the formation of vacancy clusters. In this way a high non-equilibrium concentration of vacancies is formed, and since vacancies move easily, this high concentration can extend far beyond the limits of the diffusing emitter front, into the base and beyond the base-collector transition. The phosphorus- and vacancy-concentration profiles will therefore in general be of the form shown in *fig. 9*.

The presence of the kink in the phosphorus profiles can be explained from this figure. Near the surface (a vacancy sink) the excess concentration of vacancies is much less than deeper down, which means that the effective diffusivity of the phosphorus near the surface is lower.

T. J. Parker [33] and S. M. Hu and T. H. Yeh [34] have used these ideas in an attempt to calculate theoretically the diffusivity-enhancement factor; the result was an enhancement of about six times. This hypothesis does not therefore appear to explain the effect completely. R. F. Peart and R. C. Newman [35] and also M. Yoshida and associates [36] have drawn attention to the fact that diffusion enhancement can also occur without the generation of dislocations and can moreover be strongly influenced by processes occurring at or near the surface layer on which the phosphorus source has been applied [37]. This effect can be very important. The following possible explanation has been

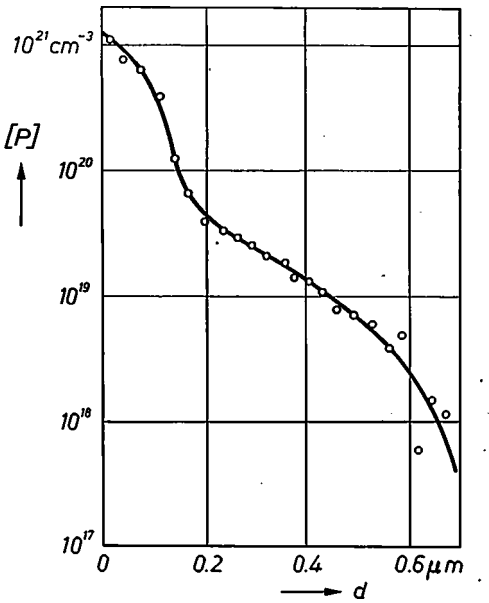


Fig. 8. Example of a 'tail' in a diffusion profile. At a depth *d* of about 0.15 μm there is a kink in the concentration profile, followed by the tail. In this case phosphorus marked with a radioactive tracer was diffused into the silicon. (After D. B. Lee [32].)

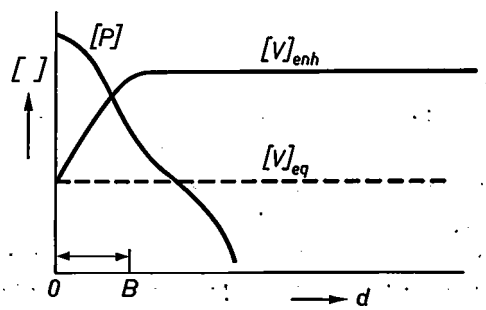


Fig. 9. The concentration profiles predicted by Lee [32] for phosphorus atoms [P] and vacancies $[V]_{enh}$ in silicon. The formation of the extra vacancies causes the push-out effect. Just below the silicon surface there is a region (*OB*) where the increased production of vacancies takes place. For simplicity it has been assumed that the production of vacancies in this region is evenly distributed. $[V]_{eq}$ is the equilibrium concentration for the vacancies in intrinsic silicon. *d* depth in the silicon.

put forward: the phosphorus diffuses only by means of the phosphorus-vacancy pairs (the *E* centres). Therefore when the phosphorus atoms enter from a surface into the bulk material they must do so as *E* centres. The number of *E* centres formed per unit time at the surface depends on the surface conditions. The *E* centres next flow into the bulk. Shortly afterwards, not far from the surface, the *E* centres dissociate into single phosphorus atoms and excess vacancies. This amounts to another mechanism for the generation of excess vacancies, which could contribute to an enhancement of the diffusivity.

Other secondary effects

In addition to the effects mentioned above, there are a number of others that affect the concentrations of the diffusants and their profiles. These include the formation of certain immobile molecules such as $SiAs_4$ [1] or of precipitates such as SiP [38], or the deviation of the direction of the normal of the actual surface plane from that of the normal of the (100) or (110) planes [39]. However, interesting as these are, they will not be dealt with here.

By now it will be clear that although the diffusion of group-III or group-V elements into silicon appears at first sight to be one of the simplest cases of diffusion, although the process is fundamental to the manufacture of many important electronic devices, and although there has been a positive flood of publications about it, the process is not yet understood. There is still uncertainty about the primary diffusion process, while it is almost impossible to estimate the various relative contributions from the secondary processes. The actual fabrication of the complicated *P-N* junctions in integrated circuits is still largely a matter of industrial empiricism. On the other hand it is fair to say that, without the help of scientific investigation and understanding, development of industrial know-how alone could not have succeeded in creating the variety of diodes, transistors and integrated circuits that exists today.

Sintering

General aspects

The three stages in sintering; the various atomic transport routes

Another process involving atomic transport is sintering, a process that is used extensively in industry. The complex geometry of a compacted powder during sintering can be dealt with in a simplified way by distinguishing between three stages. In the initial stage bonds grow between the particles, so as to form 'necks' (*fig. 10*). The intermediate stage starts when the necks approach each other. As sintering proceeds the pore structure gradually changes into a network of interconnected channels, with these channel-like pores occurring primarily at the places where three grain edges meet, or at the corners of the grains [41]. *Fig. 11* gives a schematic diagram of the geometry for the initial and the intermediate stages of sintering. In the initial stage of sintering the linear shrinkage amounts to about three to five per cent. In the intermediate stage the porosity falls to about 5%. In this intermediate stage grain growth gradually starts to become significant. In the third stage the porosity is discontinuous; the pores are separated. The removal of these completely disconnected pores, together with grain-boundary migration, resulting in grain growth, will be discussed later in this article.

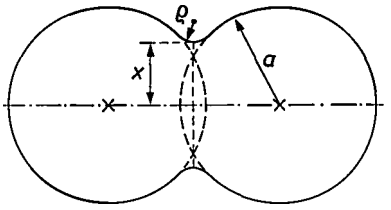


Fig. 10. In the initial stage of the sintering process neighbouring particles are bonded to one another through the formation of 'necks' (A. L. Stuijts [40]). The neck shown here has a radius of x, and the two particles are assumed to be spherical and of radius a. The surface of the neck has a radius of curvature ϱ.

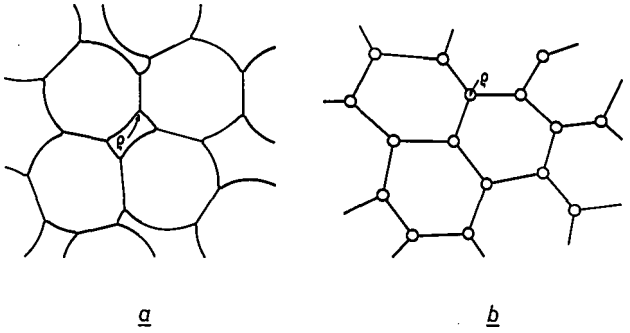


Fig. 11. A diagramatic representation of the situation during the initial stage (*a*) of the sintering process and during the intermediate stage (*b*). The small circles in (*b*) are pores (in cross-section) located between the boundaries of three grains. (After D. L. Johnson [41].)

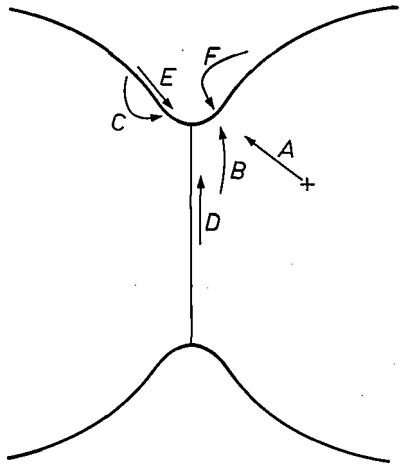


Fig. 12. Model for the geometry and the diffusion flux paths in the initial stage of sintering. *A*, *B* and *C* volume diffusion. *D* grain-boundary diffusion. *E* surface diffusion. *F* vapour transport. (After Johnson [41].)

The driving force in the first two stages of the process is the excess surface energy of a compacted powder compared with that of a dense material. This sets in motion a material-transport mechanism controlled by diffusion. Several transport routes for the diffusion are shown in *fig. 12*.

Mathematical descriptions of the first sintering stage

W. D. Kingery and M. Berg [42] have derived the expression

$$\Delta L/L_0 = C(\gamma a^3 D)^{2/5} r^{-6/5} t^{2/5} \qquad (10)$$

for linear shrinkage by considering only the volume diffusion (*A* and *B* from fig. 12). Here γ is the surface energy, a the atomic dimension, D the self-diffusivity, r the particle radius and t the diffusion or sintering time. The coefficient C is equal to $(20/\sqrt{2}\,kT)^{2/5}$.

Fig. 13 shows that the time dependence of the densification of Al_2O_3, in the first sintering stage, can in fact be described reasonably well by the expression (10).

As we have seen, other possible transport routes may also have to be considered, however. For instance, it is almost certain that the grain-boundary diffusion is important. The different kinds of transport can be divided in two classes: the fluxes *A*, *B* and *D* in fig. 12 produce broadening of the neck and shrinkage, whereas the fluxes *C*, *E* and *F*, all connected with the surface, although helping to broaden the neck, do not in the main produce shrinkage. It is generally accepted that routes *B* and *D* are most effective for shrinkage.

D. L. Johnson [41] [44–46] has developed a theory for the first stage of sintering, which is also valid, with certain assumptions, for the second stage. He took into account the fluxes *B*, *D*, *E* and, to a certain extent, *F*. The solid line in *fig. 14* shows his theoretical calculations as applied to experimental results obtained by sintering spherical iron particles with a radius of

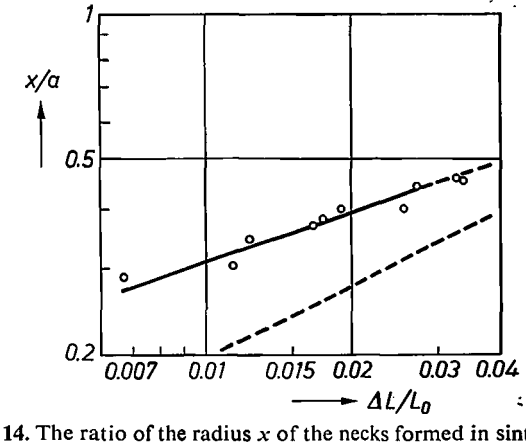


Fig. 14. The ratio of the radius x of the necks formed in sintering (fig. 10) to the radius a of the particles, as a function of the shrinkage. In this experiment the starting material consisted of iron particles of radius 12.3 μm, which were sintered at a temperature of 875 °C in an atmosphere of hydrogen. The solid line represents the result of calculations in which volume diffusion (*B* in fig. 12), grain-boundary diffusion (*D*) and surface diffusion (*E*) were taken into account; these appear to give good agreement with experimental results (small circles). The dashed line indicates the results calculated when material transport from the rounded surfaces (*E* and possibly *F*, fig. 12) was taken to be negligible. (After Johnson [45].)

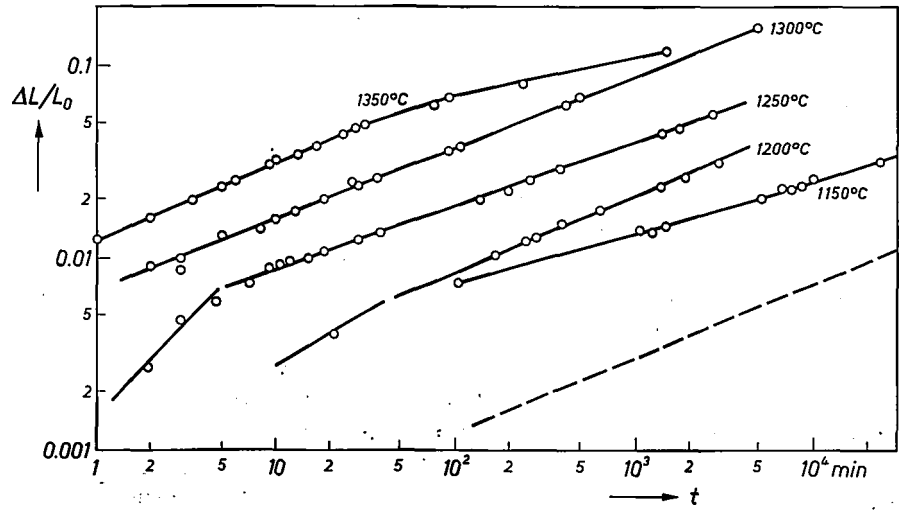


Fig. 13. The linear shrinkage $\Delta L/L_0$ measured by R. L. Coble [43] at five temperatures when sintering compacted Al_2O_3 powder, as a function of time in the initial stage of the process. The dashed line indicates the slope of a line $\Delta L/L_0 \propto t^{2/5}$ (see eq. 10).

12 μm [47]. It can be seen that these experiments can only be explained by assuming that surface diffusion is also present. The values for the three diffusivities (volume diffusion B, grain-boundary diffusion D and surface diffusion E) that Johnson calculated from P. H. Shingu's sintering data agree well with values derived from other types of experiment.

Volume diffusion in ionic compounds

In compounds, and particularly in ionic compounds, the situation concerning the flux due to volume diffusion becomes more complicated. Because of the condition of electroneutrality, the movements of the cations and the anions are related. In a binary compound $M^{2+}X^{2-}$ with Schottky disorder:

$$[V_M''] \times [V_X^{\cdot\cdot}] = K_S, \qquad (11)$$

where $[V_M'']$ and $[V_X^{\cdot\cdot}]$ denote the concentrations of the (effectively doubly charged) cation vacancies and anion vacancies respectively.

The movements for both cations and anions take place by means of a vacancy mechanism. The effective volume diffusivity $\tilde{D}$ can be written as:

$$\tilde{D} = \frac{[V_M'']D_M[V_X^{\cdot\cdot}]D_X}{[V_M'']D_M + [V_X^{\cdot\cdot}]D_X} = \frac{D_M D_X K_S}{[V_M'']D_M + [V_X^{\cdot\cdot}]D_X}. \qquad (12)$$

It has a maximum at

$$[V_M'']D_M = [V_X^{\cdot\cdot}]D_X, \qquad (13)$$

where D_M and D_X denote the diffusivities of cations and anions respectively. As D. W. Readey [48] [49] and P. J. L. Reijnen [50] [51] have pointed out, to maximize the fluxes the concentration of the more slowly diffusing vacancy has to be increased by a factor inversely proportional to the ratio of the two diffusivities. This can be achieved by adding a suitable dopant.

The third and final stage of sintering

As mentioned before, the third and final stage of sintering starts when the pores are separated and lie in a matrix of grains with intersecting grain boundaries [*]. Grain-boundary migration, with the associated grain growth, will then be one of the major processes by which the sintered compact attempts to achieve minimum energy. The force producing the migration of a given boundary is directly related to its curvature, and the direction of this force is towards the centre of that curvature. The disappearance and growth of grains by grain-boundary migration are well illustrated by a two-dimensional representation given by J. E. Burke [53] (*fig. 15*). This shows that the boundary curves are such that boundary migrations tend to favour the removal of grains with less than six sides, whereas grains with more than six sides are favoured for growth. Large many-sided grains, once formed, tend to grow very rapidly, giving 'exaggerated' grain growth. It may then happen that the grain boundary of the growing grain moves faster than the pores it encounters, with the result that the pores get trapped inside the growing grain, making a further densification virtually impossible.

Impurities, however, can sometimes hinder boundary migration and thus prevent exaggerated grain growth and pore inclusion. In this way very low porosities in sintered material can be obtained.

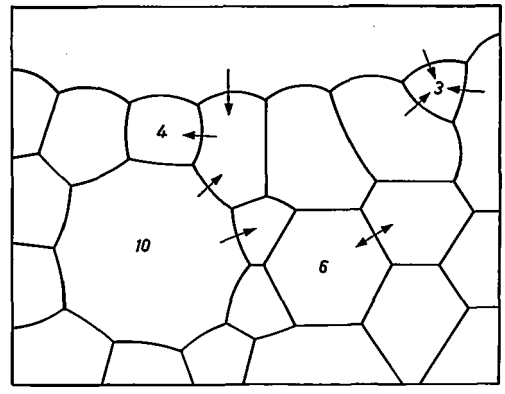


Fig. 15. Idealized grain structure in a ceramic material. The grain boundaries tend to migrate towards their centres of curvature. The numbers indicate the number of sides of the grain. (After J. E. Burke [53].)

Some industrial sintering processes and their scientific interpretation

The manufacture of dense alumina

Dense alumina is an envelope material for high-temperature high-pressure sodium lamps. These lamps emit light with a much broader and therefore more acceptable spectrum than conventional sodium lamps. Because the temperature of the sodium vapour is very high, glass cannot be used. Since it is strong and chemically stable, Al_2O_3 seems to be a good choice for a new type of lamp envelope, provided it can be made translucent or, possibly better, transparent. It appeared likely that Al_2O_3 sintered to a very high density would have these properties. R. L. Coble [54] found that sintered Al_2O_3 does indeed become translucent if 0.1 to 0.2% by weight of MgO has been added.

Explanations of this vary. Coble assumed that there was a second phase, $MgAl_2O_4$-spinel, to pin the boundaries, whereas P. J. Jorgensen [55] has suggested that the segregation of MgO at the boundaries decreases their mobility. Others [56] think that both situations can arise. Recently [57] [58] however, it has been shown that exaggerated grain growth can also be prevented by

[*] In this section use is made of an excellent review article by R. J. Charles [52].

using lower concentrations of MgO, below the maximum solubility — thus avoiding segregation. A possible explanation might be that an enhanced concentration of these impurities near the grain boundaries diminishes the excess energy of these boundaries so that the force causing their movement is also diminished. The presence of foreign atoms near the grain boundary might also cause a decrease in the mobility of the boundary. Whatever the exact mechanism, the presence of MgO appears in any case to prevent fast movements of grain boundaries and hence the inclusion of pores inside the grains.

After sintering for about 1.5 hours at 1630 °C most of the excess MgO (and possibly also a second phase consisting of MgO or $MgAl_2O_4$) has to be removed by firing. This is done by further heating for about 8 hours at 1850 °C, at very low oxygen pressures. Most of the MgO evaporates. The grain diameter continues to increase from 5 μm to about 25 μm without the occurrence of exaggerated grain growth, because some of the MgO remains in the Al_2O_3 as a solute. At least as important, however, is the fact that during this stage of the process most of the remaining pores disappear whereas the rest coalesce in the form of a small number of large pores. This is exactly what is required to produce a translucent material. For a given overall but small porosity, a relatively small concentration of large pores gives a much more translucent material than a larger concentration of smaller pores [59] [60]. Not all additives have the same kind of effect on the grain growth. The presence, for instance, of 200 ppm of Na or 40 ppm of Ca, instead of preventing exaggerated growth, enhances it [61].

Another interesting effect of MgO is that it increases the sintering rate. The addition of other divalent ions, such as Fe^{2+}, increases both the sintering rate and the creep rate. The same effects are observed on the addition of tetravalent Ti^{4+} ions [62–65]. These findings can be explained only if the diffusivity of the Al^{3+} ions is the factor that determines the material flux, and the disorder of the Al^{3+} ions is of the Frenkel type, giving rise to the existence of aluminium vacancies (V_{Al}''') and interstitial Al ions (Al_i'''). The concentration of the vacancies then will be enhanced by the addition of tetravalent ions such as Ti^{4+}, and that of the interstitials will be enhanced by the addition of divalent ions such as Mg^{2+} or Fe^{2+}. As the theory of Readey and Reijnen predicts [48–51] both effects will lead to an enhanced diffusivity of aluminium.

In accordance with this explanation it has been found that in polycrystalline Al_2O_3 with a grain diameter below 30 μm the volume diffusion of Al did indeed determine the rate, since the diffusion of O^{2-} is much faster and takes place along the grain boundaries [66].

Sintered Y_2O_3 as a ceramic material for lasers

Another example of the prevention of exaggerated grain growth is the work by C. Greskovich [67] [68] on Y_2O_3 doped with 10% by weight of ThO_2. He was able to obtain residual porosities as low as 0.1 to 1 ppm by volume. As a consequence this material, when doped with 1% Nd [69], could be used to make a highly efficient polycrystalline ceramic laser.

However, here the addition of ThO_2, which prevents exaggerated grain growth, does not give an increase in the sintering rate, unlike the addition of MgO or TiO_2 to Al_2O_3. In fact, when ThO_2 is added to Y_2O_3 it takes no less than 100 hours of sintering at 2200 °C to obtain the desired result. According to the theory of Readey and Reijnen [48–51] this could easily be explained by assuming a vacancy mechanism for the volume diffusion of Y and O. The concentration of the oxygen vacancies, which move slowly and determine the sintering rate, would be decreased by the addition of Th^{4+} ions. However, it can be inferred from the example of Al_2O_3 that simple explanations should be regarded with caution.

Mn-Zn ferrites as core material in transformers and LC filters and for magnetic recording heads

The subject of sintering will be brought to a close by a brief discussion of a ceramic material with magnetic properties. Of the ferrites with spinel structure, the Mn-Zn ferrites containing some Fe^{2+} have the smallest magnetic anisotropy and the smallest magnetostriction. These Mn-Zn-Fe^{2+} ferrites therefore have potentially the highest magnetic permeability [70]. A very high permeability, which is desirable for transformer applications, can indeed be attained in these ceramic materials if the grains are sufficiently large and do not contain pores [71–74]. In the fabrication process there are three separate stages — not to be confused with the three stages in sintering discussed earlier. In the first stage, sintering takes place at high temperature and high partial oxygen pressure. The high temperature is necessary to obtain large grains. The high oxygen pressure ensures a relatively high concentration of cation vacancies and a low concentration of oxygen vacancies. According to Reijnen [75] a pore then may achieve a velocity that is at least as high as the velocity of the movement of the grain boundaries. In this stage the pores will therefore easily move along with the grain boundaries and then coalesce between the grains. This can only happen if material can be transported quickly enough from one side of the pore to the other. *Fig. 16* shows the mechanism in detail. Because of the different curvatures of the surface of the pore — which are produced in minimizing the surface energies of the grains — vacancy gradients arise, which form the driving

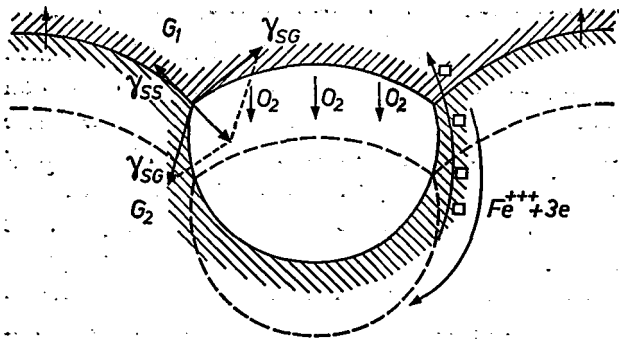


Fig. 16. The mechanism for material transport from one side of a pore to the other, as when sintering ferrites with a relatively high concentration of cation vacancies in an oxidizing atmosphere. The drawing represents a small detail of a grain boundary whose overall curvature is such that the boundary moves upwards (see fig. 15). The dashed lines represent the starting situation. The small squares denote cation vacancies. The oxygen is transported as a gas (O_2). The movement of the cations (Fe^{3+}) is equivalent to a transport of cation vacancies in the opposite direction. Electrons ($3e$) are transported at the same time as the cations, so that the net charge transport is zero. $G_{1,2}$ crystallites. At the intersection of the boundary surface between G_1 and G_2 with the pore the surface tensions γ_{SS} and γ_{SG} appear. SS solid-phase/solid-phase boundary surface, SG solid-phase/gas-phase boundary surface. The pore assumes a shape such that the three surface tensions cancel each other out. (After P. J. L. Reijnen [50].)

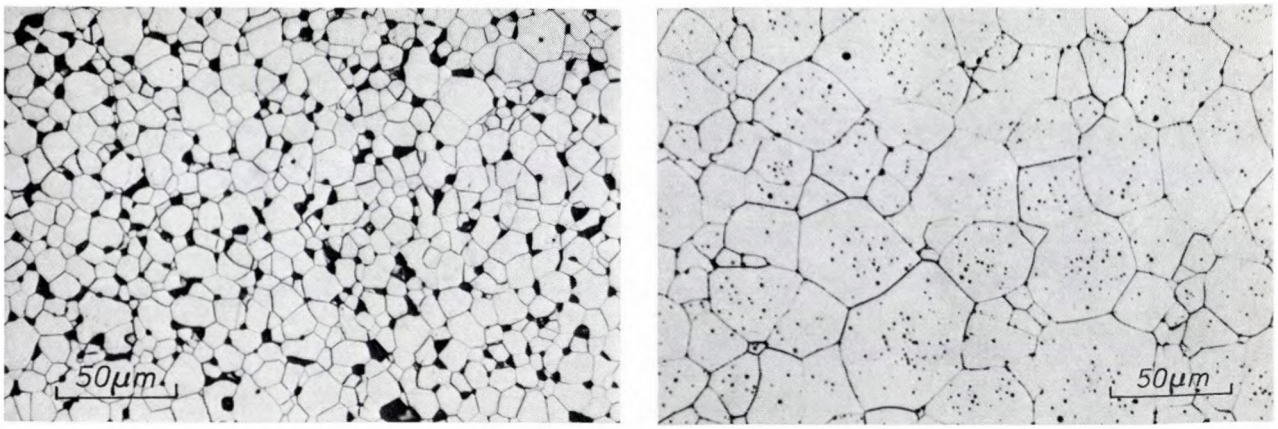
force for the transport of material. The oxygen can be quickly transported via the pore, as a molecular gas. The cations (Fe^{3+}) also move quickly via the solid phase as a result of the cation vacancies present. Electrons appear and disappear continually by the formation of oxygen from oxygen ions and vice versa. The cation transport does not create a charge across the pore, because there is an equivalent transport of electrons, which can take place via repeated electron jumps from an Fe^{2+} ion to a neighbouring Fe^{3+} ion. The movement of pores in association with grain boundaries [76], whatever the mechanism, leads to coalescence of pores between the grains [77] and hence relatively large pores. The microstructure is characterized by pore-free crystal grains and a number of large pores at the intersection of three or four grain boundaries. When the desired average dimension of the grains has been reached in this way, the second stage begins.

During this second stage the temperature is lower. As a consequence the grain growth stops. The partial oxygen pressure is also lowered, to such a value that the composition $Mn^{2+}_{1-x-y}Zn^{2+}_{x}Fe^{2+}_{y}Fe^{3+}_{2}O_4$ is obtained, which means that some of the Fe^{3+} ions are reduced to Fe^{2+} ions. The third stage is a period of cooling during which, to keep the compositions and valencies fixed, the oxygen pressure has to be changed continuously in a well defined pattern as a function of temperature and time. *Fig. 17* shows the difference between the microstructures attained by sintering with a cation deficit (*a*), as in the case just described, and with an anion deficit (*b*). In the first case (*a*) there are a small number of large pores *between* the crystallites, whereas in the second case (*b*) there are a large number of small pores *inside* the grains. Although, with the same sintering procedure, the crystallites are smaller and the total volume of pores larger in the cation-deficit case (*a*) than in the anion-deficit case (*b*), the cation-deficit case nevertheless gives higher permeabilities because the perfect crystallites in (*a*) are larger than the distances between the pores in (*b*). In case (*a*) the Bloch walls are pinned at a smaller number of anchoring points than in case (*b*). The Bloch walls can therefore move more easily (reversibly) in case (*a*).

By carefully choosing the sintering procedure and by using long sintering times it is also possible to obtain large crystallites and high densities in the cation-deficient case, permitting very high magnetic permeabilities to be obtained (*fig. 18*).

These Mn-Zn-Fe^{2+} ferrites with very high initial permeability, however, cannot be used as core material in *LC* filters in the band between 10 and 100 kHz. Because of the Fe^{2+} content these ferrites have some electrical conductivity, and this allows undesirable eddy-current losses to appear at higher frequencies. These losses can be suppressed by using small grains and surrounding each grain with a thin layer of insulating material. The material is therefore sintered at a relatively low temperature to ensure that the grains are small, and CaO and SiO_2 are added to the starting material in the correct proportions, thus causing an insulating layer to be formed around each grain during the second stage of fabrication. The nature and exact mechanism of formation of this insulating layer is not fully understood. The small size of the grains and the presence of the grain-boundary layer does in fact cause a decrease in the permeability μ, but the loss factor $\tan\delta$ is reduced so much more that a better figure of merit $(\tan\delta)/\mu$ is obtained [79–82].

Recently sintering processes have been developed that permit the use of Mn-Zn ferrites as a material for magnetic recording heads. In addition to high permeability, resistance to wear and dimensional tolerance of the air gap are also important in a recording head. For both it is highly desirable to have low porosity and small pores. The pores at the grain boundaries are particularly liable to allow crystallites to be lost because of the forces tending to pull them out. For video applications the gap length — i.e. in the direction of the tape — must be less than 1 μm so that pores of that size are already outside the tolerance for the gap length. The pores should therefore be considerably smaller than 1 μm and the total porosity should be less than 0.1 %. Recently Mn-Zn ferrites have been considered for these applications, since they have higher values of saturation magnetization and permeability and a smaller de-



a *b*

Fig. 17. Photomicrographs (G. H. Jonker and Stuijts [78]) of the etched surface of a ferrite sintered with a cation deficit (*a*) and a ferrite sintered with an anion deficit (*b*). In the second case the pores are small and are found inside the crystallites. In (*a*) the pores have been 'swept together'; they are therefore larger than in (*b*) and they all lie between the crystallites.

gradation of permeability in use than the dense Ni-Zn ferrites used so far. It might have been thought that anion-deficient sintering (fig. 17*b*) — which can after all give higher densities and smaller pores — would have been the way to get the desired specifications. This is indeed the method for manufacturing the Ni-Zn ferrites so far used for this purpose [83]. For the Mn-Zn ferrites, however, this was found not to be the right method. Although pretreatment of the starting material and changing the conditions of the sintering process brought the porosity down to 0.5-1 %, it has not as yet been possible to obtain porosities lower than 0.1 % by method (*b*) even after isostatic hot pressing. However, when material with a porosity of about 5 % (large pores at grain boundaries) obtained by method (*a*) is submitted to isostatic hot-pressing at temperatures low enough (1300 °C) to avoid exaggerated grain growth, small porosities, < 0.1 %, were indeed obtained [84] [85].

Fig. 19 shows such a microstructure. The pores at the grain boundaries have disappeared and the grains have

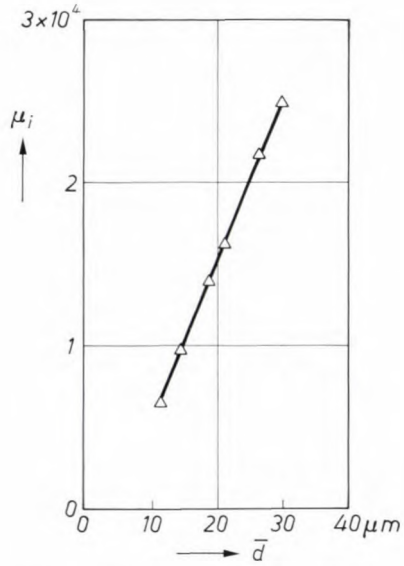
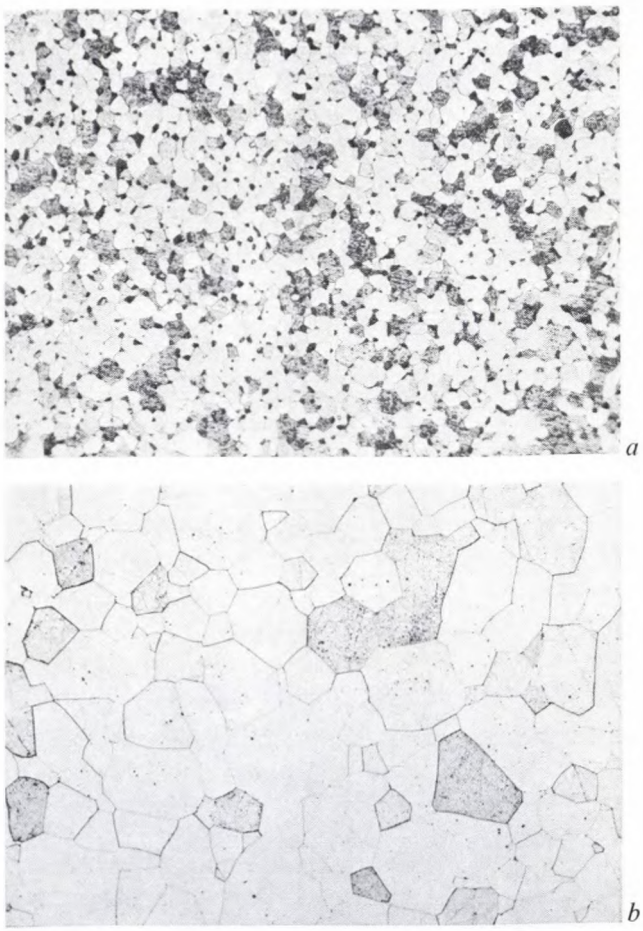
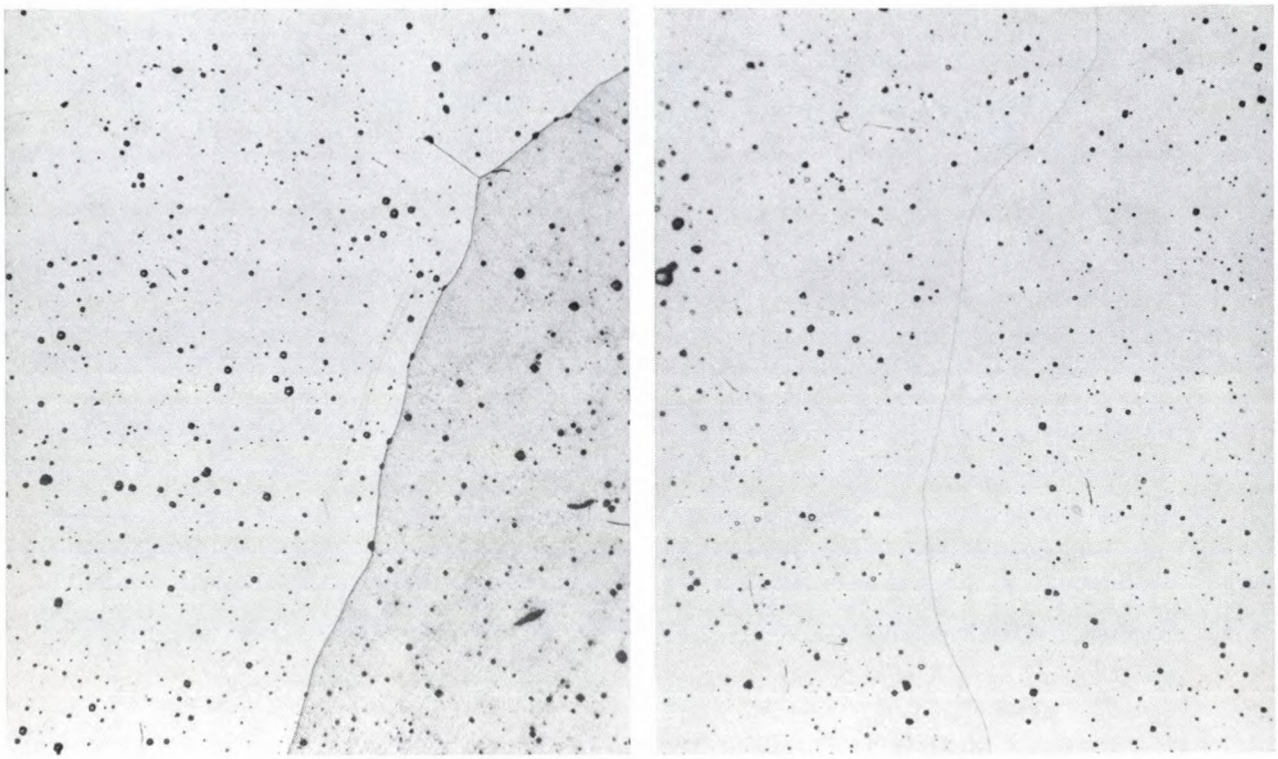


Fig. 18. The relative initial permeability μ_i of an Mn-Zn-Fe^{2+} ferrite, as a function of the average diameter $\bar{d}$ of the crystallites. (After D. J. Perduijn and H. P. Peloschek [74].)



a *b*

Fig. 19. The microstructure of a ferrite sintered with a cation deficit (see also fig. 17*a*). *a*) Before isostatic hot pressing. *b*) After isostatic hot pressing. The pores at the grain boundaries have disappeared and the grains have become larger. (After H. E. M. Stassen, J. G. M. de Lau and D. Veeneman [85].)



a *b*

Fig. 20. The microstructure of a ferrite sintered with an anion deficit (see also fig. 17*b*). *a*) Before isostatic hot pressing. *b*) After isostatic hot pressing. Only the pores in the neighbourhood of a grain boundary have disappeared. The pressing does not affect the pores inside the grains (which are large, up to 1 mm). (After Stassen, De Lau and Veeneman [85].)

become larger. Materials that have been pretreated by method (*b*), after a similar type of hot pressing, only gave a decrease in porosity from 1.4 to 1.2%. The individual grains (*fig. 20*) are very large (up to 1 mm) but, as can be seen, only the pores in the neighbourhood of the grain boundaries have disappeared: pores inside the grains do not appear to be affected by isostatic pressure, unlike those at the grain boundaries; this is in agreement with the theories of the sintering process.

By now it will have become clear that sintering is a very complicated process. It is greatly influenced by impurities, whether deliberately added or fortuitously present, and by the temperature of sintering and the composition and pressure of the ambient atmosphere as a function of time. It differs very much from one substance or composition to another. The progress and final result of the sintering is also greatly influenced by the pretreatment of the starting materials and by the way in which the powders are compacted. In view of all this it is perhaps rather surprising that valuable ideas and concepts have emerged at all. Nevertheless, the manufacturing processes for sintered materials, like those in the case of diffusion in silicon, are equally dependent on scientific understanding and on industrial empiricism.

Summary. This article is an expanded version of a lecture given in Vienna for the 1975 annual meeting of the Deutsche Bunsen-Gesellschaft; it reviews the diffusion and sintering processes in solids. The treatment in both cases is based on considerations relating to atomic transport, with emphasis on 'molecular engineering', which is now of great importance in the manufacture of materials. Various mechanisms for the diffusion of the group-III and group-V elements in intrinsic silicon are discussed. It is found that for well controlled experimental conditions the Fick diffusion laws are obeyed. During the manufacture of *P-N* junctions excessively high dopant concentrations occur, causing secondary diffusion effects (including the push-out effect), which are not as yet completely understood. The sintering process is also treated in its theoretical and experimental aspects. Examples of industrial importance are discussed in more detail: dense Al_2O_3, for lamp envelopes; dense Y_2O_3, for lasers; and the Mn-Zn ferrites, for specific applications — now made possible by appropriate manufacturing processes — in transformers, *LC* filters (telephony) and magnetic video-recording heads.

Literature

[1] S. M. Hu, in: D. Shaw (ed.), Atomic diffusion in semiconductors, Plenum Press, London 1973, p. 217.
[2] R. N. Ghoshtagore, Phys. Rev. B **3**, 389, 1971.
[3] I. R. Sanders and P. S. Dobson, Phil. Mag. **20**, 881, 1969.
[4] C. H. Lane, IEEE Trans. **ED-15**, 998, 1968.
[5] R. N. Ghoshtagore, Phys. Rev. Letters **25**, 856, 1970.
[6] R. N. Ghoshtagore, Phys. Rev. B **3**, 2507, 1971.
[7] D. L. Kendall and D. B. de Vries, in: R. R. Haberecht and E. L. Kern (ed.), Semiconductor silicon, Electrochem. Soc., New York 1969, p. 358.
[8] A. Seeger and K. P. Chik, Phys. Stat. sol. **29**, 455, 1968.
[9] S. M. Hu, J. appl. Phys. **45**, 1567, 1974.

[10] M. Yoshida and R. R. Hasiguti, in: Radiation damage and defects in semiconductors, Proc. Int. Conf., Reading 1972 (Institute of Physics Conf. Series No. 16), p. 223.
[11] R. A. Swalin, J. Phys. Chem. Solids **18**, 290, 1961.
[12] A. Scholz and A. Seeger, Phys. Stat. sol. **3**, 1480, 1963.
[13] A. Seeger and M. L. Swanson, in: R. R. Hasiguti (ed.), Lattice defects in semiconductors, Pennsylvania State University, University Park 1968, p. 93.
[14] K. H. Bennemann, Phys. Rev. **137**, A 1497, 1965.
[15] L. Elstner and W. Kamprath, Phys. Stat. sol. **22**, 541, 1967.
[16] B. J. Masters and J. M. Fairfield, Appl. Phys. Letters **8**, 280, 1966.
[17] J. M. Fairfield and B. J. Masters, J. appl. Phys. **38**, 3148, 1967.
[18] R. F. Peart, Phys. Stat. sol. **15**, K 119, 1966.
[19] R. N. Ghoshtagore, Phys. Rev. Letters **16**, 890, 1966.
[20] M. Yoshida and K. Saito, Jap. J. appl. Phys. **6**, 573, 1967, and **9**, 1217, 1970.
[21] P. Penning, Philips Res. Repts. **13**, 17, 1958.
[22] G. D. Watkins, J. Phys. Soc. Japan **18**, Suppl. 2, 22, 1963.
[23] G. D. Watkins and J. W. Corbett, Phys. Rev. **134**, A 1359, 1964.
[24] J. W. Corbett and G. D. Watkins, Phys. Rev. **138**, A 555, 1965.
[25] H. C. Casey Jr. and G. L. Pearson, in: J. H. Crawford Jr. and L. M. Slifkin (ed.), Point defects in solids, Vol. 2, Plenum Press, London 1973.
[26] R. N. Ghoshtagore, Phys. Rev. B **3**, 397, 1971.
[27] J. A. Van Vechten, Phys. Rev. B **10**, 1482, 1974.
[28] E. L. Elkin and G. D. Watkins, Phys. Rev. **174**, 881, 1968.
[29] M. Hirata, M. Hirata and H. Saito, J. Phys. Soc. Japan **27**, 405, 1969.
[30] S. M. Hu, Phys. Rev. **180**, 773, 1969.
[31] S. M. Hu and S. Schmidt, J. appl. Phys. **39**, 4272, 1968.
[32] D. B. Lee, Philips Res. Repts. Suppl. 1974, No. 5.
[33] T. J. Parker, in: J. N. Sherwood, A. V. Chadwick, W. M. Muir and F. L. Swinton (ed.), Diffusion processes, Gordon and Breach, London 1971, Vol. 2, p. 503, 1972.
[34] S. M. Hu and T. H. Yeh, J. appl. Phys. **40**, 4615, 1969.
[35] R. F. Peart and R. C. Newman, The role of defects in diffusion, in: Radiation damage and defects in semiconductors, Proc. Int. Conf., Reading 1972 (Institute of Physics Conf. Series No. 16), pp. 170-181.
[36] M. Yoshida, E. Arai, H. Nakamura and Y. Terunuma, J. appl. Phys. **45**, 1498, 1974.
[37] F. N. Schwettmann and D. L. Kendall, Appl. Phys. Letters **21**, 2, 1972.
[38] E. Kooi, J. Electrochem. Soc. **111**, 1383, 1964.
[39] S. M. Hu, J. appl. Phys. **45**, 1567, 1974.
[40] A. L. Stuijts, Synthesis of materials from powders by sintering, Ann. Rev. Mat. Sci. **3**, 363-395, 1973.
[41] D. L. Johnson, J. Amer. Ceram. Soc. **53**, 574, 1970.
[42] W. D. Kingery and M. Berg, J. appl. Phys. **26**, 1205, 1955.
[43] R. L. Coble, J. Amer. Ceram. Soc. **41**, 55, 1958.
[44] D. L. Johnson, in: H. H. Hausner (ed.), Modern developments in powder metallurgy, Vol. 4 (Processes), Plenum Press, New York 1971, p. 189.
[45] D. L. Johnson, J. appl. Phys. **40**, 192, 1969.
[46] D. L. Johnson, J. Amer. Ceram. Soc. **52**, 562, 1969.
[47] P. H. Shingu, Ph.D. dissertation, Northwestern University, 1967.
[48] D. W. Readey, J. appl. Phys. **37**, 2309, 1966.
[49] D. W. Readey, J. Amer. Ceram. Soc. **49**, 366, 1966.
[50] P. J. L. Reijnen, in: A. Rabenau (ed.), Problems of non-stoichiometry, North-Holland Publ. Co., Amsterdam 1970, p. 219.
[51] P. J. L. Reijnen, in: J. W. Mitchell, R. C. DeVries, R. W. Roberts and P. Cannon (ed.), Reactivity of solids, Proc. 6th Int. Symp., Wiley-Interscience, New York 1969, p. 99.
[52] R. J. Charles, Sintering and its role in material processing, in: Distinguished lectures in materials science, Univ. New Mexico, New Mexico State, and New Mexico Institute of Technology, April 1973.
[53] J. E. Burke, in: R. M. Fulrath and J. A. Pask (ed.), Ceramic microstructures, Wiley, New York 1968, p. 681.
[54] R. L. Coble, J. appl. Phys. **32**, 793, 1961.
[55] P. J. Jorgensen, J. Amer. Ceram. Soc. **48**, 207, 1965.
[56] N. A. Haroun and D. W. Budworth, Trans. Brit. Ceram. Soc. **69**, 73, 1970.
[57] H. L. Marcus and M. E. Fine, J. Amer. Ceram. Soc. **55**, 568, 1972.
[58] J. G. J. Peelen, to be published shortly.
[59] J. G. J. Peelen, Science of Ceramics **6**, XVII, 1973.
[60] J. G. J. Peelen and R. Metselaar, J. appl. Phys. **45**, 216, 1974.
[61] J. G. J. Peelen, private communication.
[62] R. J. Brook, J. Yee and F. A. Kröger, J. Amer. Ceram. Soc. **54**, 444, 1971.
[63] R. J. Brook, J. Amer. Ceram. Soc. **55**, 114, 1972.
[64] W. Raja Rao and I. B. Cutler, J. Amer. Ceram. Soc. **56**, 588, 1973.
[65] G. W. Hollenberg and R. S. Gordon, J. Amer. Ceram. Soc. **56**, 140, 1973.
[66] A. E. Paladino and R. L. Coble, J. Amer. Ceram. Soc. **46**, 133, 1963.
[67] C. Greskovich, Oxide ceramic laser, A.R.P.A. contract N00014-70-C-0360, Annual Technical Report, July 1972.
[68] C. Greskovich and K. N. Woods, Amer. Ceram. Soc. Bull. **52**, 473, 1973.
[69] C. Greskovich and J. P. Chernoch, J. appl. Phys. **44**, 4599, 1973.
[70] A. Broese van Groenou, P. F. Bongers and A. L. Stuijts, Mat. Sci. Engng. **3**, 317, 1968/1969.
[71] E. Röss, I. Hanke and E. Moser, Z. angew. Physik **17**, 504, 1964.
[72] E. Röss and I. Hanke, Z. angew. Physik **29**, 225, 1970.
[73] E. Röss, in: Ferrites, Proc. Int. Conf., Kyoto 1970, p. 203.
[74] D. J. Perduijn and H. P. Peloschek, Proc. Brit. Ceram. Soc. **10**, 263, 1968.
[75] P. J. L. Reijnen, Science of Ceramics **4**, 169, 1968.
[76] M. V. Speight and G. W. Greenwood, Phil. Mag. **9**, 683, 1964.
[77] W. D. Kingery and B. François, J. Amer. Ceram. Soc. **48**, 546, 1965.
[78] G. H. Jonker and A. L. Stuijts, Philips tech. Rev. **32**, 79, 1971.
[79] C. Guillaud, M. Paulus and R. Vautier, C.R. Acad. Sci. **242**, 2712, 1956.
[80] T. Akashi, NEC Res. Devel. **8**, 89, 1966.
[81] T. Akashi, NEC Res. Devel. **19**, 66, 1970.
[82] A. Hamelin and M. Paulus, in: Ferrites, Proc. Int. Conf., Kyoto 1970, p. 134.
[83] A. L. Stuijts, Proc. Brit. Ceram. Soc. **2**, 73, 1964.
[84] K. H. Härdtl, Amer. Ceram. Soc. Bull., to be published shortly.
[85] H. E. M. Stassen, J. G. M. de Lau and D. Veeneman, Ber. Dtsch. Keram. Ges., to be published shortly.

A simple and flexible automatic extractor

The increasing concern for the environment now being displayed throughout the industrialized world has led to an increasing need for analytical equipment that operates automatically [1]. Experience in the development of this equipment shows that the extraction stage so often required is difficult to automate. Extraction is desirable, for instance, for the determination of heavy metals and pesticides in surface water, since they occur in very low concentrations (0.1-100 ng/ml) in the presence of varying and often unknown other constituents. Extraction not only permits separation of the substances to be analysed from others that would interfere with the measurement; it also enables their concentrations to be increased to such an extent that they can be measured reliably with simple equipment.

We have now developed an extractor that not only can be used for automatic determination of heavy metals and pesticides but also offers an elegant solution to numerous other extraction problems because it satisfies a combination of requirements: extractions are effected quickly and with great reproducibility and the apparatus is extremely reliable. It is also very simple and flexible and will handle large or small quantities of liquid equally well.

The possibility of extracting a substance from a solution depends on the difference in the solubility of the substance in two immiscible liquids, one of which is usually polar (e.g. water) and the other nonpolar (e.g. an organic solvent such as carbon tetrachloride). The two liquids are first brought into intimate contact with each other — e.g. by shaking in a separating funnel — and then separated again. The efficiency of an extraction process, i.e. the ratio of the amount of the substance extracted to the original total quantity, depends on the partition coefficient and the ratio of the volumes of the two liquids. The extraction efficiency can therefore always be increased by changing the volume ratio, but this can only be done at the expense of the attainable concentration. Chemical methods (choice of pH, formation of complexes, etc.) are therefore often used to make the partition coefficient as high as possible. If a single extraction is nevertheless insufficient, the extraction process can be repeated a number of times until a total efficiency of greater than 99% is attained. The extraction procedure can be conducted either in steps or continuously by two different methods: either by the cross-current principle or by the counter-current principle [2], as illustrated in *fig. 1*.

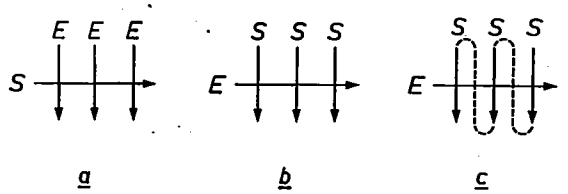


Fig. 1. Schematic diagram of extraction employing the cross-current principle (*a* and *b*) and the counter-current principle (*c*). *S* liquid to be extracted. *E* extraction medium.

In cross-current extraction one of the two liquids brought into contact with each other is always in its *original* concentration. If the intention is to remove from a solution as much as possible of a substance dissolved in it, a supply of 'fresh' extraction medium has to be maintained (*a*). A continuous fresh supply of the solution to be extracted is required if the intention is to obtain the highest possible concentration of the dissolved substance in the extraction medium (*b*).

In counter-current extraction the concentrations of *both* liquids always have a different value at each new contact, with one liquid in its most 'refined' state and the other in its least refined state. Counter-current extraction can also be used for both purification and concentration.

The work on the development of an automatic extractor at Philips Research Laboratories was originally concentrated mainly on the *reproducibility* of the extraction process. Our starting point was that the transport of the substance to be extracted depends on the contact surface area between the two liquids, their contact time and the rate of transport. The extractor is based on an extremely simple scheme for controlling these three parameters.

In the new extractor, which is shown schematically in *fig. 2*, the two liquids (e.g. water and carbon tetrachloride) are fed into a narrow tube, at the same rates of flow, via a Y-piece, made typically of PTFE. Depending on the surface tensions of the liquids both in contact with each other and against the wall of the tube, there is a narrow range of flow ratios within which the size of the drops formed is such that they completely seal off a tube of a given diameter. If this condition is satisfied, and provided also that the Y-piece is of a suitable shape, a regular and reproducible pattern of drops of the polar liquid (water) alternating with 'columns' of the non-polar liquid (carbon tetrachloride) is formed in the tube.

In this extraction method, unlike those generally used previously, the drops cannot move freely. In particular, the fact that the size of the contact surface is fixed by the tube diameter plays an important part in ensuring the reproducibility of the material transport. Also, the fact that the drops are forced along a tube wall has another advantage: the friction between the

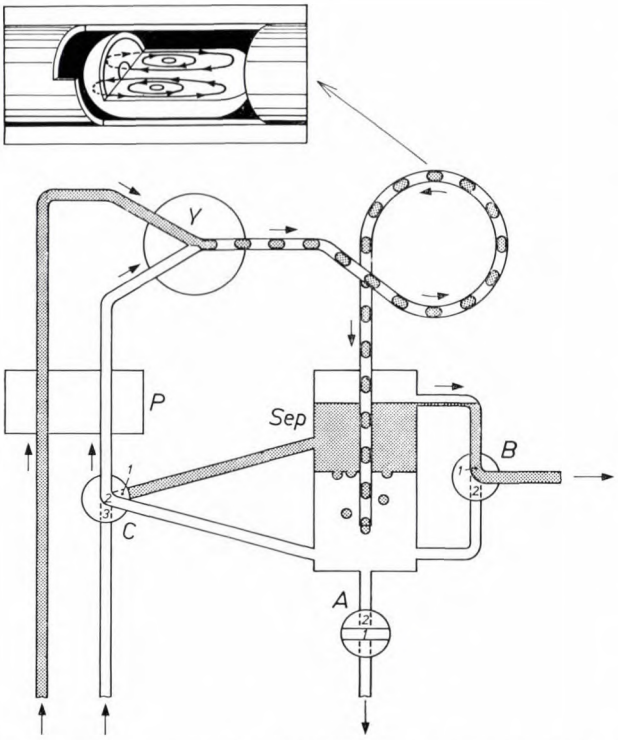


Fig. 2. The operating principle of the extractor. Two liquids (one polar, one non-polar) are fed into a narrow tube, through the Y-piece, by the peristaltic pump *P*, giving the drop pattern shown. This is followed by separation in the separating vessel *Sep*. The following extraction modes are possible: through flow of both liquids (valve *A* in position *2*, valve *C* in position *3*); recirculation of one of the liquids and removal of the other, i.e. a discontinuous cross-current extraction (*A1*, *B1*, *C2* or *A1*, *B2* and *C1*); recirculation of both liquids (either of the last two combinations of valve settings, but with the outflowing liquid led back via the left-hand supply tube).

If every extraction stage in the process is carried out in a separate extractor — by connecting a number of extractors appropriately — the process can be run continuously. With this arrangement extraction can be performed in either the cross-current or the counter-current mode (see also fig. 4).

is shown in *fig. 3*. Three rollers, mounted on bearings at intervals of 120° around the circumference of a wheel, force the liquid through a flexible tube lying in a bed concentric over the same distance with the circumference of the wheel. This type of pump is ideally suitable for the synchronous propulsion of two liquids as required here; all that need be done is to place another tube alongside the first in the same bed. Once set to a particular flow, the peristaltic pump maintains that flow accurately within certain limits whatever the load, i.e. for any length and diameter of the feed tubes or the extraction tube. The flow is also independent of the density and viscosity of the liquid, so that it is not difficult to propel two liquids that differ in this respect at the same flow rate. Other advantages of the peristaltic pump are that it is simple, self-priming, can be made entirely from non-contaminating plastics and has no dead space.

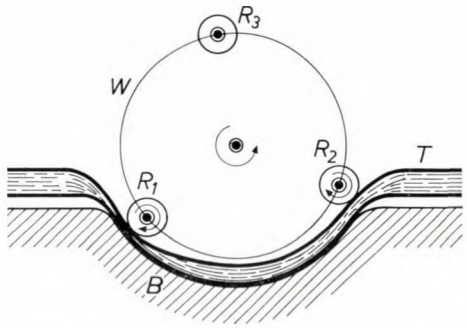


Fig. 3. The operating principle of the peristaltic pump. Three rollers *R*, mounted on bearings fastened to the wheel *W*, force liquid through the flexible tube *T* in the bed *B*. The pump is shown at the moment when R_1 has pinched the tube shut and R_2 has almost completely released pressure on it again.

tube wall and the drops and columns moving along it sets up a vigorous — and reproducible — internal stirring in the drops and columns (inset in fig. 2). This ensures a high extraction rate.

The arrangement of our extractor prevents emulsification and the formation of foam — two familiar but undesirable effects when liquids such as polluted surface water are shaken in a separating funnel: it is very important to avoid these effects in automatic measuring equipment. Emulsification does not occur here because the drop size in the device is more or less predetermined. Another important consideration is that because of the high extraction rate a relatively large drop size can be accepted. The formation of foam is prevented by the absence of air in the extraction tube.

We use a peristaltic pump (*P* in fig. 2) for propulsion of the liquids in the extractor. One form of such a pump

When two liquids are separated in the separating chamber (*Sep*), the heavier one goes to the bottom and the lighter to the top. Valves *A*, *B* and *C* can be set so that either of the two liquids is recirculated while the other is drained off. In fact they give a choice of the following four extraction modes: 1. through flow of both liquids (i.e. a single extraction stage); 2. recirculation of the non-polar liquid and through flow of the polar liquid, or vice versa (cross-current extraction); in both cases either liquid may be heavier; 3. recirculation of both liquids (sometimes desirable with a solution that is difficult to extract); 4. extraction on the counter-current principle by linking up two or more extractors (*fig. 4*). Gas-liquid extraction is also possible.

[1] See for example D. J. Kroon and M. Q. Mengarelli, Philips tech. Rev. **34**, 113, 1974.

[2] See L. Alders, Liquid-liquid extraction, Elsevier, Amsterdam 1955.

Finally, a practical example to illustrate what the new device can do. Lead present in a litre of water taken from the River Waal (20 ng/ml) was concentrated 30 times within half an hour. Extraction was effected at a pH of 4.6 with 30 ml of continuously recirculated carbon tetrachloride to which 0.03 g of the sequestering agent diethyl ammonium dithiocarbamate had been added. Over 98% of the lead was extracted with a contact time in the tube of 16 s. The extraction tube was 8 metres long, with a diameter of 1.5 mm, and the velocity of the liquid in the tube was 0.5 m/s (corresponding to a flow rate of 1.8 l/h).

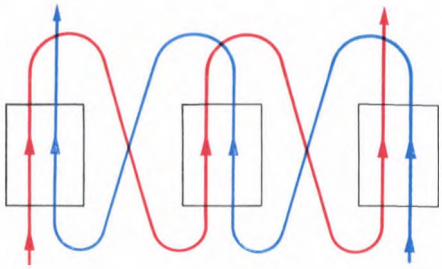


Fig. 4. Diagram showing how three extractors (the three blocks) are connected for extraction in the counter-current mode. Although the extraction procedure as a whole takes place in the counter-current mode, the liquids in the extraction process in any one extractor both flow in the same direction. This is the only kind of arrangement possible with our extractor, but this in no way detracts from its effectiveness.

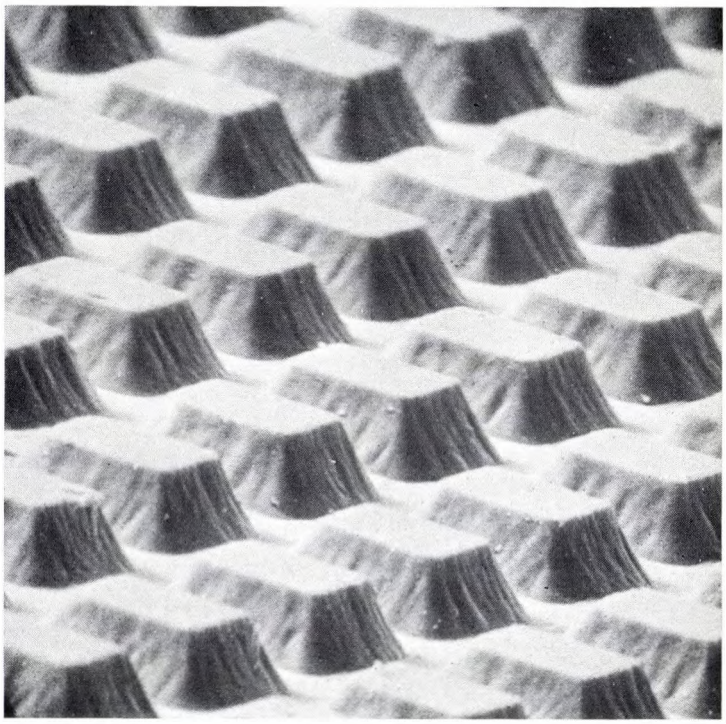
A. van Tooren

A. van Tooren is with Philips Research Laboratories, Eindhoven.

An investigation of ion etching

H. Dimigen and H. Lüthje

The ever-increasing demand for further miniaturization continues to result in the development of new processes. The article below describes a method of etching based on the use of accelerated inert-gas ions, which is already in use in many laboratories and in pilot production. Compared with the chemical etching methods that have normally been used until now, this method improves the resolution in a lateral direction by a full order of magnitude for the same etch depth, or gives a considerably greater etch depth for the same lateral resolution. The method has previously been mainly employed in semiconductor technology, but its value has also been demonstrated in a wide range of other applications.



Introduction

Etching is an important step in the semiconductor technology for the manufacture of integrated circuits [1]. Chips of semiconductor material, usually silicon, are first oxide-coated by means of heat treatment and openings are etched in the oxide layer. Diffusions through these openings form zones of the appropriate *P*- or *N*-type silicon at the desired locations. In the past the etching has usually been done with reagents such as hydrogen fluoride (chemical etching methods) although recently other methods have gained ground.

One of these recent methods, ion etching, is the subject of this article. The increased attention that this method has attracted in recent years is connected with the increasingly exacting requirements that progressive miniaturization places on etching technology—requirements that can no longer be met by the conventional chemical etching methods. These chemical methods will only give a smallest separation of about 5 μm between adjacent grooves of an etched pattern with a desired etch depth of about 1 μm. With ion etching, however, patterns can be made under the same conditions with groove spacings of only 1 μm.

Ion etching consists in bombarding the surface of the substrate with ions of an inert gas (*sputtering*) which causes an erosion of the surface layer. This erosion is a direct result of the collision process, i.e. the physical interaction between the incident ions and the surface of the substrate. In this article we shall not consider the chemical attack on surfaces by certain reactive ions sometimes used in methods such as plasma etching.

Ion etching is useful in many other fields besides semiconductor technology. Its applications include the production of spiral-groove bearings and of integrated optics, and also the fabrication of the patterns used in bubble and magneto-optical memories. The title photograph shows an example of its use in a magneto-optical memory. The pattern shown has a periodicity of 20 μm, while the grooves between the individual elements are 10 μm wide and 5.5 μm deep [2].

Dr H. Dimigen and H. Lüthje are with Philips Forschungslaboratorium Hamburg GmbH, Hamburg, West Germany.

[1] See for example A. Schmitz, Philips tech. Rev. **27**, 192, 1966.
[2] J.-P. Krumme and H. Dimigen, IEEE Trans. **MAG-9**, 405, 1973.

In the course of our investigation reported here into ion etching we made frequent use of specimens of the type illustrated in *fig. 1*. These consist of a silicon chip of diameter of about 30 mm, covered in some cases by a thin oxide layer, to which an etching mask that might typically be a titanium pattern of parallel strips 20 μm wide and 1-2 μm thick has been applied. The effect of ion bombardment on these specimens is easily investigated by breaking them at right angles to the strips after etching and studying the exposed cross-section under a microscope.

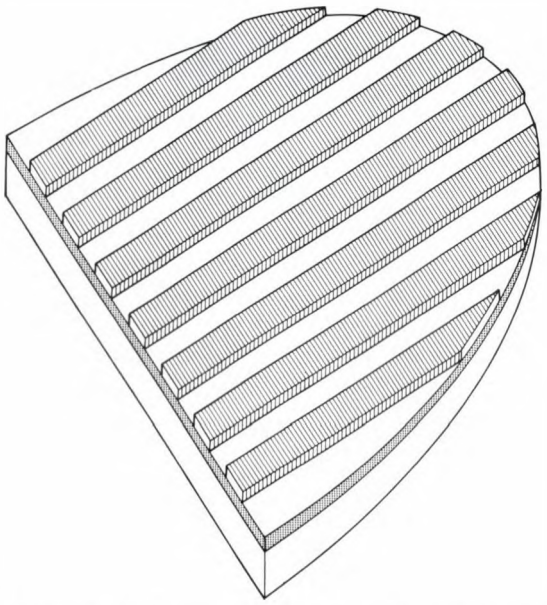


Fig. 1. The type of specimen on which most of the investigations described were carried out: a silicon chip covered with a thin coating of oxide, with an etching mask consisting of parallel strips of titanium or other substance on top of it. After ion etching, the specimen is broken at right angles to the longitudinal direction of the strips and the resultant cross-section is compared under a microscope with that of the unetched specimen.

Fig. 2 shows such a cross-section before and after etching (a titanium mask on a silicon substrate). We conclude from our examination of this cross-section that there has been etching of both the mask and the substrate, so that the original pattern of both the substrate surface and the mask have been altered by the etching.

It is understandable that both should have been affected, since the bombardment energy required to ensure a sufficiently high etch rate is several orders of magnitude greater than the bond energy of atoms in the solid state. There are therefore no solids that would be unaffected by such a bombardment.

The extent to which the resultant etched pattern differs from the original mask pattern is the main thing that determines the practibility of ion etching as a production method. In the last few years we have conducted extensive investigations with the aim of determining the effect of various parameters on the result of etching, in order to make the process as suitable as possible for practical use.

Our investigations showed that it is not necessary at the outset to consider in detail the mechanism of the microscopic collision processes that occur between the etching ion and the surface of the substrate [3]; it is sufficient to know the average value of the sputtering rate S (the number of ions sputtered per incident ion) that each material exhibits on bombardment with ions possessing a well defined energy.

We shall now begin by giving a description of two ion-etching methods now in use. This will be followed by a section dealing with the etch rates that can be attained. Finally, we shall analyse the changes produced in the mask and substrate by the etching, from which we shall draw a number of conclusions relating to the requirements placed upon the mask material.

Ion-etching methods now in use

The ions most commonly used for etching the surfaces of solids are those of argon, at an energy between 500 and 5000 eV, most of which is converted into heat in the substrate. The heat flux that the substrate will withstand is therefore the quantity that determines the maximum permissible energy density of the ions. It is generally in the region of several watts per square centimetre.

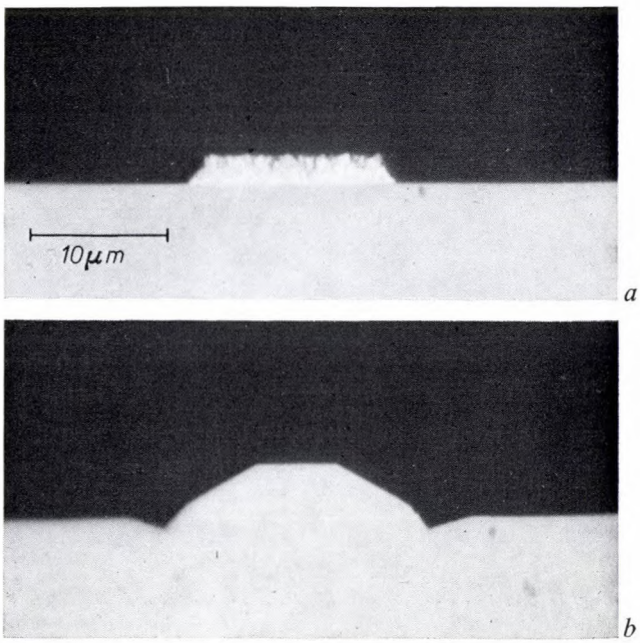


Fig. 2. Cross-section of a silicon chip with titanium strips, (*a*) before and (*b*) after etching. It can be seen that the mask as well as the substrate has been attacked by the etching.

The production of the plasma, the extraction of the ions required for etching and their acceleration can be effected in a number of ways. There are two main methods: radio-frequency cathode sputtering and ion-beam etching, each of which has its own special features and its own field of application.

Radio-frequency cathode sputtering

Fig. 3 illustrates the principle of an arrangement for radio-frequency (r.f.) cathode sputtering. The gas pressure in the conducting 'bell-jar' *B* is reduced to less than 10^{-4} Pa by a vacuum-pump system. Argon is then admitted through the valve *V* until the pressure is about 1 Pa. The substrate *S* is mounted on the cooled target electrode *T*, with which it must be in good thermal contact. The bell-jar is always kept at the same potential as the anode *A*.

If an r.f. source of sufficient power (about 1 kW) is connected to the target in the way shown in the figure, a self-maintaining discharge will be set up between the target and the combination of anode and bell-jar. Because of the difference in the mobilities of the electrons and the ions, and the difference in area between the target and the anode-bell-jar combination, the plasma and the anode will have approximately the same potential and the target will take up a much lower potential than the plasma. Positively charged argon ions from the plasma are consequently accelerated in the direction of the target, where they cause etching of the substrate.

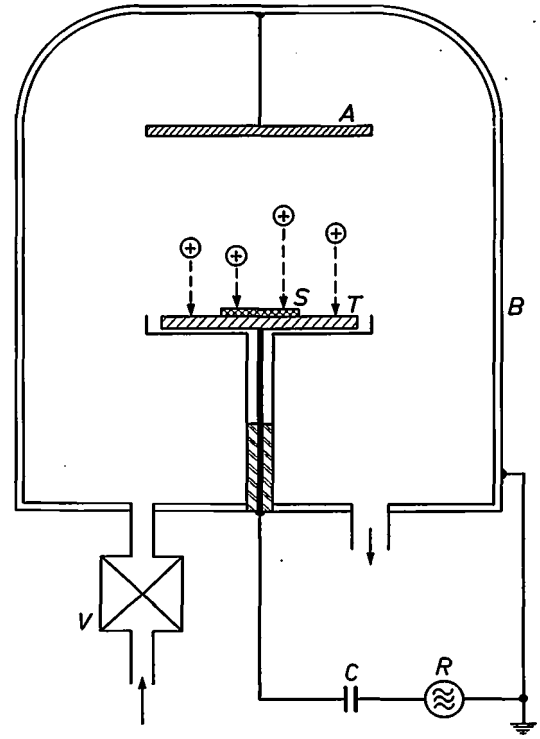


Fig. 3. Schematic diagram of a device for r.f. cathode sputtering. *B* bell-jar of conducting material with a vacuum pump connected at the lower right (to reduce the pressure to about 10^{-4} Pa); argon is admitted via valve *V* until a pressure of about 1 Pa is reached. *R* r.f. source. *C* capacitor. *A* anode in electrical contact with the bell-jar. *T* target cathode with the substrate *S* attached.

The advantages of the method described here may be brought out more clearly by a more general examination of the problem. We shall consider first the hypothetical case of two electrodes of the same area and with a d.c. voltage between them. If the voltage is sufficiently high there will be a discharge. Near the cathode there is the well known 'dark space', corresponding to the region the electrons emitted by the cathode have to pass through in order to produce ionization. In this case the sputtering takes place at the cathode side. Non-conducting substrates cannot be etched in this way, since they do not pass a current and therefore become charged.

If the d.c. voltage is replaced by a low-frequency a.c. one the system behaves as if it had two cathodes: a dark space is formed at each of the electrodes. Sputtering now occurs at both sides, which is generally undesirable. In this case the substrate also has to be conducting.

When an r.f. voltage (frequency greater than 10 MHz) is applied, discharges of the kind described above do not appear, but another ionization mechanism comes into operation. This comes about because the electrons in the high-frequency field perform forced oscillations and, as a result of elastic collisions with gas atoms, can take up sufficient energy to cause ionization. On its own, however, this mechanism is unable to impart to the much heavier ions sufficient energy to produce sputtering of any significance.

Another effect also occurs. It has been observed that the plasma resulting from such a high-frequency discharge assumes a strongly positive potential with respect to the electrodes. This is a consequence of the much greater mobility of the electrons compared with that of the ions, so that the electrons readily find their way to an electrode in the parts of the radio-frequency cycle during which the electrode is positive with respect to the plasma. This causes a depletion of electrons in the dark space, and this in turn leads to the increase in the potential of the plasma. As a result of this potential distribution positive ions are attracted to the electrodes. Once they have arrived in a dark space, they reach the electrode within several cycles of the r.f. voltage and cause sputtering there — this again occurs at *both* electrodes.

Even if the potential difference between the plasma and the electrodes in the appropriate part of the a.c. voltage cycle is small, sufficient electrons can flow from the plasma to the electrodes to compensate for the ion flow. The r.f. current through the dark space is therefore mainly a displacement current, so that the r.f. coupling across the dark space is to a first approximation capacitive.

This capacitive coupling allows the voltage across the dark space at the target to be made much larger than that at the anode. This can be done by making the anode area greater than the target area; in our arrangement we did this by connecting the anode to the bell-jar. Calculations based on a model showed that the voltage across the dark spaces at the electrodes is inversely proportional to the fourth power of their area [4]. The area of

[3] Discussions of the mechanism of the impact process are given in M. Kaminsky, Atomic and ionic impact phenomena on metal surfaces, Springer, Berlin 1965, and in G. Carter and J. S. Colligon, Ion bombardment of solids, Heinemann, London 1968.

[4] H. R. Koenig and L. I. Maissel, IBM J. Res. Devel. 14, 168, 1970.

the electrode formed by the anode and bell-jar need not therefore be much larger than that of the target in order to reduce the voltage across the dark space at the anode to practically zero and ensure a high voltage across the dark space at the target, so that the ions accelerated in the direction of the target acquire the energy required for sputtering, while no sputtering occurs at the anode.

It will also be necessary to compensate the voltage across the dark space at the target by a d.c. voltage somewhere between the target and the voltage source. One way of doing this is to include a capacitor in the external part of the circuit (see fig. 3), or to use a non-conducting substrate for the target. This means that a non-conducting substrate is not now merely permissible, it is advantageous.

There are two particularly interesting aspects of r.f. cathode sputtering: the equipment required for etching is very simple, and it is possible to use very large targets (with diameters of 30 cm or more), so that many substrates can be etched simultaneously. This is extremely important in production applications. As we have seen, the method is also suitable for etching non-conducting substrates, while the use of an a.c. voltage prevents the charging of the substrate that would otherwise result from the ion bombardment, since the charge is cancelled by a periodic supply of electrons.

There is however the disadvantage that, as a result of the potential distribution around the target, the etching ions can only strike the target and substrate at right angles to the surface. This makes it impossible to carry out ion etching at any other angle of incidence.

Another disadvantage is that a gas discharge of the type used here requires a relatively high argon pressure, since the mean free path of the ions has to be appreciably smaller than the distance between the target and the anode. As we shall show later, this can have a very adverse effect on the results of etching.

Ion-beam etching

The second etching method we wish to discuss — etching with an ion *beam* — can be carried out with an arrangement of the kind shown schematically in *fig. 4*. An arc discharge is set up in the upper part of the bell-jar *B*, which contains argon, with the aid of a heated cathode *Ca* and an anode *A*. The gas pressure in an arc discharge can be very low; in this case it is about 10^{-2} Pa. Ions are drawn away from the arc by a sieve-like electrode *E*, then accelerated and fired into the lower part of the bell-jar. The current density of the ion beam is increased by a magnetic field produced by the coil *Co*. The ions thus released finally strike the substrate *S*, mounted on a cooled substrate holder *SH*, which can be rotated about two axes.

This arrangement has a number of advantages. The energy and current density of the ions can be varied independently, since the arc discharge, the voltage on the electrode *E* and the strength of the magnetic field can all be adjusted separately. Since the region around the substrate is nearly free of field, the substrate can be bombarded at any desired angle. The gas pressure can be kept so low that the mean free path of the ions is large with respect to the distance between the substrate and the bell-jar.

As in r.f. cathode sputtering, it is possible to etch dielectrics, provided that the positive surface charge created in the process is compensated by a source of electrons.

These beneficial features are to some extent offset by disadvantages. The system is relatively expensive, while there is at present no device that can deliver a sufficiently homogeneous and dense beam with a diameter much greater than 8 cm. Simultaneous etching of a large number of substrates with a large area is therefore not possible.

Etch rate

Anyone interested in ion etching will obviously wish to know what kind of etch rates can be achieved, i.e. the layer depth removed per unit time. The etch rate depends on many parameters, such as the kind of ion, the energy and density of the ions, and the nature of the

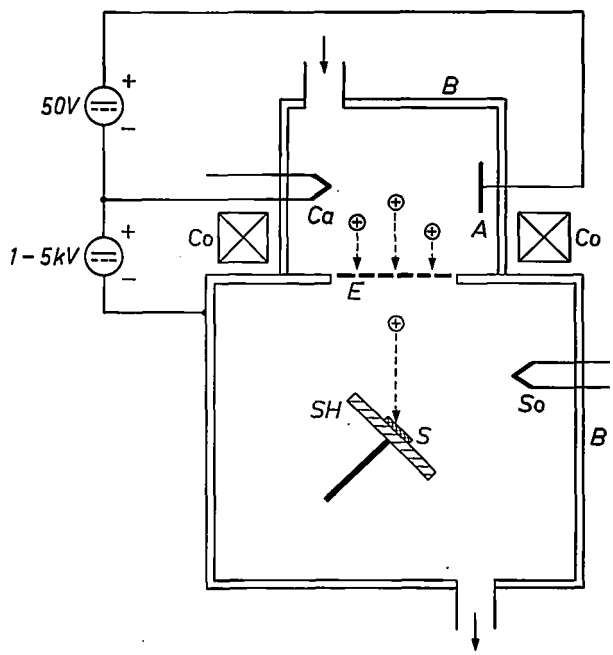


Fig. 4. Schematic diagram of an arrangement for ion-beam etching. Ions are removed from an arc discharge between the filamentary cathode *Ca* and the anode *A* by means of the sieve-like electrode *E* and the coil *Co*, accelerated and fired into the lower compartment of the bell-jar *B*. The ion beam produced strikes the substrate *S*, located on the substrate holder *SH*, which can rotate about two axes. A pump is connected to the lower aperture, and argon is admitted via the upper aperture. In this arrangement the pressure can be lower (about 10^{-2} Pa) than in the device described in fig. 3. *So* electron source used to compensate for any surface charge produced on the substrate.

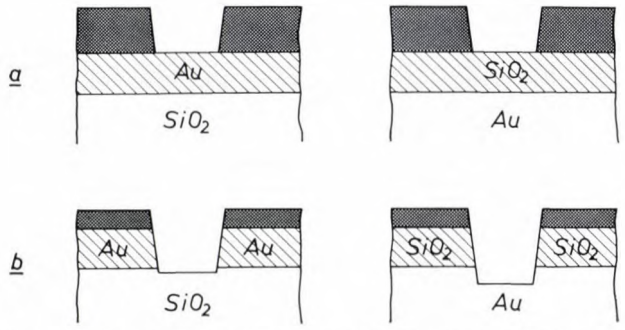


Fig. 5. Specimens consisting of a mask and a layer of gold (*left*) or SiO_2 (*right*) which has to be etched away from a substrate of SiO_2 or gold, *a*) before etching, *b*) after etching. The unwanted but unavoidable etching of the substrate is much deeper in the specimen on the right, because the etch rate of gold is higher than that of SiO_2.

substrate to be etched. Under normal conditions it is in the region of 10 to 300 nm/min.

This applies for *any* material, which means that ion etching is not very selective in comparison with other methods. It is always necessary to be careful that the material underneath the layer to be removed does not become attacked once the etching time required for the layer has elapsed. The higher the etch rate of the substrate relative to that of the material to be etched, the greater the care that must be taken with the control of the etching conditions (*fig. 5*).

So far it has been tacitly assumed that the etch rate for a particular material has a certain well defined value under given conditions. However, this is only true on a macroscopic scale. If we examine the etching of a polycrystalline layer more closely we observe that etching occurs preferentially at grain boundaries. What is more, different rates also apply for the different crystallographic directions of the crystallites. Examination of etched polycrystalline layers therefore generally reveals a roughening of the surface. *Fig. 6* shows an etched polycrystalline layer with two parts that have not been etched, because they were masked; the effect of the etching can be seen clearly. (The importance of this effect will be obvious when we consider that if the aluminium layer is etched away completely the resultant surface roughness is passed on to the substrate layer underneath.)

A final factor affecting the etch rate is the back-diffusion of sputtered material. The mechanism of this process, which oddly enough has received little attention to date, is shown schematically in *fig. 7* for r.f. cathode sputtering. If the argon pressure is low and the mean free path λ of the atoms in the bell-jar is small compared with the distance d between the target and the anode, some of the material removed by etching returns after collisions to the target or to the substrate being etched. There is therefore a diffusion of substrate material to the target and of target material to the substrate.

This gives a reduction in the etch rate, as shown in *fig. 8*. The etch rates are shown in this figure as functions of the argon pressure for a gold substrate on a copper target and for a copper substrate on a copper

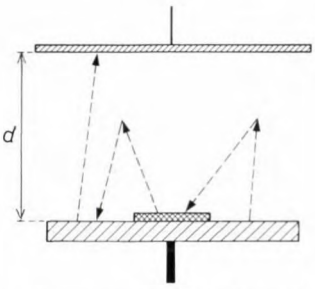


Fig. 7. Diagram to illustrate back-diffusion. Sputtered particles, after collision with gas atoms, can diffuse from the target towards the substrate and in the opposite direction. Back-diffusion affects the etch rate and increases as the free path λ of the ions becomes smaller in relation to the distance d between the anode and the target, i.e. with increasing gas pressure.

Fig. 6. An etched layer of polycrystalline aluminium that was partly screened and partly not. The effects of local variations in the etch rate can be clearly seen in the unprotected part; these must be ascribed to the higher values resulting at the grain boundaries and in certain crystallographic directions.

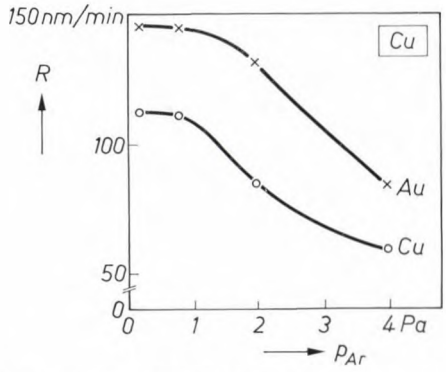


Fig. 8. The etch rate R of gold and copper substrates, both on a copper target, measured as a function of the argon pressure p_{Ar} in a cathode-sputtering equipment, at constant ion-current density. As a result of increasing back-diffusion the etch rate decreases markedly with increasing argon pressure.

target. It can be seen that both rates decrease in about the same proportion with increasing argon pressure, i.e. with increasing back-diffusion. In this case it apparently makes no difference whether the atoms back-diffused are of a different material (copper on gold) or of the same material (copper on copper).

Another complication that can arise with back-diffusion is that the back-diffused target material can adhere to the substrate surface in such a way that the etch rate is lowered more by this material in its new environment than might be expected from the results shown in fig. 8 [5]. A case of this kind is illustrated in *fig. 9*. Titanium targets with substrates of either titanium or silicon were used for the experiments. It was found that

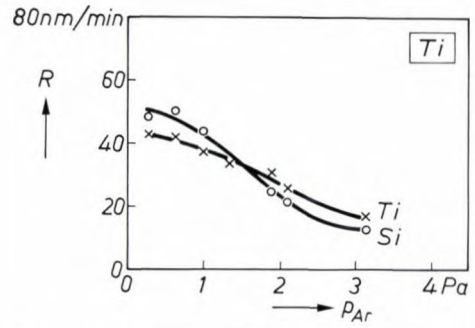


Fig. 9. As fig. 8 but with titanium and silicon substrates on a titanium target. In this case we find that as a result of back-diffusion titanium deposited in a silicon environment produces a greater reduction of the etch rate than when it is deposited in a titanium environment. It would appear that some kind of bond is formed between titanium and silicon.

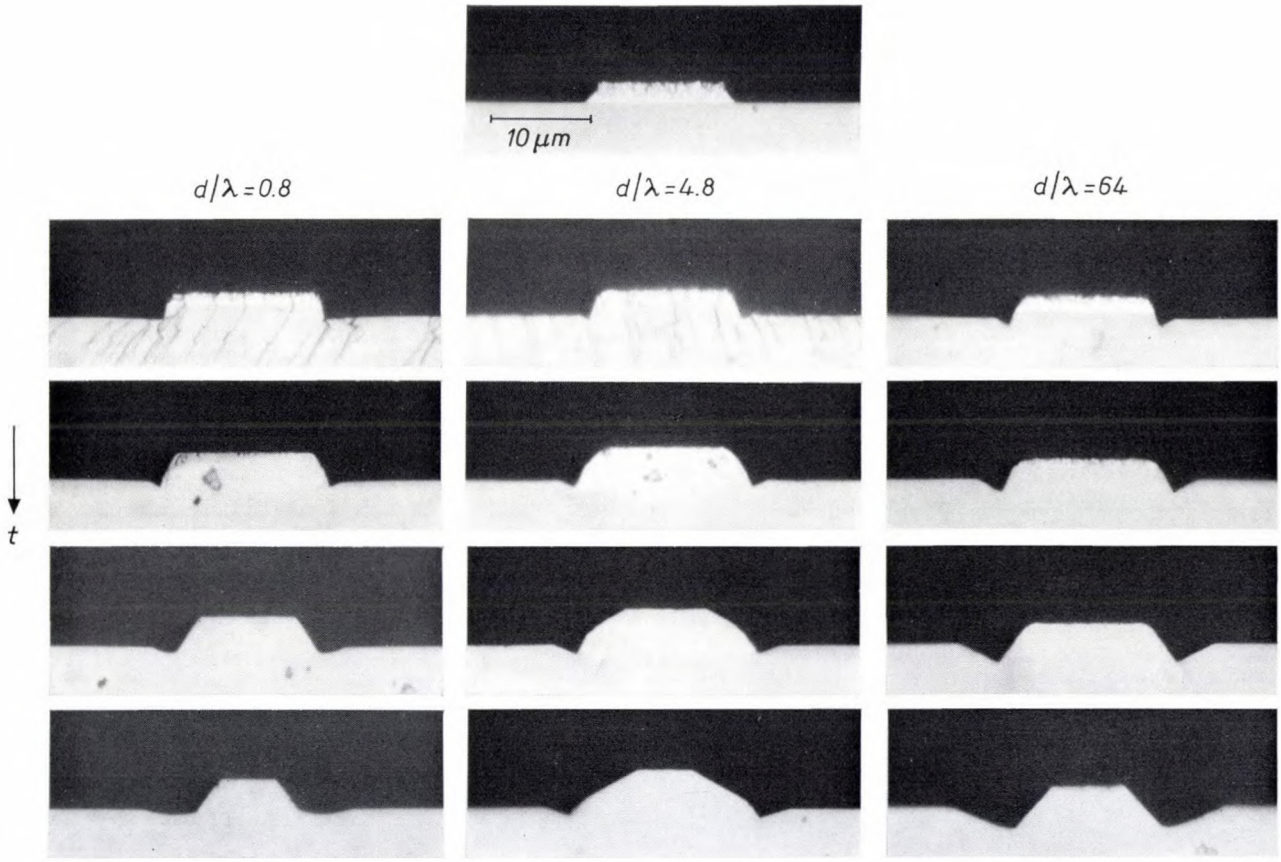


Fig. 10. The effect of an ion bombardment on a silicon substrate coated with a titanium mask (strips 20 μm wide, 3 μm in thickness). Etch time t increases from top to bottom, while the argon pressure, and hence the ratio d/λ (see fig. 7), increases from left to right. With increasing etch time the width of the strips becomes smaller and the slope at the edges decreases. 'Ditches' also appear on either side of the strips; the ditches increase in depth and width with increasing back-diffusion.

with negligibly small back-diffusion the etch rate of the silicon is higher than that for titanium; however, the converse situation comes into effect as the back-diffusion increases. It appears that titanium has an appreciably lower etch rate in the presence of silicon than on its own.

The solution to many etching problems can be considerably complicated by the effect of back-diffusion. Surfaces that have to be cleaned by ion etching are in fact contaminated by back-diffusion, which again is undoubtably responsible for many of the uncertainties encountered in comparisons of the etch rates, or ratios

of the etch rates, reported by many investigators. The situation will be clarified when back-diffusion is sufficiently taken into account in ion-etching experiments.

The imaging of the mask in the substrate

The quality of an etching method used to apply a pattern depends on the accuracy with which the mask pattern originally applied is reproduced in the layer to be etched. In fig. 2 it could be seen that this is a problem in ion etching, because under ion bombardment the dimensional stability of the mask is never perfect. We shall now take a closer look at this problem.

Fig. 10 shows some of the stages in the ion-etching process. The picture right at the top shows a cross-section of a substrate (Si) as yet unetched and with a mask on top of it (a titanium strip 20 μm wide and about 3 μm thick). Other cross-sections are arranged beneath this first one in order of increasing etch time, for three different values of the gas pressure and hence of the ratio d/λ. The following points emerge immediately. The mask strip becomes narrower and narrower during etching and the slope of the sides of the strip also changes. These effects depend on the ratio d/λ, but also arise when the back-diffusion of the sputtered material is negligible ($d/\lambda < 1$).

The changes undergone by the mask strip are to a large extent determined by the way in which the etch rate depends on the angle of incidence of the ions on the substrate. *Fig. 11* shows how most materials behave in this respect. The quantity $S(\Theta)$, i.e. the number of sputtered particles per ion incident at the surface, goes to a maximum and then decreases rapidly (fig. 11*a*). The significance of this for the shape of the mask strip during etching is illustrated in fig. 11*b*. In addition to a change in slope angle, there is also a marked narrowing of the mask strip, because the sides are etched away faster than the top.

The way in which the width reduction ΔL depends on $S(\Theta)$ and the etch depth Δd [6] is given by:

$$\Delta L = \Delta d\left\{\frac{S(\Theta)}{S(0)} - 1\right\}(\tan\Theta)^{-1}.$$

The results of measurements that we made confirm this relation; they are shown in *Table I*. The inconsistencies that we found in these measurements when etching layers of titanium and molybdenum are associated with the peculiar anisotropic structure of these layers.

The explanation for the 'ditches' seen in the etched material near the edges of the mask strip in fig. 10 is probably a locally higher etch rate due to ions rebounding elastically from the sides. These ditches are narrow when the back-diffusion is negligible, but become appreciably wider and deeper with increasing back-diffusion. *Fig. 12* gives an explanation for this. While point *A* receives back-diffusing material from the solid angle Φ_1, point *B*, which is closer to the side of the mask strip, receives material from the much smaller solid angle Φ_2. The difference between sputtered and back-diffused material results in a higher etch rate for

Table I. The ratio of the width reduction ΔL of the mask strip in fig. 1, occurring as a result of ion etching, to the etch depth Δd, and also the ratio $S(\Theta)/S(0)$ (see fig. 11*a*), calculated and measured for six values of the angle of incidence Θ.

Material	Θ (°)	$\Delta L/\Delta d$ measured	$S(\Theta)/S(0)$ calculated	$S(\Theta)/S(0)$ measured
Silicon (111)	31	0.98	1.59	1.72
Silicon (111)	57	2.34	4.60	4.02
SiO_2	35	1.46	2.02	2.25
Photoresist (AZ 1350 H)	68.7	1.35	4.46	3.87
Titanium	45	0.04	1.04	1.89
Molybdenum	60	0.04	1.07	1.48

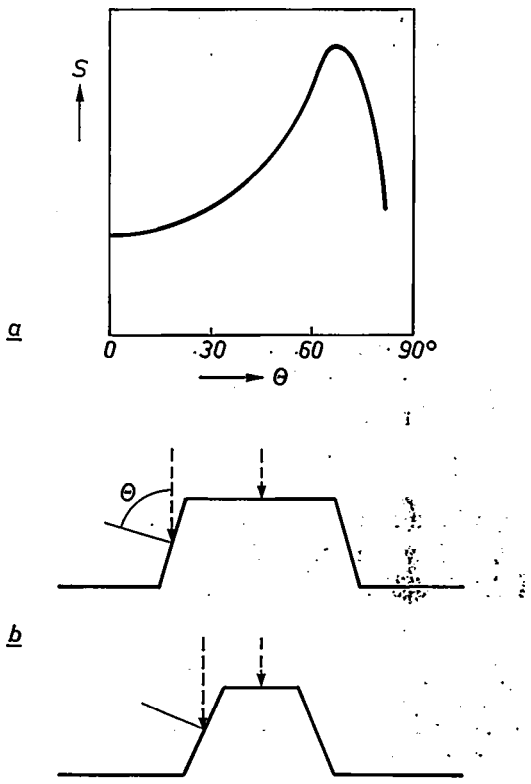


Fig. 11. *a*) Example showing how the sputtering rate $S(\Theta)$, i.e. the number of sputtered particles per incident ion, depends on the angle of incidence Θ. *b*) The low value of sputtering rate for vertical incidence compared with that obtained with bombardment at an angle accounts for the reduction of the width of the mask strips during etching. The fact that the top and side of the mask strip never meet at a sharp corner but merge smoothly means that the etch rate is faster at the top of the side surface than underneath, resulting in a reduction of the slope during etching. (The upper figure in (*b*) is the unetched specimen, the lower figure the etched specimen.)

[5] M. L. Tarng and G. K. Wehner, J. appl. Phys. 43, 2268, 1972.
[6] A. D. G. Stewart and M. W. Thompson, J. Mat. Sci. 4, 56, 1969.

point *B*, i.e. in a deeper ditch. Quantitatively, the situation is in fact much more complicated because the width reduction of the mask strip referred to above continuously changes the conditions for the formation of ditches during the etching process.

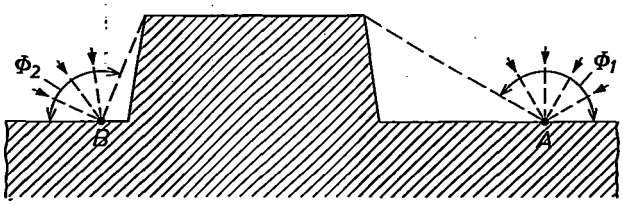


Fig. 12. Digram illustrating the deepening of the 'ditches' formed during the etching as a result of back-diffusion. The solid angle Φ_2 through which ions back-diffuse to point *B* is smaller than Φ_1, which is the corresponding solid angle for point *A*. This results in a higher etch rate at *B* than at *A*, and hence in a deeper ditch.

Requirements to be met by the mask

Because of the ditch formation it is often desirable to restrict the back-diffusion of sputtered material to a minimum by proper choice of the etching method (usually ion-beam etching). When that condition is observed, the accuracy with which the mask pattern is reproduced in the etched pattern depends almost entirely on the behaviour of the mask during etching.

It will also be clear that if it is desired to etch relatively *deep* with a *thin* mask, the etch rate of the mask will have to be low compared with that of the material to be etched. The desirability of using thin masks arises from the fact that only thin masks can be made with high lateral resolution and that therefore only thin masks can give the high-resolution etch patterns that, as stated earlier, constitute the great attraction of ion etching.

The two most important requirements that mask material must satisfy follow from what has been said above. Firstly, its etch rate must be as low as possible. Secondly, sputtering should depend as little as possible on the angle of incidence of the ions, for only under that condition can changes in the mask pattern during etching be kept to a minimum.

If we consider mask materials that might be suitable for the first requirement, the following picture emerges. Photoresist masks give a relatively low etch rate provided the pressure of the residual gas in the bell-jar is very low (photoresists are usually organic and susceptible to attack by reactive residual gases). This condition can be met either with a good pump system or by using a target of a material possessing good gettering properties. So far, however, our best results have been obtained by using masks made of certain metals. Titanium in particular, etched in an argon and oxygen atmosphere, is extremely satisfactory. This is because the oxygen reacts with the titanium, a highly absorbent getter material, to form titanium oxide, which is found to be highly resistant to ion etching, resulting in particularly low etch rates. The etch rates of titanium and silicon, and the ratios of these rates, are shown as a function of the oxygen pressure in *fig. 13*. The figure

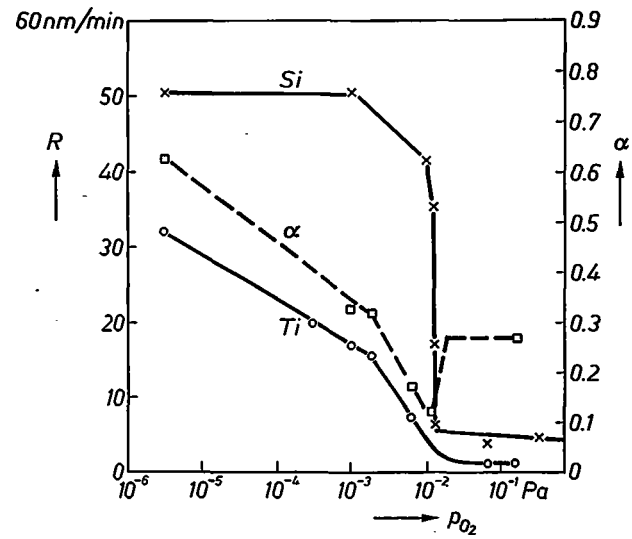


Fig. 13. The etch rates *R* of titanium and silicon and their ratio α as a function of the partial pressure p_{O2} of the oxygen added to the argon in which ion etching is performed. It is clear from the figure that for a given thickness of the titanium mask the largest etch depth in silicon is obtained for an oxygen pressure of about 10^{-2} Pa.

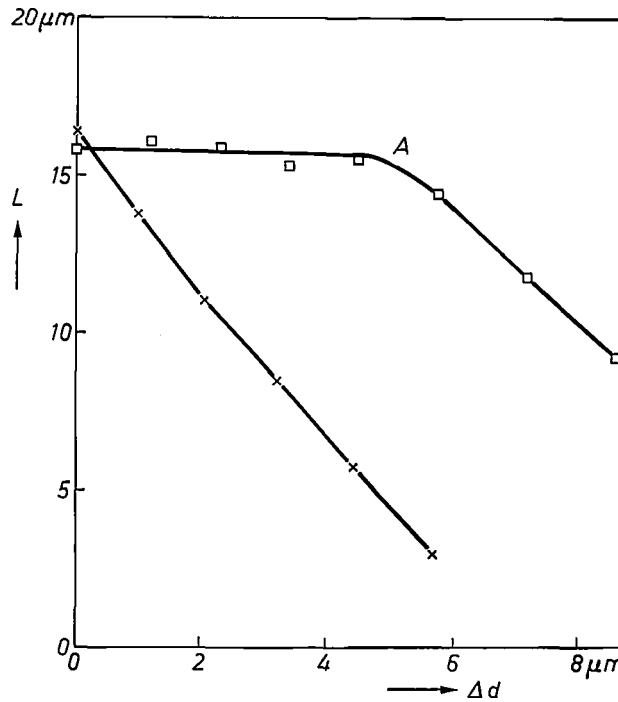


Fig. 14. The width *L* of the mask strips in fig. 1 as a function of the etch depth Δd in silicon, with a photoresist (Shipley AZ 1350 H) and titanium (upper curve) as mask materials. In the case of titanium the width of the strip does not start to decrease until point *A* is reached, i.e. only when the mask has been almost entirely removed by etching. This shows that titanium is considerably more suitable as a mask material than the particular photoresist it is compared with here.

illustrates how important it is to choose the oxygen pressure correctly in order to combine the greatest possible etch depth in silicon with the thinnest possible titanium mask. The etch rate of titanium is low even in the absence of oxygen but not so low as with it.

Now let us consider the desire for mask materials whose sputtering rate is almost independent of the angle of incidence of the ions. Table I has already given some idea of the width reduction occurring in the profiles of masks made of certain materials. *Fig. 14* shows the width L of the mask strips of fig. 1 as a function of the etch depth in the silicon substrate for a photoresist (Shipley AZ 1350 H) and for titanium, in an oxygen-*free* atmosphere. In titanium L remains practically constant up to point A (by which time almost all of the material to be removed by etching has been removed), whereas in the photoresist it immediately starts to decrease perceptibly. The photoresist must therefore be considered unsuitable for applications where the etching has to be both deep and accurate. If there is oxygen in the atmosphere, the direction dependence of the sputtering rate of titanium is not so good.

Titanium is thus found to be an excellent material, although it is a somewhat expensive material for masks. The mask is prepared by depositing a layer of titanium on the substrate to be etched, applying a coating of photoresist and forming the desired pattern by chemical etching.

The two desirable qualities of a mask material such as titanium — a low etch rate, which is also independent of the angle of incidence — permit further optimization by operation with an *oblique* angle of incidence. In most cases, however, the substrate would then have to be rotated about the normal to the surface in order to restrict shadow effects to a minimum.

Let us suppose that a pattern has to be formed in a substrate whose sputtering rate as a function of the angle of incidence is represented by the curve S_S in *fig. 15*, and that the curve S_M applies to the mask material. Curve d gives the relative mask thickness, as calculated from S_S and S_M, which is necessary to etch a substrate to a certain depth. It shows that for larger angles of incidence the mask may be made thinner down to a certain limit. The more the sputtering rate of the substrate at a given angle of incidence exceeds the sputtering rate of the mask for the same angle, the thinner the mask required for etching.

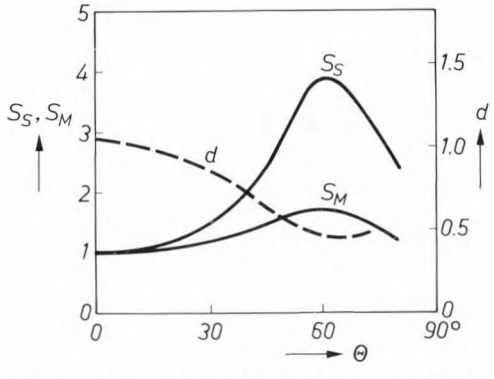


Fig. 15. The sputtering rates S_S and S_M of the substrate and mask materials as a function of the angle of incidence Θ, normalized for vertical incidence. Curve d is the relative depth, as calculated from S_S and S_M, which is necessary to obtain a certain etch depth. It can be seen that to attain this objective with the thinnest possible mask the angle of incidence must be about 60°.

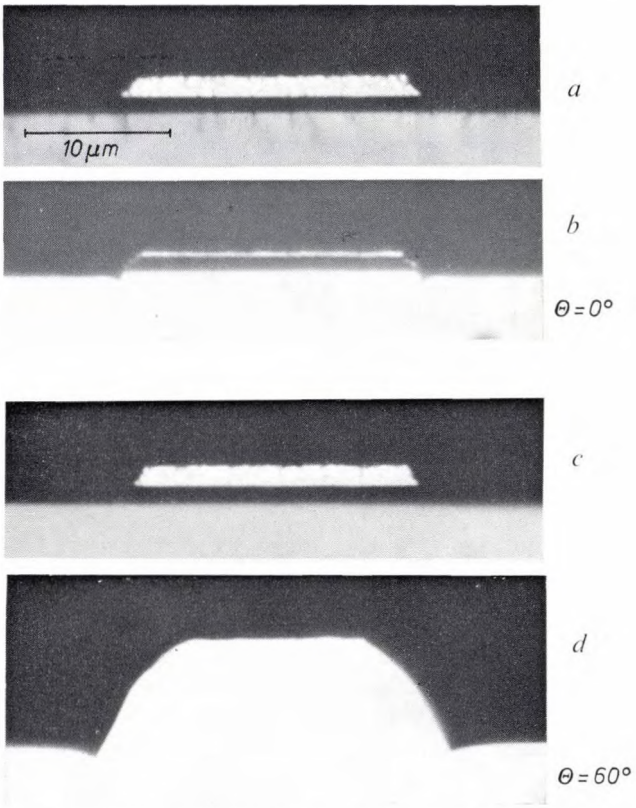


Fig. 16. Sections through 1.5-μm titanium masks deposited on thermally oxidized silicon wafers. (*a*) and (*c*) show the situation before etching (*b*) the situation after etching vertically to the substrate surface and (*d*) after etching at an angle of 60°. The cross-sections clearly demonstrate the much greater etch depth (about 6 μm) that is obtainable in the last case.

This point is further illustrated in *fig. 16*. A thermally oxidized silicon chip is covered by a 1.5-μm titanium mask. If etching is continued until the mask has disappeared, substrate etching to a depth of about 2 μm is possible with vertical ion bombardment. For an angle of incidence of 60° and bombardment in the longitudinal direction of the strips the depth attainable is about 6 μm (rotation is not necessary here because of the line pattern).

This possibility of reducing the thickness of the mask required is of course confined to the ion-beam process, which is the only one permitting bombardment at any desired angle.

Finally, to demonstrate the possibilities of our method, we should like to refer to the title picture again. The mosaic-like pattern shown there of an epitaxial garnet layer (with a periodicity of 20 μm and an etch depth of 5.5 μm) was achieved with the optimized ion-etching method just described. The mask was a titanium layer approximately 3 μm thick, and ion bombardment was carried out at an angle of 45° on a rotating substrate. Had the 'optimum' angle of 60° been used, shadow effects would have restricted the attainable depth.

Summary. With ion etching it is possible, for the same etch depth of, say, 1 μm, to form patterns with lateral dimensions (also 1 μm) that are approximately five times smaller than is possible with conventional chemical etching methods. Two currently practised ion-etching processes are r.f. cathode sputtering and ion-beam etching. Research carried out at Philips Forschungslaboratorium Hamburg, has led to the following conclusions: back-diffusion must be avoided as far as possible (which means that the ion-beam process will often be the most suitable method); the required ratio of etch rates (low for the mask, high for the substrate) can be achieved very effectively with a titanium mask and etching in an argon and oxygen atmosphere; sputtering must depend as little as possible on the angle of incidence of the ions. With regard to this last point, titanium is again found to be an excellent mask material. The best results are obtained with a titanium mask and bombardment at an oblique angle.

Ceramic-to-metal bonding with sputtering as a metallization technique

Ceramic insulators or radiation-transmitting windows that have to be sealed hermetically into the metal wall of a vacuum container are often used in vacuum technology and electron-tube manufacture. Brazing, whether or not on a previously metallized surface, is the principal technique for making such seals. Ceramic-frit techniques and diffusion methods are processes also used for this purpose [1].

Metallization of ceramic surfaces is usually done by applying a suspension of finely divided metal powder and subsequent firing of this layer. A suspension of either a molybdenum-manganese mixture or an active — i.e. readily oxidized — metal, such as titanium, is used (*fig. 1*).

During the past fifteen years we have developed a sputtering technique that offers an attractive additional possibility for applying a metallic coating to a ceramic substrate [2].

The sputtering, which is usually done in argon, allows thin metal coatings of even thickness (0.5-2 μm) to be applied. Automation of the various stages makes the sputtering process more reproducible than metallization with metal-powder suspensions. Other notable advantages of the process are that no firing is required — which is of special importance in sealing infrared windows, which do not always withstand the firing temperatures required — and that small holes can be successfully metallized.

Sputtering is achieved by maintaining a gas discharge between planar electrodes in a gas atmosphere at a pressure of about 25 Pa (2×10^{-1} torr) [3]. At an inter-electrode distance of 45 mm a potential difference of 2-3 kV is required. Cathode material is sputtered by ion bombardment, and the sputtered material is deposited as a thin coating on the surrounding surfaces. The components to be metallized are placed on the water-cooled anode. Parts that should not be coated can be covered with metal shields or cleaned afterwards by grinding, sandblasting, etc.

In our experience the metallization should consist of a series of layers of two or three different metals for successful brazing.

The initial layer has to provide the actual bond to the ceramic substrate and the metal employed must therefore have a high oxygen affinity [4]. On top of this initial layer a second one is applied to serve as a barrier to protect the bonding layer from penetration by the braze metal. A third layer may also be needed to avoid oxidation of the underlying layers and to enhance the flow of the braze metal. The metals used for the different layers should fulfil a number of requirements. In view of the widely differing functions of the layers the essential properties of the metals to be used will be quite different.

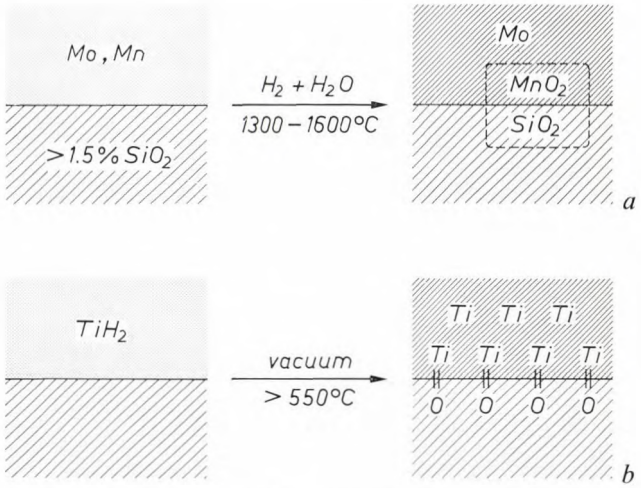


Fig. 1. *a*) Principles of the Mo-Mn-suspension metallizing technique. A suspension of Mo and Mn particles, finely dispersed in a suitable binder, is applied to the ceramic surface by screen-printing techniques or simply with a brush. The ceramic is then fired at a temperature of 1300-1600 °C in a wet hydrogen atmosphere. Mn is oxidized to MnO_2, which forms a low-melting phase with SiO_2 from the ceramic substrate. This melt penetrates the interstices between the Mo particles that have sintered together. After application of a thin nickel coating the metallization is ready for brazing. In this process it is essential for the ceramic to have an SiO_2 content of at least 1.5%.
b) Principle of the active-metal process for metallization of a ceramic. A suspension of finely divided active metal (such as Ti, Ta, Zr, Cr etc.) is applied to the ceramic, usually as a hydride powder suspended in a suitable binder. Sintering in vacuum or in an inert gas decomposes the hydride and establishes a chemical bond between metal atoms and oxygen atoms in the ceramic surface. An SiO_2 content of the ceramic is not essential in this process. If necessary, the sintered metallization can be covered with a thin gold coating to assist the wetting by the brazing filler.

The initial bonding layer, as we noted earlier, must have a high oxygen affinity — free-energy decrease upon

[1] W. H. Kohl, Handbook of materials and techniques for vacuum devices, Reinhold, New York 1967, Chap. 15.
A. Roth, Vacuum sealing techniques, Pergamon Press, Oxford 1966, pp. 197-224 and 704-744.
J. T. Klomp, Powder Metall. Int. **3**, 142 and 193, 1971.

[2] R. L. Bronnes, R. C. Hughes and R. C. Sweet, U. S. Patents 3 339 267, 3 371 406, 3 412 455, 3 418 423, 3 432 913, 3 464 725, 3 505 041 and 3 518 066.

[3] The arrangement used in sputtering is rather similar to that described in the article by Dimigen and Lüthje, this issue, page 201, fig. 3; in our case a d.c. discharge is used.

[4] J. E. McDonald and J. G. Eberhart, Trans. Met. Soc. AIME **233**, 512, 1965.

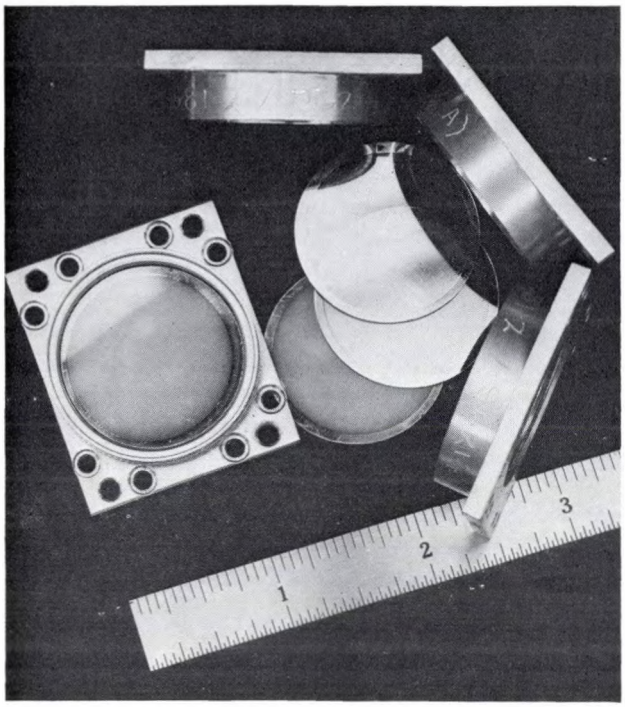


Fig. 2. A number of 'Irtran 2' windows; 'Irtran 2' is an infrared-transparent material, consisting of ZnS. Three windows are shown metallized along the edge. At the left a window is brazed into a metal plate like those shown in a side view at the right. The ruler is calibrated in inches.

oxidation at least 300 kJ per gram atom of oxygen — to ensure good adhesion to the ceramic substrate. The oxide must be non-volatile and both metal and oxide should have sufficient tensile strength. The barrier layer must not completely dissolve the initial layer and both metals must not form brittle intermetallic compounds.

The second layer must function as the barrier to penetration by the braze metal. The third layer is selected to provide compatibility between the metallized surface, the metal to be joined, and the braze metal. It may have an extremely low solubility in the braze metal, or it may dissolve in the braze metal to a limited extent but produce a high-melting solid solution. It is also desirable that each layer of metal be capable of alloying and producing such solutions with the adjoining layer(s). Otherwise, the surface layer would be dissolved completely by the braze metal to expose a layer that cannot be wetted by the braze metal. A two-layer composite of metals that have limited mutual solubility, however, will provide a graded alloy at their interface; this alloy will resist penetration by the braze metal, but will remain wettable. Properly formulated composite metallizing films, even though only a few microns thick, are remarkably resistant to erosion and stripping by molten brazing metal. A good metallization will even withstand a prolonged exposure to molten braze metals such as silver, copper or gold.

Table I. Metals used successfully in the three layers of a multilayer structure that gives a readily brazed metallization on a ceramic. Preferred thicknesses of the layers are also given.

First layer (thickness 0.5 μm)	Ti, Cb (Nb), Ta, Ni-Cr, Zr, stainless steel, 'Inconel'
Second layer (thickness 0.5 μm)	Ni, Au, Pt, Mo, stainless steel
Third layer (thickness 1-2 μm)	Ni, Au, Cu, Pt, Pd, stainless steel

Metals that have been found to be suitable for the three layers are listed in *Table I*, with the preferred thickness of the layers.

The configuration of the components to be brazed is determined in the usual way by factors such as thermal expansion and ductility of the components. A slight compressive stress in the ceramic is normally to be preferred, but excessive stresses due to differences in thermal expansion should be avoided by using feather-edged metal parts or other ductile structures.

A wide range of solders and braze metals can be used. The choice should, of course, be based on factors such as the temperature requirements and the wettability of the metal surfaces.

The kind of protective gas atmosphere that should be used in brazing the ceramic components metallized in the way described is determined in the first place by the materials used. The metallization layer itself introduces hardly any limitations here. Brazing has been carried out successfully in vacuum, in reducing atmospheres and in a noble-gas atmosphere. Cracked ammonia and dry nitrogen can also be used, or exothermic gas (i.e. pure or partly oxidized coal gas). This freedom of choice in brazing atmosphere is a significant advantage of the process.

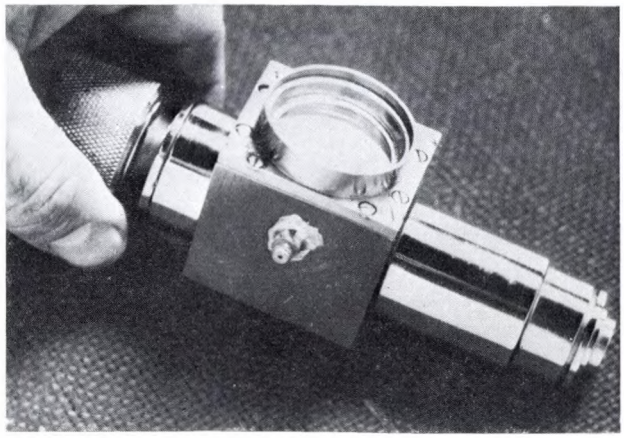


Fig. 3. One of the windows shown in the previous illustration is incorporated here in the wall of a dewar vessel for cooling an infrared detector.

Fig. 4. Dewar vessel for cooling an infrared-detector array consisting of a large number of elements. The synthetic sapphire window has been brazed into the metal dewar wall at the top. A large number of ceramic-insulated connectors are brazed into the wall.

Finally, we give a number of examples of structures that we were able to make with the sputtering technique, taken mainly from work on infrared-detector devices. *Fig. 2* shows a group of 'Irtran 2' (zinc-sulfide) window blanks (metallized) and finished windows. These windows are employed on a side-windowed dewar for an infrared detector as shown in *fig. 3*.

Fig. 4 shows a dewar designed for use with infrared-detector arrays that consist of a large number of elements. A synthetic-sapphire window has been brazed in at the top. The cylindrical surface contains a large number of ceramic-insulated electrical lead-throughs. Sealed-off devices of this general type have retained vacuum for a period of three years.

The strength of seals obtained by the process described here depends upon the identity and properties of the materials employed. Standard ASTM [*] alumina tensile-test specimens (*fig. 5*) brazed with copper-silver

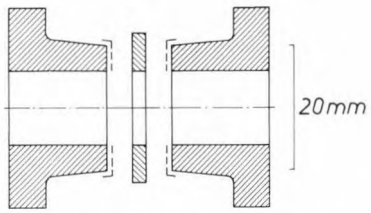


Fig. 5. Geometry of the test specimens — as specified in the ASTM [*] Standard F 19-64 — used for tensile testing of the ceramic-to-metal bonds produced by sputtering composite metallization layers.

eutectic exhibit a tensile strength of 70×10^6 N/m^2 (10 000 psi) or greater. Seals to other materials, if broken by mechanical forces, will usually fracture in the non-metal component.

Vacuum-tightness of seals is regularly tested on a helium mass-spectrometer leak detector. Leaks are seldom detected, and those that do occur are always accompanied by visible defects in the non-metal component or the brazed joint. Evacuated and sealed-off devices, as mentioned above, have retained vacuum for one or more years.

R. L. Bronnes
R. C. Hughes
R. C. Sweet

[*] American Society for Testing and Materials.

R. L. Bronnes and Dr R. C. Sweet are with Philips Laboratories, Briarcliff Manor, N.Y., U.S.A.; R. C. Hughes, now retired, was with the same laboratories.

Digital integrated circuits with low dissipation

N. C. de Troye

The rapid development taking place in digital integrated-circuit technology is directed at integrating more and more elements on a single chip. In this context the dissipation of the individual element has to be reduced without adversely affecting the switching speed. New techniques and technologies are appearing on the scene with great regularity, and the initiated have labelled them with abbreviations or acronyms that leave the outsider somewhat baffled. We are therefore pleased to be able to publish the following survey of new developments in the field; it is based upon an invited paper read by the author at the Sixth International Micro-Electronics Congress held at Munich in November 1974.

The wide range of modern digital electronics — from measuring instruments to computers and communication satellites — is only a practical possibility because of integrated-circuit technology and its extension to 'Large-Scale Integration' (LSI). Significant progress has been made in this field in the past few years; a single silicon chip can now accommodate an entire microprocessor for a small computer. The unit may consist of both MOS and bipolar transistors.

This article presents a survey of a number of technologies and processes suitable for LSI. Only low-power circuits can be used for LSI and subjects such as emitter-coupled logic will not therefore be discussed. Some of these circuit technologies, such as the aluminium-gate MOS logic, are already well established, whereas others such as the silicon-gate technology and the V-groove MOS logic are new. A new bipolar technology is represented by Integrated Injection Logic (I^2L) or Merged-Transistor Logic (MTL). This technology gives high packing density; it is extremely versatile and digital and analog functions can even be combined on the same chip.

Transistors are essentially charge-controlled devices. The charge can be supplied quickly by a high current or more slowly by a low current. In the low-current case the delay time per logic operation is longer but the dissipation is less. The product of the two — the 'speed-power' product — is fairly consistent for a particular technology and is therefore frequently employed as a quality rating. It has the dimension of an energy [1]. It will be employed here as a means of comparing the various types of circuits.

Dr N. C. de Troye is with Philips Research Laboratories, Eindhoven.

Integrated MOS circuits

The number of process steps in the fabrication of a MOS transistor is much smaller than for other types of transistors; it was this inherent simplicity that provided an impetus to work on MOS transistors in the 1960s. MOS transistors are virtually the only element in MOS digital integrated circuits. *Fig. 1a* shows a digital building-block consisting of a NOR gate with two inputs. It comprises three *P*-channel enhancement transistors [2]; the layout is shown in fig. 1*b*. The sources of driver transistors T_1 and T_2 are at ground potential V_{SS}. The logic employed here is 'negative', i.e. the negative state corresponds to '1'. The output Z assumes ground potential ('0') if either or both of gates A and B become negative ('1'). Transistor T_3 acts as a nonlinear load, and will give load values as high as 200 kΩ. The load gate can be connected to the drain, but if the drain voltage V_{DD} is reduced and the gate connected to a separate supply voltage that is more negative than V_{DD}, lower dissipation and greater speed are obtained. The speed-power product for this type of MOS circuit is between 30 pJ and 150 pJ.

Silicon-gate technology

The first MOS integrated circuits had *P*-channel enhancement-type transistors with an aluminium gate [3]. The switching speed of these transistors is not particularly high because of the high threshold voltage that has to be applied before the transistor starts to conduct and because of the parasitic capacitance of the gate, which overlaps the source and drain diffusions. This overlap is necessary to ensure that the gate still covers the channel completely if the masks shift slightly during fabrication.

Efforts were of course made to devise a technology that would give a lower threshold voltage and also ensure that the source and drain diffusions would be produced automatically at the right places. One of the solutions was the silicon-gate technology [4]. In this approach the gate electrode is made of polycrystalline silicon, which has a smaller contact potential relative to the substrate than aluminium, and also acts as a mask during the application of the source and drain diffusions; overlaps are limited to the lateral diffusion depth of the impurities. *Fig. 2* shows a cross-section of a *P*-channel MOS transistor with a silicon gate; the insulation consists of a layer of silicon dioxide.

This use of polycrystalline silicon not only gives an increased switching speed, owing to the smaller parasitic capacitance, it also makes it easy to incorporate double-layer interconnection. This is because the polycrystalline silicon can be oxidized on one side and hence insulated from the aluminium interconnection pattern deposited on top of it later. In *fig. 3* the polycrystalline

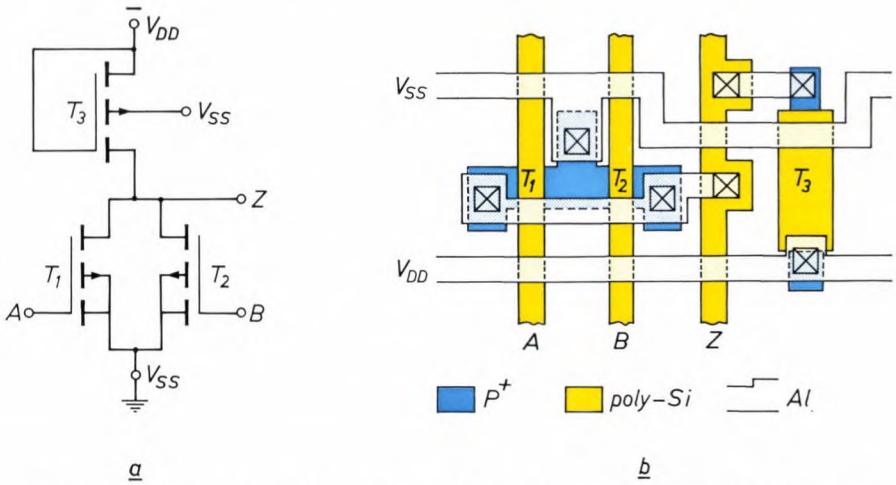


Fig. 1. Circuit diagram (*a*) and layout (*b*) of a NOR gate with two inputs *A* and *B*. The gate consists of three *P*-channel MOS transistors, T_1, T_2 and T_3, all of the enhancement type. Output *Z* is at ground potential if either or both of the inputs are negative. V_{SS} ground potential. V_{DD} negative supply voltage.

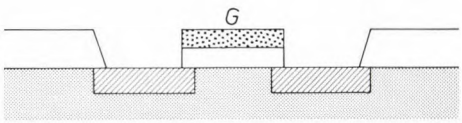


Fig. 2. MOS transistor with polycrystalline-silicon gate *G*. The gate acts as a mask during the diffusion of the source and drain and large overlaps are thus avoided. The materials used are indicated in the key at the foot of the page.

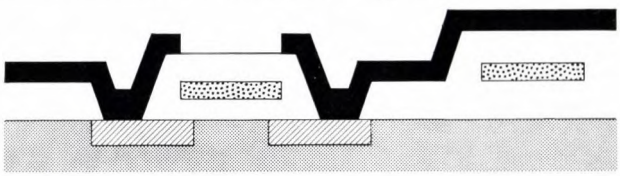
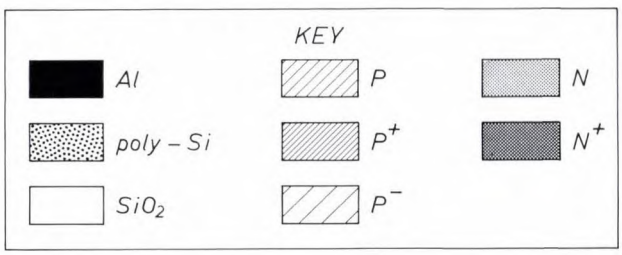


Fig. 3. Use of polycrystalline silicon as a gate (*left*) and as the lower layer of a double-layer interconnection pattern (*right*). The upper layer is of aluminium. The materials used are given in the key.



silicon on the left acts as a gate and that on the right as the bottom layer of a double-layer interconnection pattern.

The next step was to apply the silicon gate to *N*-channel transistors. Since the mobility of the electrons in an *N*-type channel is about twice that of the holes in a *P*-channel, a device of this sort was expected to double the switching speed. In practice, however, doubling the mobility in a circuit does not give anything like twice the switching speed. The dissipation in the resultant circuit also turned out to be twice as high as in *P*-channel transistors, and the speed-power product of the *N*-channel transistors was twice that of the *P*-channel devices [5].

[1] The term 'speed-power product' is a misnomer; the product is in fact that of a time and a power.

[2] An introduction to the subject is given in J. A. van Nielen, Operation and d.c. behaviour of MOS transistors, Philips tech. Rev. **31**, 209-215, 1970.

[3] S. R. Hofstein and F. P. Heiman, The silicon insulated-gate field-effect transistor, Proc. IEEE **51**, 1190-1202, 1963.

[4] F. Faggin and T. Klein, Silicon gate technology, Solid-State Electronics **13**, 1125-1144, 1970.

[5] R. B. Seeds and R. Badertscher, *N* versus *P* channel for single-supply TTL compatibility, 1973 IEEE Int. Solid-State Circuits Conf. Dig. tech. Papers, pp. 32-33.

Ion implantation

It was another new technology, ion implantation, that enabled us to combine depletion and enhancement transistors on the same chip [6]. The implantation was carried out in two stages. The first stage lowers the threshold voltages of all the transistors, though they still remain of the enhancement type. A selectively applied second implantation then changes the load transistors into transistors of the depletion type. A typical result is that of the inverter circuit shown in *fig. 4*.

The advantage of a depletion-type transistor as a load is that there is no need for a separate gate voltage V_{GG} to enable the output voltage to reach the value V_{DD}, so that the gate can be connected to the source. This means not only that the circuit is simpler but also that the transistor always has the same gate voltage, whatever the existing voltage levels. This is not the case with the enhancement-type transistor (fig. 4*a*), where if the driver transistor is conducting and the output is '0', i.e. close to ground potential, the gate of the load transistor consequently becomes more negative in relation to the source. This makes the load transistor conduct more, although this is exactly the opposite of what is required. To bring the output sufficiently close to ground potential, the driver transistor therefore has to be given a much greater conductivity, which is done by making its channel many times wider than that of the load transistor.

The need for this when a depletion-type transistor is used is much less. As soon as the voltage between the drain and the source is above the pinch-off voltage, the transistor is in a state of saturation [2]; it then has a higher resistance and behaves more or less as a current generator. The smaller channel width which is now adequate for the driver transistor leads to smaller capacitances and hence a higher switching speed. In addition, since the depletion-type transistor is effectively a constant-current generator, it provides the same amount of current for charging the capacitance connected to the output, even when there is a small voltage across it, i.e. when the output is negative ('1'); this again increases the switching speed.

Compared with the situation with enhancement-type transistors, the narrower channels that result when depletion-type transistors are used give a greater packing density than is possible with enhancement-type transistors, while the higher switching speeds yield speed-power product values two to ten times smaller: between 6 and 15 pJ.

D-MOS

Another factor, in addition to the parasitic capacitances, that affects the switching speed of a MOS transistor appreciably is the length of the channel: for a high switching speed the channel has to be short. The need for short channels led to the technologies known as D-MOS and V-MOS [7]. The main point of similarity between the two is that the length of the channel is determined by the thickness of a *P*-type layer.

In D-MOS technology (the D stands for 'double-diffused') two diffusions are made successively through the same mask opening. The first diffusion produces a *P*-type region round the source; the *N*-type channel is formed later in the same region. Before the second diffusion the drain connection is also exposed, and an N^+-type diffusion is then applied through both openings. The length l of the channel (*fig. 5*) is equal to the difference between the lateral depths of the *P* and N^+ diffusions; the length is determined solely by the diffusion process and can be controlled in much the same way as the base thickness in a bipolar transistor. Channels with any length between 0.4 and 2 μm can easily be made in this way.

Either N^- or P^- substrate can be used. In both cases an N^- drift space with a length l' will be formed (fig. 5).

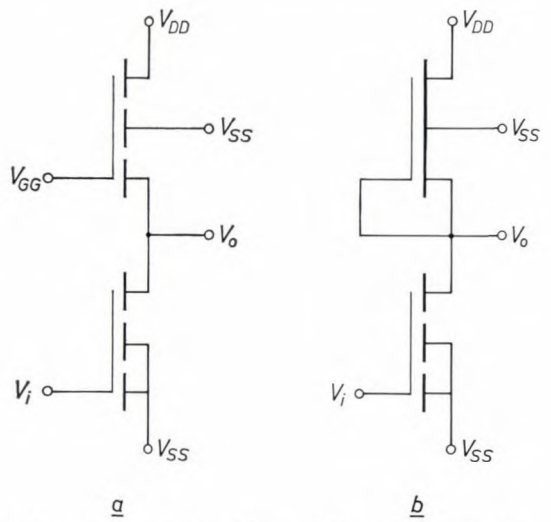


Fig. 4. Inverter circuit, *a*) with two MOS transistors of the enhancement type, *b*) with a depletion-type transistor as load. The shift in the threshold voltage of this depletion-MOS transistor was achieved by ion implantation.

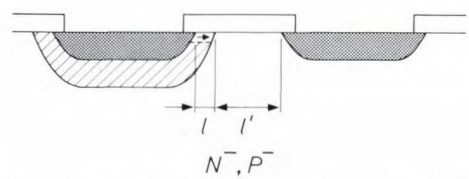


Fig. 5. Cross-section of a MOS transistor made with the aid of double diffusion (D-MOS technology). The length l of the channel equals the difference in lateral depth between the *P* and the N^+ diffusions. There is a drift region of length l' next to the channel.

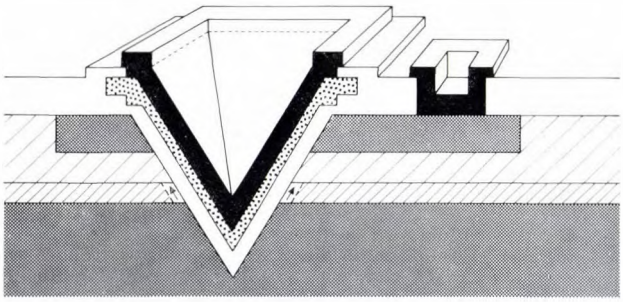


Fig. 6. Cross-section of a V-MOS transistor. The gate is deposited on the crystal faces of a V-shaped etch pit. The length of the channel (*arrows*) is determined by the thickness of the *P*-type layer.

It is clear why this will happen if an N^- substrate is used. If a P^- substrate is adopted, the lightly doped P^- region beneath the oxide of the gate is easily converted into material with the opposite type of conductivity by the positive oxide charge; inversion produces an N^--type region. In both cases the doping of this part of the transistor is at least an order of magnitude less than that in the channel, giving a sufficiently high breakdown voltage.

V-MOS

Fig. 6 shows a cross-section of a V-MOS transistor; the name is derived from the V-shaped profile. These transistors are made by growing first a *P* and then a P^- layer on an N^+ substrate. The substrate forms the common source electrode for all the transistors on the chip and is grounded. N^+ regions are then diffused locally into the surface to form the drains of the transistors. Next, V-shaped channels extending downwards to the N^+ substrate are made in the surface by anisotropic etching, a process in which material is preferentially removed along certain crystal planes. The walls of the channel form (111) planes. These walls are oxidized and then the gate is applied to them. The thickness of the *P* layer now determines the length of the channel. The lightly doped P^- region is inverted by the oxide charge to an *N* channel, which forms the drift region and gives a breakdown voltage of 12 V.

Speed-power products of 0.25 pJ for D-MOS and 3 pJ for V-MOS transistors have been reported [7]. Both configurations have the disadvantage that all the transistors have a common ground connection. They are also electrically asymmetrical, which means that the source and drain cannot be interchanged.

C-MOS

In a logic circuit there are advantages to be gained by connecting an *N*-channel and a *P*-channel MOS transistor, both of the enhancement type, in series. The same gate voltage is supplied to both transistors and, depending on the polarity, it makes either one or the other transistor conduct, so that current never flows in the steady state, apart from the leakage current, which is very small. The dissipation of this arrangement, which is called a complementary MOS circuit (C-MOS) is consequently very low in the steady state. *Fig. 7a* shows the circuit diagram of a NOR gate again, but this time in the form of a C-MOS circuit. It can be seen that four transistors are necessary. Fig. 7*b* shows the layout

[6] B. Crawford, Implanted depletion loads boost MOS array performance, Electronics **45**, No. 9, 85-90, 24 April 1972.
[7] T. P. Cauge, J. Kocsis, H. J. Sigg and G. D. Vendelin, Double-diffused MOS transistor achieves microwave gain, Electronics **44**, No. 4, 99-104, 15 Feb. 1971.
T. J. Rodgers and J. D. Meindl, VMOS: high-speed TTL compatible MOS logic, IEEE J. **SC-9**, 239-250, 1974.

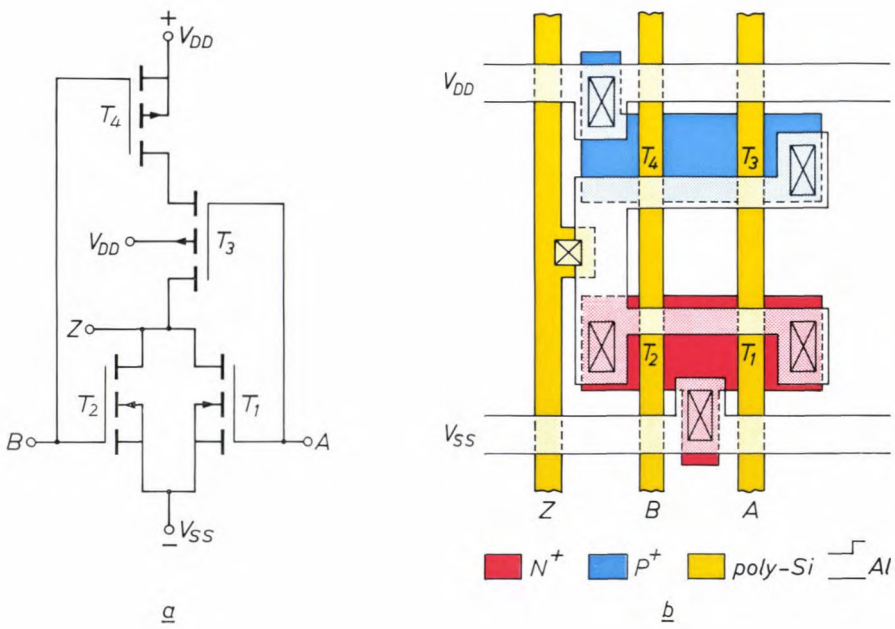


Fig. 7. NOR gate with complementary MOS transistors. Four transistors are required. *a*) Circuit diagram. *b*) Layout. Transistors T_1 and T_2 are formed in a *P*-type island but because of the LOCMOS technology employed this is not visible in the layout.

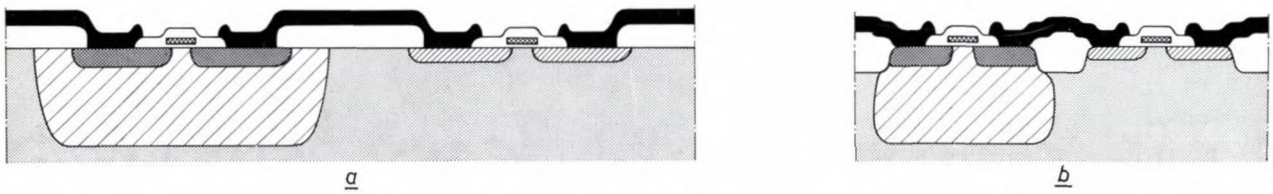


Fig. 8. Cross-section of a pair of complementary MOS transistors. The *N*-channel transistor is on the left, the *P*-channel one on the right. The gates are of polycrystalline silicon. *a*) Standard technology. *b*) LOCMOS technology. Thick local oxidation is used as insulation; less chip area is consequently required.

of the circuit (obtained using LOCMOS technology, see below).

Fig. 8a is a cross-section of a pair of complementary MOS transistors. A deep *P*-type island has to be diffused in the *N*-type substrate for the *N*-channel transistor. The low dissipation (and the high switching speed and noise insensitivity) of a complementary MOS configuration is thus obtained at the expense of a more complicated production process and extra space on the chip. (No extra space is required, however, if the complementary transistors are made by the LOCMOS technology (figs. 7*b* and 8*b*).)

LOCOS, LOCMOS, SOS and ESFI

There are several more recent technologies that enable the parasitic capacitances in MOS circuits to be reduced, thereby increasing the switching speed. LOCOS technology (LOCal Oxidation of Silicon) consists, as the name suggests, in oxidizing the silicon surface locally. This step produces a thick oxide, e.g. under the part of the gate that overlaps the source and drain diffusions [8]. This has the effect of reducing the capacitance of the overlaps.

If LOCOS technology is applied to complementary MOS circuits the result is referred to as LOCMOS [9]. As fig. 8 shows, this gives an appreciable saving of space.

The parasitic capacitances between the source and drain regions and the substrate are also important. They can be drastically limited by diffusing the MOS transistors in thin epitaxial layers on an insulating substrate consisting of sapphire or spinel; the *N*- and *P*-type regions are then diffused through to the substrate. This technology is called SOS [10] (Silicon On Sapphire or Silicon On Spinel) or ESFI [11] (Epitaxial Silicon Film on Insulators). A pair of complementary transistors that have been formed in this way in an epitaxial P^--type layer is illustrated in *fig. 9*. All the superfluous silicon is removed after the diffusion process. Since the substrate limits the inversion layer to the channel length between the source and the drain, SOS technology allows the capacitance to be reduced to about a twentieth of that in transistors in bulk silicon material.

It should of course be noted that the *N*-channel transistor in fig. 9 is of the enhancement type but that the P^+-P^--P^+ configuration conducts at 0 V; a sufficiently high positive voltage stops it conducting, because the entire P^- region is then depleted.

Since the leakage currents between the source and the drain are generally very small, the dissipation in the steady state is also very low, typically 1 nW. At intermediate frequencies and above the dissipation occurring during switching predominates and may increase up to say 0.25 mW. A value of 2.2 pJ for the speed-power product has been reported for SOS circuits at 50 MHz [10].

Bipolar integrated circuits

Schottky TTL

The family of bipolar integrated circuits most widely used at present is the one known as TTL (Transistor-Transistor Logic). TTL is based on a logic element consisting of a multiemitter transistor and an output stage formed by a number of transistors (see fig. 15 below). There are now standard TTL circuits, low-power TTL circuits and Schottky TTL circuits. The last of these combine the high speed of standard TTL circuits with the low dissipation of the second group [12]. The following values have been obtained for the speed-power product: 100 pJ for standard TTL; 30 pJ for low-power TTL; and 20 pJ for Schottky TTL.

Fig. 10 shows a cross-section of a Schottky transistor of a type that can be used in Schottky TTL. The base contact metal *B* extends over the *N*-type collector region. It thus forms a Schottky barrier, so that there is a diode in parallel with the base-collector junction. This Schottky diode has a lower forward voltage than

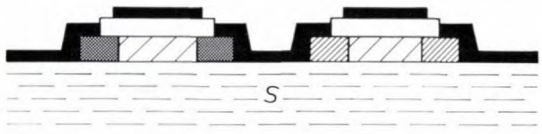


Fig. 9. Cross-section of a pair of complementary MOS transistors formed in a thin epitaxial single-crystal P^- silicon layer on an insulator *S*. There are no parasitic capacitances to a semiconducting substrate. The superfluous silicon between the transistors has been removed.

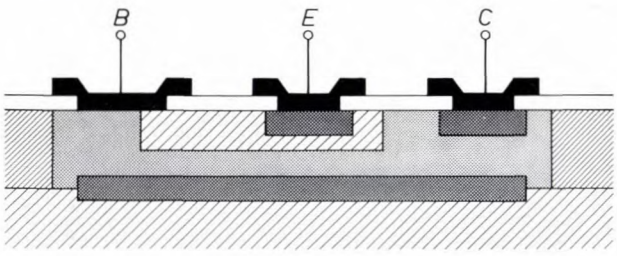


Fig. 10. Cross-section of a Schottky transistor. The base contact *B* extends over the *N*-type collector region and the two form a Schottky diode parallel to the base-collector junction. *E* emitter. *C* collector contact.

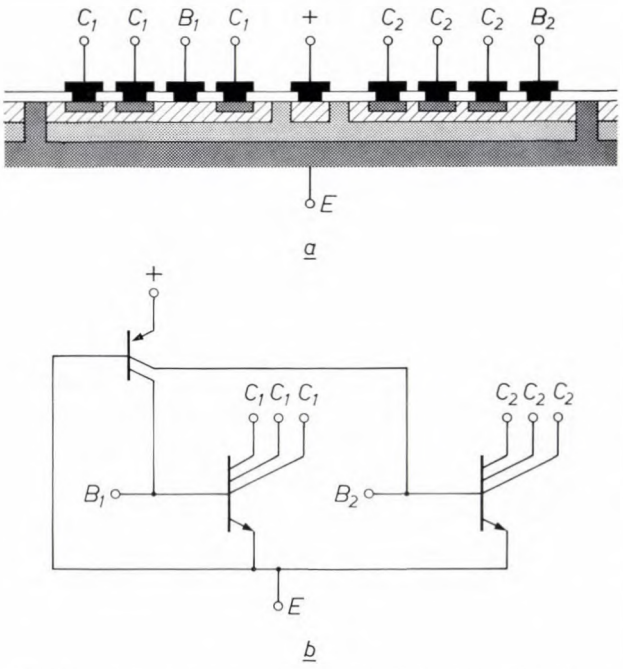


Fig. 11. *a*) Cross-section of two vertical *N-P-N* multicollector transistors based on the I^2L technique, combined with a lateral *P-N-P* transistor. The *N*-type region is the common emitter *E* of the vertical transistors; it is connected to the outside world via the N^+ substrate. The *P*-type region connected to + is the emitter of the lateral *P-N-P* transistor. *b*) Schematic representation.

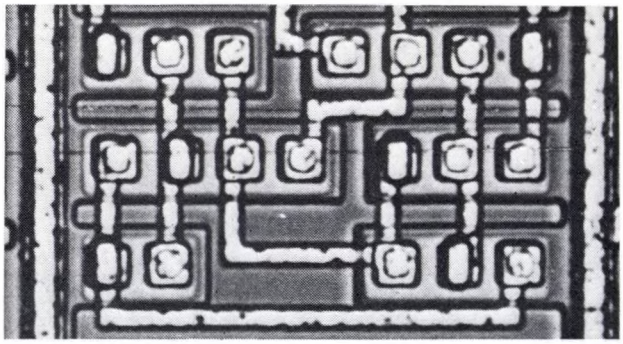


Fig. 12. Layout of I^2L circuit. The multicollector transistors are located on either side of the long, narrow *P*-type regions, the injectors, which are the common emitters of all the *P-N-P* current generators. The position of the base contact among the collectors can be chosen to suit the interconnection pattern. The transistors are surrounded on three sides by N^+ separation regions.

the base-collector junction and thus limits the voltage drop across this junction and offers an alternative path for surplus base current so that the transistor does not saturate. There is no charge storage in the transistor or the diode, so that the switching speed is increased and the dissipation reduced.

I^2L

A completely new approach to large-scale integration is seen in the technology known as I^2L (Integrated Injection Logic) [13], or MTL (Merged Transistor Logic) [14]. This logic will be discussed in somewhat greater detail because it is one of the most recent developments. In I^2L vertical *N-P-N* transistors, each with a number of collector regions and a common emitter, are combined with lateral *P-N-P* transistors that act as current generators. *Fig. 11* shows a cross-section through two of these multicollector transistors and their current generators, and a representation in circuit-diagram symbols. The number of collectors per transistor — three — has been chosen arbitrarily; in practice it will depend on the logic circuit. The circuit can be manufactured with existing standard processes: only five masks are needed: one each for the deep N^+ diffusion, the shallow *P*-type diffusion, the shallow N^+ diffusion, the contact openings and the interconnection pattern. Instead of the deep N^+ isolation diffusion, it is also possible to oxidize locally using the LOCOS technology, without increasing the number of masks required. I^2L is a new digital circuit technique, not a new technology.

Fig. 12 shows a typical layout for an I^2L circuit. Long, narrow *P*-type regions, the injectors, act as emitters for the *P-N-P* transistors of all the I^2L gates along both sides of this region. The arrangement of the base contact and the collectors is adapted to the intercon-

[8] J. A. Appels, E. Kooi, M. M. Paffen, J. J. H. Schatorjé and W. H. C. G. Verkuylen, Local oxidation of silicon and its application in semiconductor-device technology, Philips Res. Repts. **25**, 118-132, 1970.
J. A. Appels, H. Kalter and E. Kooi, Some problems of MOS technology, Philips tech. Rev. **31**, 225-236, 1970.
[9] B. B. M. Brandt, W. Steinmaier and A. J. Strachan, LOCMOS, a new technology for complementary MOS circuits, Philips tech. Rev. **34**, 19-23, 1974.
[10] E. J. Boleky and J. E. Meyer, High-performance low-power CMOS memories using silicon-on-sapphire technology, IEEE J. **SC-7**, 135-145, 1972.
See also J. A. van Nielen, M. J. J. Theunissen and J. A. Appels, MOS transistors in thin monocrystalline silicon layers, Philips tech. Rev. **31**, 271-275, 1970.
[11] K. Goser and M. Pomper, Five-transistor memory cells in ESFI MOS technology, IEEE J. **SC-8**, 324-326, 1973.
[12] R. N. Noyce, R. E. Bohn and H. T. Chua, Schottky diodes make IC scene, Electronics **42**, No. 15, 74-80, 21 July 1969.
[13] C. M. Hart and A. Slob, Integrated injection logic: a new approach to LSI, IEEE J. **SC-7**, 346-351, 1972.
C. M. Hart and A. Slob, Integrated injection logic (I^2L), Philips tech. Rev. **33**, 76-85, 1973.
[14] H. H. Berger and S. K. Wiedmann, Merged-transistor logic (MTL) — a low-cost bipolar logic concept, IEEE J. **SC-7**, 340-346, 1972.

nection patterns required. The logic functions are obtained by connecting the collectors.

I^2L does not require much area on the chip because there are no resistors, which take up more space than transistors. The usual isolation diffusion is also omitted. This compactness means that the parasitic capacitances are kept to a minimum. In addition, since the logic voltage swing is also small (0.6 to 0.7 V), the parasitic capacitances can be rapidly charged and discharged. The speed-power product can therefore be expected to be low in I^2L.

τP product measured on this register was 1.75 pJ per gate, with a dissipation of between 5 nW and 6 μW per gate (curve *5* in fig. 13). In another complicated circuit comprising 964 transistors on a single chip the speed-power product was 0.85 pJ per gate; the smallest details in this case were 7 μm. These τP values are about twenty times as small as those of Schottky TTL circuits.

The curve *lim* in fig. 13 indicates the 'theoretical limit' of the product τP; it is about 0.001 pJ if we base our calculations on the optimum values that the transistor parameters can attain, having regard to the

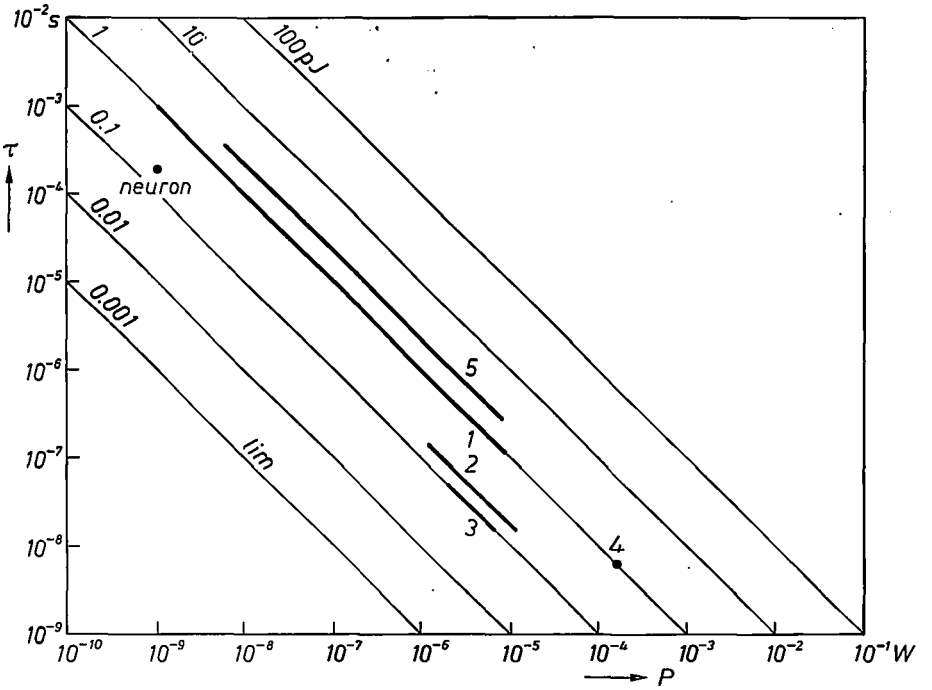


Fig. 13. Propagation delay τ, power dissipation P per transistor, and the product of the two for various I^2L configurations. *1* Minimum details 10 μm. *2* Minimum details 5 μm. *3* The *N-P-N* transistor has current applied to it from three sides by the injector. *4* Shallow-diffusion process. *5* Practical values in shift register, details 10 μm. *lim* Theoretical minimum. *neuron* A neuron in the human brain has a speed-power product of the same order as I^2L; it uses very little energy but is slow.

Fig. 13 shows a plot of the propagation delay time τ against the dissipation P (per gate) for five I^2L device geometries. A number of lines on which the speed-power product is constant have been drawn in for guidance. The product was found with the aid of a test circuit consisting of a closed-loop chain of inverter circuits. As measurements *1* to *4* show, values between 1 and 0.1 pJ were found.

In practice the delay time is actually longer because of the inclusion in the circuits of more inputs and outputs (fan-in and fan-out) and of wiring. To measure practical values a static shift register consisting of 820 I^2L transistors was made. Standard layout rules were adhered to and a standard fabrication process followed, with smallest details measuring 10 μm. The maximum permissible current densities, the required breakdown voltages and the minimum field-strengths required [15]. The neurons in the human brain have a speed-power product of 0.2 pJ, which is of the same order as the value attainable with I^2L. The packing density of neurons, however, is 1000 times higher, although their speed is fairly low.

To determine the chip area required by I^2L and various MOS systems for a 101-gate circuit (part of the control logic of a calculator) we prepared a comparative layout, using standard circuit modules (e.g. three-input NAND, four-input NAND, flipflop, etc.), which are all the same height within a single system, while their length depends on the logic function. The same layout rules were applied throughout; the basic elements were ar-

Table I. The chip area of a 101-gate circuit using I^2L and various MOS techniques.

Type of circuit	Area (mm^2)	Ratio to I^2L
I^2L	0.75	1
C-MOS	1.11	1.48
Static $\overline{\text{AND-OR}}$ gates with *P*-channel enhancement-type transistors	1.30	1.73
Static $\overline{\text{AND-OR}}$ gates with depletion-type transistors as load	1.17	1.56
Dynamic $\overline{\text{AND-OR}}$ gates with *P*-channel enhancement-type transistors	1.08	1.44
Dynamic NAND gates with *P*-channel enhancement-type transistors	1.41	1.88

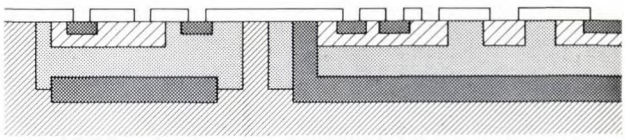


Fig. 14. Cross-section of an I^2L circuit in which the use of a seven-mask technique makes *N*-type islands with a P^+ insulation (*left*) available for other forms of logic or analog circuits.

layer has been grown, a deep P^+ diffusion can be made, followed by a deep N^+ diffusion. *N*-type islands surrounded by N^+ walls and other *N*-type islands surrounded by P^+ walls are then formed. I^2L circuits can be arranged in the first group of islands, and other logic and also analog circuits can be arranged in the second group.

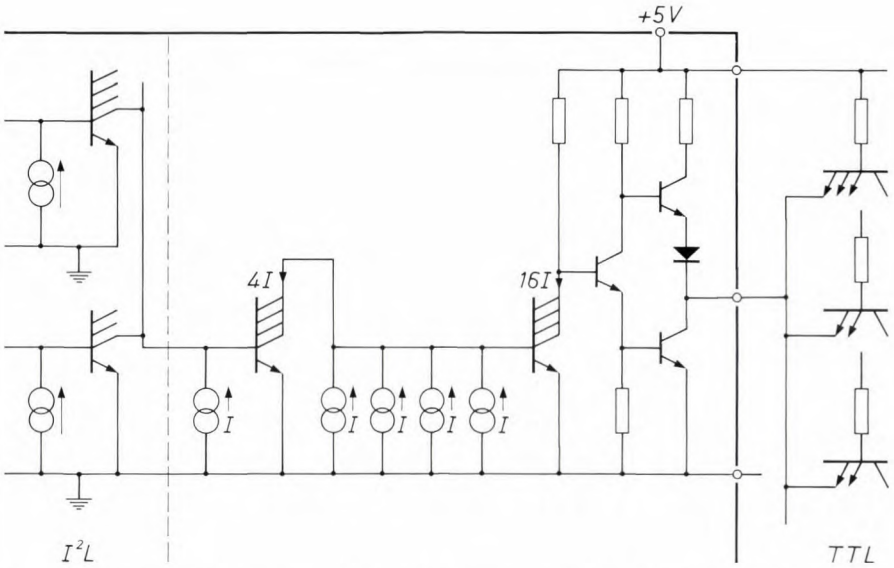


Fig. 15. Interface circuit integrated with I^2L for connection to TTL, for which larger currents are necessary. The interface circuit comprises a current amplifier followed by a totem-pole output stage.

ranged in rows with room for interconnections between them. Two-layer interconnections were permitted; local oxidation was used for isolation. The results are listed in *Table I*. The conclusion reached was that the area is smallest when I^2L is used [16]. The comparison was not extended to the speed of the circuits.

If I^2L technology is extended to incorporate a few extra steps, as in the case of the C-MOS technology, it becomes possible to combine other types of logic and analog circuits with I^2L on a single chip [17]. All kinds of interface circuits can then be integrated. *Fig. 14* shows a cross-section of a circuit made using a seven-mask technology. It starts from a P^+ substrate in which buried N^+ layers are diffused. After an epitaxial *N*-type

Fig. 15 is an example of an interface circuit for matching I^2L to TTL. An arrangement of this kind is necessary because I^2L works at a low current level, e.g. 1-10 μA per transistor, while in TTL the current is 1.9 mA per input. The interface circuit consists of two parts: a current amplifier followed by a totem-pole output stage built up from isolated transistors. This output stage is like that used in the basic logic circuit

[15] O. G. Folberth, Entwicklungstendenzen der logischen Bauelemente elektronischer Rechenanlagen, Elektron. Rechenanl. **8**, 175-180, 1966.

[16] N. C. de Troye, Integrated injection logic — present and future, IEEE J. **SC-9**, 206-211, 1974.

[17] C. M. Hart, A. Slob and H. E. J. Wulms, Bipolar LSI takes a new direction with integrated injection logic, Electronics **47**, No. 20, 111-118, 3 Oct. 1974.

in TTL. The current amplifier makes use of the fact that, up to a certain limit, the available output current is directly proportional to the N^+ collector area; current amplification can be obtained by cascading I^2L gates, with an increase in collector area at each stage. *Fig. 16* illustrates the layout of a current amplifier. The four parallel-connected collectors of the first transistor provide four times the base current for the next transistor, where there is another multiplication by a factor of four. The four current generators in parallel are realized by making the total injecting area in the second stage four times larger than in the first stage. The current density is the same everywhere.

Until the present there has been a continuous improvement in switching speed and power dissipation. These improvements have been made possible by new technologies and new circuit techniques. If we consider the circuit technique for a moment, it is clear that the reduction of the logic-voltage swing has given an improvement in switching speed. The minimum logic-voltage swings for modern digital circuits lie between 0.15 V and 0.7 V. Appreciable further reductions do not seem very likely in silicon integrated circuits, because the minimum swing is mainly determined by junction and diode voltages [18]. Increases in switching speed must therefore be expected to come from the development of new processes, e.g. very shallow diffusions, combined if necessary with dielectric insulation (LOCOS).

[18] An element with an extremely small voltage swing is at present in development. This is the Josephson element, which is based on tunnel effects and has a voltage swing of 2.25 mV. Propagation delay times of 175 ps have been measured, giving a speed-power product of only 0.004 pJ at an average dissipation of 22 μW. The element functions at cryogenic temperatures. See D. J. Herrell, Femtojoule Josephson logic gates, 1974 IEEE Int. Solid-State Circuits Conf. Dig. tech. Papers, pp. 116-117.

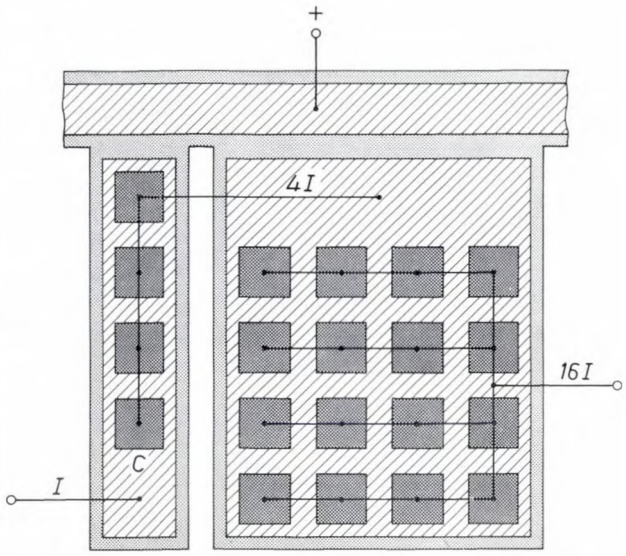


Fig. 16. Layout of a current amplifier in I^2L. Each collector C delivers a current I; the quadruple transistor delivers a current $16I$.

Summary. Larger and larger circuits are being integrated on individual single chips. This is only possible because the dissipation per transistor is being continually reduced. A number of new technologies are described which combine low dissipation with high switching speeds. A polycrystalline-silicon gate is incorporated in MOS transistors for this purpose, while ion implantation permits enhancement-type and depletion-type transistors to be combined in the same circuit. Higher switching speeds are achieved in D-MOS and V-MOS technologies by shortening the channel; the d.c. dissipation in C-MOS circuits is very low. SOS technology gives extremely low parasitic capacitances. The I^2L circuit technique represents an important development in bipolar transistor circuits, where it enables low dissipation, high switching speeds and high packing density to be achieved. Interface circuits for adaptation to other logic families such as TTL can be integrated on the same chip.

Recent scientific publications

These publications are contributed by staff of laboratories and plants which form part of or cooperate 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, 3 avenue Descartes, 94450 Limeil-Brévannes, 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*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.)	*N*

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, The Netherlands.

G. Armand (Centre d'Etudes Nucléaires de Saclay, Gif-sur-Yvette) **& J. B. Theeten:** Surface phonons on the (001) surface of a fcc crystal covered with a $C(2\times2)$ monolayer.
Phys. Rev. B 9, 3969-3984, 1974 (No. 10). *L*

H. Arnould-Nétillard & F. Rondelez: Electrohydrodynamic instabilities in cholesteric liquid crystals with negative dielectric anisotropy.
Mol. Cryst. liq. Cryst. **26**, 11-31, 1974 (No. 1/2). *L*

V. Belevitch & Y. Genin: Rectifier circuits for optimal frequency conversion.
Int. J. Circuit Theory & Appl. **2**, 111-131, 1974 (No. 2). *B*

C. Belouet: Growth of large single crystals and perfection problems connected with the gravitational field.
Proc. Symp. on Space processing, Frascati 1974, pp. 247-259. *L*

A. Bouwknegt, J. de Kok & J. A. W. de Kock (Philips Lighting Division, Eindhoven): Evolved gas detection apparatus for quantitative vacuum thermal analysis of alkaline-earth carbonates.
Thermochim. Acta **9**, 399-407, 1974 (No. 4).

M. J. G. Braken: Lassen en solderen van onderdelen uit roestvrij staal voor ultrahoogvacuümdoeleinden.
Lastechniek **40**, 211-214, 1974 (No. 8). *E*

F. J. A. den Broeder, G. D. Westerhout & K. H. J. Buschow: Influence of the stability of RCo_5-phases on their permanent magnetic properties.
Z. Metallk. **65**, 501-505, 1974 (No. 7). *E*

H. H. Brongersma: Threshold electron-impact spectroscopy.
Chemical spectroscopy and photochemistry in the vacuum-ultraviolet, Proc. Adv. Study Inst., Valmorin (Quebec) 1973 (NATO ASI Ser. C 8), pp. 287-304; 1974. *E*

H. H. Brongersma & W. J. Schouten: Ion scattering spectroscopy for surface analysis.
Acta Electronica **18**, 47-50, 1975 (No. 1). *E*

T. M. Bruton: The growth of single crystals of $Bi_4Ti_3O_{12}$.
Ferroelectrics 7, 259-260, 1974 (No. 1/2/3/4). *M*

M. R. Burgess, H. B. Haanstra & C. Schiller: Scanning electron microscopy and its application to semiconductor device assessment.
Acta Electronica **18**, 15-25, 1975 (No. 1). *M, E, L*

K. H. J. Buschow & A. Oppelt (Technische Hochschule Darmstadt): Magnetic and NMR studies of the CsCl type compounds $GdMg_{1-x}Zn_x$.
J. Physics F **4**, 1246-1255, 1974 (No. 8). *E*

K. L. Bye, P. W. Whipps, E. T. Keve & M. R. Josey (Mullard Ltd., Southampton): Alanine doped TGS/TGSe crystals for pyroelectric applications.
Ferroelectrics 7, 179-181, 1974 (No. 1/2/3/4). *M*

B. Cassanhiol, J. Cornet & D. Rossier: Transverse photoconductivity in Te-based glasses and application to arrays of light sensors, I. Recombination and trapping processes.
Amorphous and liquid semiconductors, Proc. 5th Int. Conf., Garmisch-Partenkirchen 1973, Vol. 1, pp. 571-579; 1974. *L*

M. Cathelin: Dispositif de prise de vues à accès aléatoire.
Onde électr. **54**, 349-352, 1974 (No. 7). *L*

T. A. C. M. Claasen, W. F. G. Mecklenbräuker & J. B. H. Peek: Some considerations on the implementation of digital systems for signal processing.
Philips Res. Repts. **30**, 73-84, 1975 (No. 1). *E*

J. B. Clegg & E. J. Millett: Analysis of semiconductor surfaces by spark source mass spectrometry.
Acta Electronica **18**, 27-32, 1975 (No. 1). *M*

L. W. Cooprider (Carnegie-Mellon University, Pittsburgh), **F. Heymans, P. J. Courtois & D. L. Parnas** (Technische Hochschule Darmstadt): Information streams sharing a finite buffer: Other solutions.
Inform. Process. Letters 3, 16-21, 1974 (No. 1). *B*

J. Cornet & D. Rossier: Structural ageing in bulk and thin films of vitreous selenium.
Amorphous and liquid semiconductors, Proc. 5th Int. Conf., Garmisch-Partenkirchen 1973, Vol. 1, pp. 267-275; 1974. *L*

A. Daniels: The Stirling engine, a promising power source for communications equipment.
Signal **28**, No. 9, 65-68, May/June 1974. *N*

P. Delsarte: Distance distribution of functions over Hamming spaces.
Philips Res. Repts. **30**, 1-8, 1975 (No. 1). *B*

D. den Engelsen & B. de Koning: Ellipsometric study of organic monolayers, Part 1. Condensed monolayers.
J. Chem. Soc., Faraday Trans. I **70**, 1603-1614, 1974 (No. 9). *E*

E. Fischer: The column of a dc arc burning in a metal halide atmosphere.
J. appl. Phys. **45**, 3365-3373, 1974 (No. 8). *A*

R. C. French: Binary transversal filters in data modems.
Radio & electronic Engr. **44**, 357-362, 1974 (No. 7). *M*

M. J. C. van Gemert: Evaluation of dielectric permittivity and conductivity by time domain spectroscopy. Mathematical analysis of Fellner-Feldegg's thin cell method.
J. chem. Phys. **60**, 3963-3974, 1974 (No. 10). *E*

M. J. C. van Gemert: Theoretical analysis of the lumped capacitance method in dielectric time domain spectroscopy.
Adv. mol. Relax. Proc. **6**, 123-137, 1974 (No. 2). *E*

C. J. Gerritsma & P. van Zanten: Electro-optical properties of imperfectly ordered planar cholesteric layers.
Liquid crystals and ordered fluids, Vol. 3, editors J. F. Johnson & R. S. Porter, publ. Plenum Press, New York 1974, pp. 437-448. *E*

K. Gijsbers, L. Bastings & R. van de Leest: The coulometric determination of trace levels of sulphur in gallium phosphide.
Analyst **99**, 376-380, 1974 (No. 1179). *E*

J. M. Goethals & H. C. A. van Tilborg (Eindhoven University of Technology): Uniformly packed codes.
Philips Res. Repts. **30**, 9-36, 1975 (No. 1). *B*

M. Goscianski: Contribution to the study of electrohydrodynamic instabilities in a nematic above a smectic-*A* or -*C*/nematic transition.
Philips Res. Repts. **30**, 37-55, 1975 (No. 1). *L*

S. H. Hagen & A. W. C. van Kemenade: Infrared luminescence in silicon carbide.
J. Luminescence **9**, 9-17, 1974 (No. 1). *E*

J. A. G. Halé & P. Saraga: Digital image processing.
Opto-electronics **6**, 333-348, 1974 (No. 5). *M*

H. W. Hanneman & L. J. M. Esser: Field and potential distributions in charge-transfer devices.
Philips Res. Repts. **30**, 56-72, 1975 (No. 1). *E*

P. Hansen: Anisotropy and magnetostriction of gallium-substituted yttrium iron garnet.
J. appl. Phys. **45**, 3638-3642, 1974 (No. 8). *H*

P. Hansen: Field dependence of the wall width and wall energy of compensation walls.
Appl. Phys. Letters **25**, 241-244, 1974 (No. 4). *H*

D. Hennings & H. Pomplun: Evaluation of lattice sites and valencies of manganese and iron in lead titanate ceramics by ESR and thermogravimetry.
Ferroelectrics **7**, 345, 1974 (No. 1/2/3/4). *A*

O. F. Hill & J. C. Brice: The composition of crystals of bismuth silicon oxide.
J. Mat. Sci. **9**, 1252-1254, 1974 (No. 8). *M*

W. K. Hofker (Philips Research Labs., Amsterdam Division), **H. W. Werner, D. P. Oosthoek** (Philips Res. Labs. Amsterdam) **& N. J. Koeman** (Philips Res. Labs. Amsterdam): Boron implantations in silicon: a comparison of charge carrier and boron concentration profiles.
Appl. Phys. **4**, 125-133, 1974 (No. 2). *E*

D. W. Hutton & W. H. Kautz (Stanford Research Institute, Menlo Park, Calif.): A simplified summation array for cellular logic modules.
IEEE Trans. **C-23**, 203-206, 1974 (No. 2). *E*

H. D. Jonker: Photomagnetic effects in single-crystal Ru-doped lithium ferrite.
J. Solid State Chem. **10**, 116-121, 1974 (No. 2). *E*

J. G. Kloosterboer: Differential kinetic analysis of cations — The continuous pH-variation method.
Anal. Chem. **46**, 1143-1145, 1974 (No. 8). *E*

E. Klotz, R. Linde & H. Weiss: A new method for deconvoluting coded aperture images of three-dimensional X-ray objects.
Optics Comm. **12**, 183-187, 1974 (No. 2). *H*

E. Klotz & H. Weiss: Three-dimensional coded aperture imaging using nonredundant point distributions.
Optics Comm. **11**, 368-372, 1974 (No. 4). *H*

J. de Kok, J. A. W. de Kock & A. Bouwknegt (Philips Lighting Division, Eindhoven): Dissociation of alkaline-earth triple carbonates in the pressure range 10^{-4}-10 torr.
Thermochim. Acta **9**, 409-418, 1974 (No. 4).

E. Kooi: MOS effects and silicon device technology.
Proc. 5th Int. Conf. on Solid state devices, Tokyo 1973 (Suppl. J. Japan Soc. Appl. Phys. **43**, 1974), pp. 427-436. *E*

E. Krätzig: Ultrasonic attenuation of current-carrying superconducting films.
Phys. Stat. sol. (a) **23**, K 143-145, 1974 (No. 2). *H*

G. Krekow & J. Schramm: A glow discharge for uniform charging of electrophotographic layers.
IEEE Trans. **ED-21**, 378-379, 1974 (No. 6). *H*

U. Krüger & R. Memming: Formation and reactions of long lived xanthene dye radicals: I. Photochemical studies on reactions of semireduced fluorescein, II. Photochemical reduction of rhodamine-B and fluorescein derivatives, III. Spectroelectrochemical investigations on the reduction of xanthene dyes.
Ber. Bunsen-Ges. phys. Chemie **78**, 670-678, 679-685, 685-692, 1974 (No. 7). *H*

P. R. Locher: Ratio of the magnetic dipole moments of ^{63}Cu and ^{65}Cu and the hyperfine-structure anomalies $^{63}\Delta^{65}$.
Phys. Rev. B **10**, 801-804, 1974 (No. 3). *E*

R. Metselaar & P. K. Larsen: High-temperature electrical properties of yttrium iron garnet under varying oxygen pressures.
Solid State Comm. **15**, 291-294, 1974 (No. 2). *E*

F. Meyer & G. J. Loyen: Ellipsometry applied to surface problems.
Acta Electronica **18**, 33-38, 1975 (No. 1). *E*

C. Michel & A. Sicignano: Observation of field-induced microstructure in β phase of 8/65/35 PLZT electro-optic ceramics.
Appl. Phys. Letters **24**, 559-562, 1974 (No. 11). *N*

A. R. Miedema, P. Bloembergen (University of Amsterdam), **J. H. P. Colpa** (Univ. Amsterdam), **F. W. Gorter, L. J. de Jongh** (Kamerlingh Onnes Laboratory, Leiden) **& L. Noordermeer:** Heisenberg ferromagnetism in two dimensions: an experimental study.
AIP Conf. Proc. **18**, 806-818, 1974 (Part 2). *E*

K. Pasedach: Ein Verfahren zur schnellen Approximation von Summen mit vielen Summanden.
Computing **12**, 323-331, 1974 (No. 4). *H*

G. W. Rathenau: Met z'n allen research doen voor de industrie.
Ned. T. Natuurk. **40**, Annex to No. 13, pp. V-VII, 1974. *E*

H. Rau: Homogeneity range of cubic high temperature cuprous sulfide (digenite).
J. Phys. Chem. Solids **35**, 1415-1424, 1974 (No. 10). *A*

P. J. Rijnierse & F. H. de Leeuw: Domain wall dynamics in low-loss garnet films.
AIP Conf. Proc. **18**, 199-212, 1974 (Part 1). *E*

T. E. Rozzi & J. H. C. van Heuven: Quasi-power algebraic invariants of linear networks.
1974 Eur. Conf. on Circuit theory and design, London (IEE Conf. Publ. No. 116), pp. 406-411. *E*

F. L. J. Sangster: The bucket brigade and other charge transfer devices.
Proc. 5th Int. Conf. on Solid state devices, Tokyo 1973 (Suppl. J. Japan Soc. Appl. Phys. **43**, 1974), pp. 275-281. *E*

G. B. Scott, D. E. Lacklison & J. L. Page: Absorption spectra of $Y_3Fe_5O_{12}$ (YIG) and $Y_3Ga_5O_{12}:Fe^{3+}$.
Phys. Rev. B **10**, 971-986, 1974 (No. 3). *M*

J. P. Scott: Design of high homogeneity split pair magnets with side access.
J. Physics E **7**, 574-578, 1974 (No. 7). *M*

J. L. Sommerdijk, J. M. P. J. Verstegen (Philips Lighting Division, Eindhoven) **& A. Bril:** Luminescence of $MeFX:Eu^{2+}$ (Me = Sr, Ba; X = Cl, Br).
J. Luminescence **8**, 502-506, 1974 (No. 6). *E*

A. L. N. Stevels & A. T. Vink: Fine structure in the low temperature luminescence of Zn_2SiO_4:Mn and $Mg_4Ta_2O_9$:Mn.
J. Luminescence **8**, 443-451, 1974 (No. 6). *E*

J. B. Theeten: Analysis of the surface crystallography of a solid: LEED and RHEED techniques.
Acta Electronica **18**, 39-45, 1975 (No. 1). *L*

G. E. Thomas & E. E. de Kluizenaar: Bombardment-induced light emission for the analysis of surfaces.
Acta Electronica **18**, 63-68, 1975 (No. 1). *E*

H. Vantilborgh: On the working set size and its normal approximation.
BIT **14**, 240-251, 1974 (No. 2). *B*

A. W. C. van Veelen, G. F. Kinds, H. S. Wijnands (all with Christelijk Instituut voor Blinden en Slechtzienden 'Bartimeus', Zeist) **& H. Bouma** (Institute for Perception Research, Eindhoven): Verkennend onderzoek naar de TV-loep als nieuw hulpmiddel voor visueel gehandicapte kinderen bij het lezen.
Ned. T. Geneesk. **118**, 664-671, 1974 (No. 18).

J. van der Veen & Th. C. J. M. Hegge: Niedrigschmelzende, flüssig-kristalline *o*-Hydroxyazo- und *o*-Hydroxyazoxybenzole.
Angew. Chemie **86**, 378-379, 1974 (No. 10). *E*

A. Venema: Analysis of surfaces.
Acta Electronica **18**, 9-13, 1975 (No. 1). *(Also in French.)* *E*

W. F. van der Weg (Philips Research Labs., Amsterdam Division), **D. Sigurd** (Research Institute for Physics, Stockholm) **& J. W. Mayer** (California Institute of Technology, Pasadena): Ion beam induced intermixing in the Pd/Si system.
Applications of ion beams to metals, editors S. T. Picraux, E. P. EerNisse & F. L. Vook, publ. Plenum Press, New York 1974, pp. 209-218.

H. W. Werner: Secondary ion mass spectrometry and its application to thin film and surface analysis.
Acta Electronica **18**, 51-62, 1975 (No. 1). *E*

J. van Wershoven & J. Vervest: Glazen O-ringkraan.
Polytechn. T. Werktuigbouw **29**, 467-469, 1974 (No. 14). *E*

P. W. Whipps & K. L. Bye: Polycrystalline tri-glycine sulphate-selenate materials for pyroelectric applications.
Ferroelectrics **7**, 183-185, 1974 (No. 1/2/3/4). *M*

H. P. J. Wijn: Bij het afscheid van G. W. Rathenau van de Philips Research.
Ned. T. Natuurk. **40**, Annex to No. 13, pp. III-IV, 1974. *E*

S. Wittekoek, T. J. A. Popma & J. M. Robertson: Faraday and Kerr effect of bismuth substituted iron garnets: applications and physical origin.
AIP Conf. Proc. **18**, 944-948, 1974 (Part 2). *E*

Contents of Philips Telecommunication Review **32**, No. 4, 1974:

W. Lulofs: A Dutch satellite for astronomical research (pp. 217-224).
G. David: Digital airborne equipment for AEROSAT (pp. 225-236).
C. G. den Hertog: DS 7 programmable controller system (pp. 238-245).
D. W. Singleton & H. H. A. Heesterbeek: Tunnel and motorway supervision system REYERS - Brussels (pp. 246-257).

Contents of Philips Telecommunication Review **33**, No. 1, 1975:

J. J. M. de Vries & A. Ros: Philips intersection controller type 86 AD 182 (pp. 1-7).
C. J. Krayenbrink & A. Vlaanderen: The Philips all-purpose traffic processor system (pp. 8-15).
A. Brinkman, J. Nijmeyer & H. van Poeteren: Application of the all-purpose traffic processor (pp. 16-33).
R. Gierer & H. Günther: Automation of the public mobile radiotelephone system (pp. 35-49).

Contents of Electronic Applications Bulletin **32**, No. 4, 1975:

T. S. Christian: Radiometry, photometry, and optoelectronic devices (pp. 125-136).
W. M. Dorn, W. Th. Hetterscheid, C. W. J. v. d. Pol & W. Smülders: SOAR — the basis for reliable power circuit design, Part I: Power transistor limitations and ratings (pp. 137-153).
T. S. Christian: Developments in permanent magnet materials. A report on the proceedings of the Third European Conference on Hard Magnetic Materials (pp. 154-157).
Nomogram for determining the inductance of iron-cored coils (pp. 158-159).

Contents of Mullard Technical Communications **13**, No. 125, 1975:

M. J. Prescott: 20AX 110° colour television: a brief outline (pp. 186-197).
R. E. F. Bugg: Full-wave thyristor power supply for television receivers (pp. 198-209).
J. R. Kinghorn: Flyback converters with current transformer drive (pp. 210-224).
FX3700 series transformer cores and coil formers for SMPS (pp. 225-228).

Contents of Mullard Technical Communications **13**, No. 126, 1975:

J. R. Kinghorn: Character rounding for alphanumeric video display (pp. 230-239).
A. Cinciura: A low-cost data display (pp. 240-246).
Image intensifiers: aids to night vision (pp. 247-251).
Photocouplers (pp. 252-254).
High-grade long-life aluminium electrolytic capacitors (pp. 255-256).

Contents of Valvo Berichte **19**, No. 2, 1974:

K. Ruschmeyer: Permanentmagnete und ihre Weiterentwicklung (pp. 29-40).
E. Scharrer, L. Grimm, K.-J. Hilke & U. Schlenker: Physikalisch-chemische und chemische Vorgänge bei der Herstellung von Farbbildröhren, 2. Teil (pp. 41-65).

Contents of Valvo Berichte **19**, No. 3, 1974:

W. Weltersbach & T. Rabeler: ZF-Verstärker-Baustein für Farbfernsehempfänger (pp. 67-96).
D. Höhne: PXE-Analogspeicher für fernbedienbare Rundfunk- und Fernsehempfänger (pp. 97-106).
J. Goerth: Stromspiegel-Schaltungen (pp. 107-114).
Dimensionierung von Transformatoren und Speicherdrosseln mit Ferroxcube-Kernen in Schaltnetzteilen (Durchflußwandlern) (pp. 115-132).

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 9

Magnetic anisotropy and magnetostriction of Ru and Ir ions in yttrium iron garnet

P. Hansen

Ever since magnetic garnets were first made, in the fifties, they have played a prominent role in fundamental research on the magnetism of the magnetic oxides. In practical applications, however, the garnets are far less widely used than the hexagonal and spinel ferrites. The potential applications of thin epitaxially grown garnet layers in memory systems have now stimulated a renewed interest in the garnets.

Philips Forschungslaboratorium Hamburg have for many years been making garnets for fundamental research and for use in microwave components [*]. In recent years they have been investigating the anisotropic and magnetostrictive properties of ions of the transition metals Ru and Ir in the garnet lattice. These properties can be used for 'tailoring' the properties of the garnet to the requirements of a particular application.

Introduction

Since the discovery of the non-conducting magnetic ferrites and garnets these materials have been used in many applications, including permanent magnets, computer-memory cores, magnetic tapes and microwave devices. Until now hexagonal ferrites and ferrites with spinel structure have mainly been used in these applications, with the exception of the microwave devices that now include garnet in 'YIG' components (YIG = Yttrium Iron Garnet). However, the most promising application of garnet materials has emerged only in recent years. It relates to memory systems based on the properties of cylindrical magnetic domains (magnetic bubbles) in thin garnet layers [1], or domains in garnet layers with 'compensation planes' [2], where the magnetization of the garnet is zero. To use garnets in devices it must be possible to 'tailor' such basic physical properties as saturation magnetization, Curie temperature, anisotropy, magnetostriction, Faraday rotation and optical absorption to meet the conditions for a particular application. This can be done within certain limits by varying the preparation technique and the composition of the garnet. The physical properties we shall consider more particularly here are the magnetic anisotropy and the magnetostriction. Since both properties involve a change in magnetic energy when the direction of magnetization has changed, they play an important part in applications that depend on rotations of magnetization. A widely used material parameter is the stress sensitivity, which is the ratio of the magnetostriction to the anisotropy. The stress sensitivity is required to be small to obtain a sufficiently square hysteresis loop and a high remanence ratio (the ratio of the remanent to the saturation magnetization).

In this article it will be shown how the anisotropy and magnetostrictive properties of garnets can be selectively controlled by substituting small amounts of strongly anisotropic ions without significantly affecting the other properties such as Curie temperature and saturation magnetization. The strongly anisotropic 3d transition-metal ions Co^{2+} and Fe^{2+} are generally used for this. In the past few years we have made an extensive study of the 4d transition-metal ion Ru^{3+} and the 5d transition-metal ion Ir^{4+}. These ions were expected to be even more anisotropic than the 3d ions. In the following the theory for the anisotropy of these ions will be presented and the results of the theory will be compared with experimental results obtained for single crystals of yttrium iron garnet doped with small quan-

Dr P. Hansen is with Philips GmbH Forschungslaboratorium Hamburg, Hamburg, West Germany.

[*] See for example P. Röschmann, Philips tech. Rev. 32, 322, 1971.

[1] L. J. Varnerin, IEEE Trans. MAG-8, 329, 1972.

[2] J.-P. Krumme and P. Hansen, Appl. Phys. Lett. 23, 576, 1973.

tities of Ru and Ir. The article will conclude with an example showing how the anisotropic properties of these ions can be put to practical use.

Phenomenology

In single crystals of magnetic materials it is always found that the magnetization orients itself in a preferred crystallographic direction, called the 'easy direction'. The magnetic energy apparently depends on the direction of the magnetization with respect to the crystallographic axes, and is at a minimum when the magnetization is in the easy direction. The intrinsic anisotropic contribution to the energy, i.e. disregarding external sources of anisotropic energy such as the shape effect, is called the magnetocrystalline energy.

When a magnetic crystal is magnetized, the dimensions of the crystal are observed to change from the dimensions in the demagnetized state. This effect, called magnetostriction, implies a relation between magnetic and elastic properties. The well known magnetostriction constants λ_{100} and λ_{111} are defined as the relative change in length $\Delta L/L$ in the direction of the applied magnetic field when a saturating magnetic field is applied in the [100] or [111] directions, as appropriate.

The anisotropic part of the free energy F of a magnetically saturated single crystal is usually written as the sum of the anisotropy energy $F_K(\alpha)$ in the unstrained state and the magnetoelastic energy $F_{me}(\alpha,\varepsilon)$, usually much smaller, which is proportional to the strains. The tensor ε is the strain tensor and α is a unit vector parallel to the direction of magnetization. The magnetocrystalline energy is usually expressed as a series of powers of the direction cosines α_1, α_2 and α_3, the components of α along the crystal axes. For crystals with cubic symmetry it follows from symmetry considerations that

$$F_K(\alpha) = K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) + K_2\alpha_1^2\alpha_2^2\alpha_3^2 + \ldots, \quad (1)$$

where K_1 and K_2 are the well known first and second anisotropy constants. A power series of this type is useful because it may often be sufficient to consider only the first two terms. The sign of K_1 then determines the easy direction of magnetization, provided K_2 is not too large, which is often the case. With K_1 positive the easy directions are then the $\langle 001\rangle$ directions (cubic axes) and with K_1 negative the $\langle 111\rangle$ directions (body diagonals).

The term $F_{me}(\alpha,\varepsilon)$ represents, as we have seen, the magnetoelastic energy. The most general series can also be found for this term, again by considering the symmetry. For cubic crystals it is:

$$F_{me}(\alpha,\varepsilon) = b_1\sum_{i=1}^{3}\alpha_i^2\varepsilon_{ii} + b_2\sum_{i<k-1}^{3}\alpha_i\alpha_k\varepsilon_{ik} + \ldots . \quad (2)$$

The constants b_1 and b_2 are the magnetoelastic constants. They are related to the magnetostriction constants λ_{100} and λ_{111}:

$$\left.\begin{aligned}\lambda_{100} &= -\frac{2}{3}\frac{b_1}{C_{11}-C_{12}},\\ \lambda_{111} &= -\frac{1}{3}\frac{b_2}{C_{44}}.\end{aligned}\right\} \quad (3)$$

The constants C_{11}, C_{12} and C_{44} are the elastic constants (for YIG at room temperature $\lambda_{100} = -0.41\times10^{-12}\,b_1$ and $\lambda_{111} = -0.44\times10^{-12}\,b_2$). In many cases only the first two terms of the series expansion (2) need be considered.

The garnet structure

Before discussing the anisotropic and magnetostrictive properties of garnets we shall first take a look at their crystal structure. The composition of the garnets is $\{C_3\}(B_3)[A_2]O_{12}$. The different sets of brackets indicate that the ions occupy three crystallographically different sites. The trivalent C ions are dodecahedrally surrounded by eight oxygen ions. These dodecahedral sites are occupied by large ions such as yttrium ions or rare-earth ions. The A ions are octahedrally surrounded by six oxygen ions and form a body-centred cubic lattice (see *fig. 1*). The B ions are tetrahedrally surrounded by four oxygen ions. The octahedral and tetrahedral sites are occupied by smaller ions, such as ions of the transition metals or diamagnetic ions like Ga^{3+}, Al^{3+} and In^{3+}. The best-known magnetic garnet is yttrium iron garnet with the composition $Y_3Fe_5O_{12}$. Garnet crystals of high quality can be grown at about 1100 °C from a solution of the metal oxides [4] [5].

Although the structure of garnet is cubic, there are local deviations from cubic symmetry. It can be seen in fig. 1 that there are two sets of three Y^{3+} ions near the Fe^{3+} ion A at an octahedral site. The local symmetry at the site of the ion A is therefore uniaxial. This local threefold axis of symmetry coincides with one of the body diagonals of the body-centred lattice, the crystallographic $\langle 111\rangle$ directions. In the right-hand part of fig. 1 it can be seen that the garnet structure contains equal numbers of octahedral sites with each of the $\langle 111\rangle$ directions as the local axis. The tetrahedral sites in the garnet have a twofold axis parallel to one of the three $\langle 100\rangle$ directions.

Microscopic origin: the single-ion mechanism

In the foregoing the phenomenological expressions (1) and (2) have been introduced for the anisotropic part of the free energy of a cubic magnetic crystal. The main source of this energy in the cubic ferrites and garnets has a single-ion character [6] [7], that is to say

the energy of a single ion depends on the direction of the magnetic moment (spin moment) of the ion with respect to the crystallographic axes. This originates from 'spin-orbit coupling'. The orbitals of the electrons of the free ion are affected by the electric fields produced by the surrounding ions in the crystal lattice (the crystal field [8]), thus determining the magnitude and direction of the orbital momentum. As a result of the spin-orbit coupling the energy of an ion will depend on the direction of the spin with respect to the orbital momentum, and hence with respect to the crystallographic axes. The crystal field, and likewise the electron orbits, are sensitive to strains in the crystal. Spin-orbit coupling therefore also accounts for the dependence of the free energy of the ion upon strains present in the crystal, the magnetoelastic energy.

An obvious source of magnetocrystalline energy is the classical dipole-dipole interaction between the magnetic moments of the ions, which are generally arranged anisotropically in the crystal lattice. It is found, though, that in cubic crystals the contribution made by this interaction to the anisotropy energy is much smaller than the contribution of the single-ion anisotropy of the ions [6]. This does not apply to the magnetoelastic energy, however. In the presence of a strain the crystal symmetry is lower than cubic, and therefore the dipole-dipole interaction makes a substantial contribution to the magnetoelastic energy [9].

The important consequence of the dominant role of single-ion anisotropy is that the anisotropy energy of a crystal is simply the sum of the free energy of the magnetic ions:

$$F_K = \sum_i F_K^{(i)}. \qquad (4)$$

Experimental confirmation of this relation is given in *fig. 2*, which shows how the values of K_1, b_1 and b_2 in YIG depend on the extent to which diamagnetic ions are substituted for Fe^{3+} (Ga [10] and In [11] [12]). The In^{3+} ions, for example, substitute for Fe^{3+} ions at

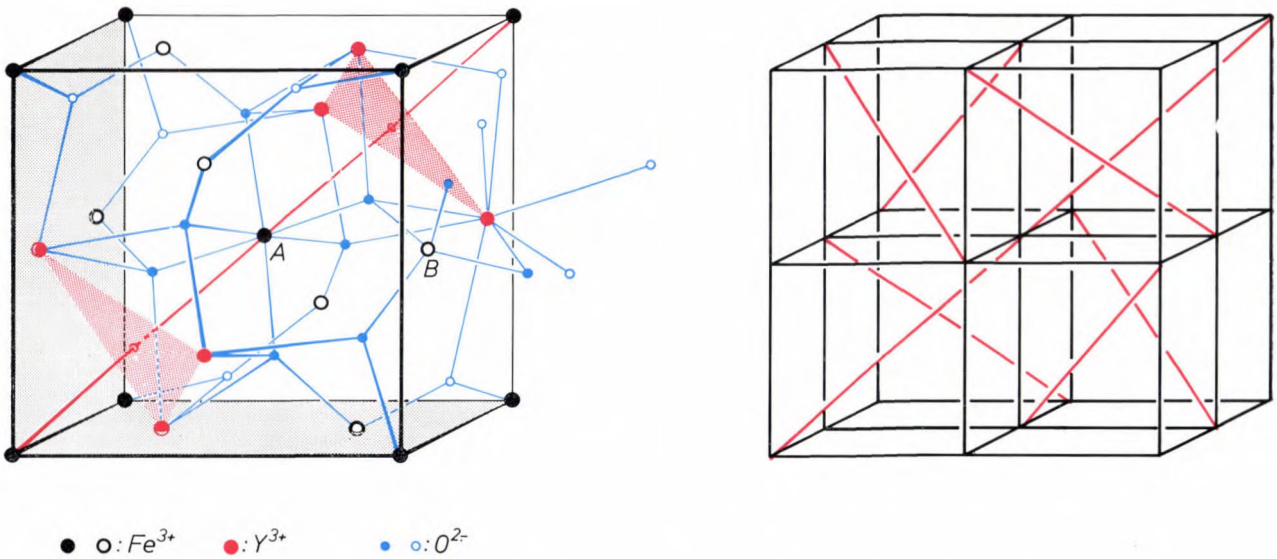


Fig. 1. Crystal structure of $Y_3Fe_5O_{12}$ [3] (yttrium iron garnet or YIG). The diagram on the right shows a complete unit cell; the diagram on the left gives the details of an eighth part of the cell. The ordering of the ions in the complete cell is obtained when the octant on the left is made to overlap the octants on the right-hand figure so that the red diagonals coincide.

The Fe^{3+} ions, shown as black points, are surrounded octahedrally by six oxygen ions and form a body-centred cubic lattice. The Fe^{3+} ions, shown as black circles, are surrounded tetrahedrally by four oxygen ions. The sets of four and six oxygen ions surrounding the two Fe^{3+} ions *A* and *B* are shown as blue points, the other oxygen ions as blue circles.

The two groups of three Y^{3+} ions, linked by a red plane, give rise to threefold symmetry at the location of the Fe^{3+} ion *A*. The local threefold axis coincides with the red diagonal. The diagram on the right shows that the local axis of an octahedral site has four possible directions, so that four different kinds of octahedral site are possible.

[3] From P. P. Ewald and C. Hermann, Strukturbericht 1913-1926, p. 364.
[4] W. Tolksdorf, J. Crystal Growth **3/4**, 463, 1968.
[5] W. Tolksdorf, Acta Electronica **17**, 57, 1974.
[6] K. Yosida and M. Tachiki, Progr. theor. Phys. **17**, 331, 1957.
[7] W. P. Wolf, Phys. Rev. **108**, 1152, 1957.
[8] P. F. Bongers, Philips tech. Rev. **28**, 13, 1967.
[9] G. A. Petrakovskii, E. A. Petrakovskaia and E. M. Smokotin, IEEE Trans. **MAG-8**, 476, 1972.
[10] P. Hansen, J. appl. Phys. **45**, 3638, 1974.
[11] B. E. Rubinshtein and G. M. Galaktionova, Sov. Phys. Solid State **9**, 2124, 1968.
[12] R. F. Pearson and A. D. Annis, J. appl. Phys. **39**, 1338, 1968.

octahedral sites. Since the In ions are diamagnetic, the effect of this substitution is to decrease the anisotropy energy by the single-ion contribution of one Fe^{3+} ion for every Fe^{3+} ion replaced by an In^{3+} ion. This linear behaviour is indeed found in fig. 2. In the case of YIG doped with Ga it is seen that K_1 and b_2 also show the expected linear behaviour, but b_1 behaves differently. The reason for this is that the magnetoelastic constant b_1 contains a significant contribution from dipolar energy (dipole-dipole interaction) [9]. It also appears from fig. 2 that a high degree of substitution by diamagnetic ions is necessary to produce a perceptible change in anisotropy and magnetostriction. But such a substitution will also significantly affect other magnetic properties, such as magnetization and Néel temperature.

It is also possible, however, to change the anisotropy and magnetostriction selectively without affecting other properties. This can be done by doping the garnet with ions that make a much stronger contribution to the anisotropy and magnetostriction than the Fe^{3+} ions, such as Fe^{2+} and Co^{2+} ions. These ions have an electronic ground state with an orbital momentum, unlike Fe^{3+} ions, which have no orbital momentum, since the d shell is only half-filled. The single-ion anisotropy and magnetostriction of these ions is therefore more than two orders of magnitude greater than for the

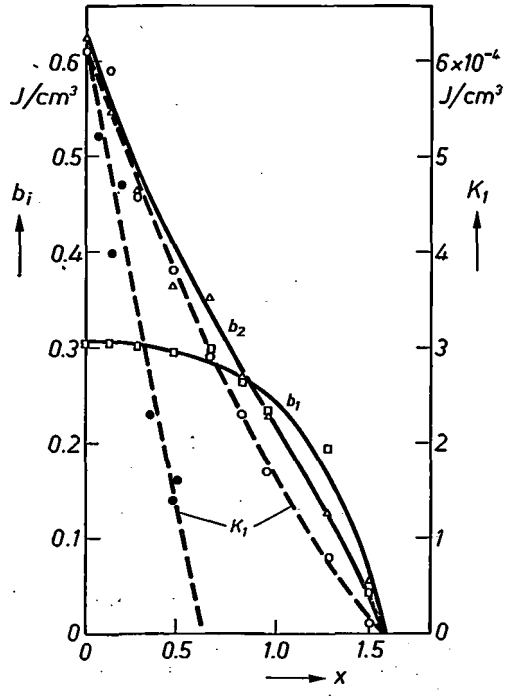


Fig. 2. The anisotropy constant K_1 (right-hand scale) and the magnetoelastic constants b_1 and b_2 (left-hand scale) at room temperature as a function of the quantity x of diamagnetic ions (Me) in $Y_3Fe_{5-x}Me_xO_{12}$. Me stands for In (points) [11] [12], which replace Fe^{3+} at octahedral sites, or Ga (circles) [10], which mainly replaces Fe^{3+} at tetrahedral sites.

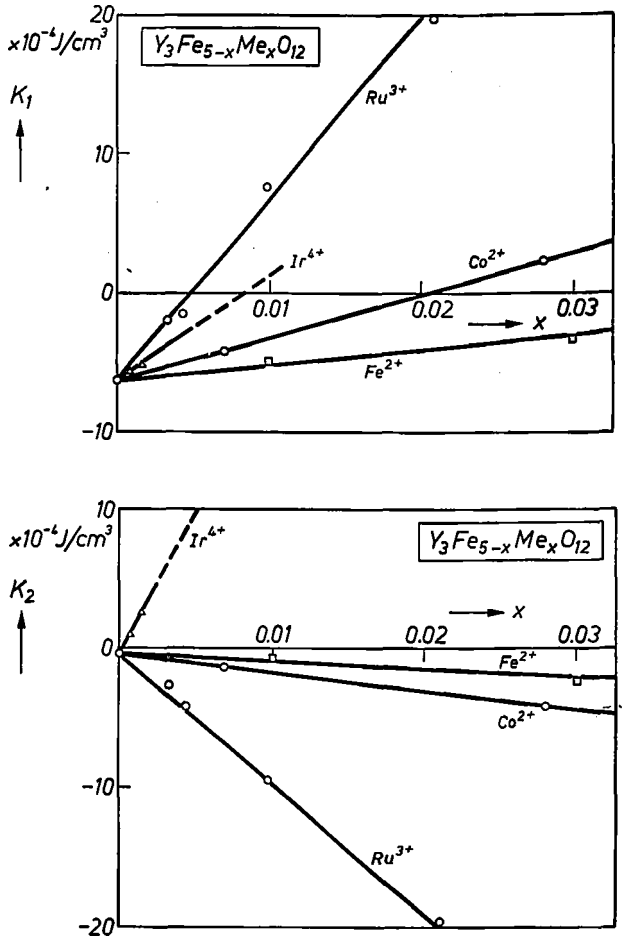


Fig. 3. The anisotropy constants K_1 and K_2 of $Y_3Fe_{5-x}Me_xO_{12}$ at room temperature as a function of the concentration of strongly anisotropic Me ions. Me stands for Fe^{2+} [14], Co^{2+} [17], Ru^{3+} [19-21] or Ir^{4+} [19] [24]. In the case of Fe^{2+}, Co^{2+} and Ir^{4+} charge-compensating ions were added (Si^{4+}, Ge^{4+} or Ca^{2+}, Zn^{2+}).

Fe^{3+} ions. Even very small concentrations of these ions in garnets can have a very strong effect on the values of K and b without significantly affecting other properties. The change in the magnetoelastic energy, ΔF_{me}, produced by doping garnets with such ions is given by an equation similar to eq. (4), since the dipole-dipole contribution is scarcely affected by such small substitutions whereas the single-ion contribution of the dopants is very large. Thus

$$\Delta F_{me} = \sum_i F_{me}^{(i)}, \qquad (5)$$

where the summation is over the dopant ions only.

During recent years we have extensively studied some alternative ions that could provide a selective control of the anisotropy and magnetostriction of garnets. These ions are the transition-metal ions with incomplete 4d or 5d shells: Ru^{3+} ($4d^5$) and Ir^{4+} ($5d^5$). The spin-orbit coupling of these ions is very much greater than that

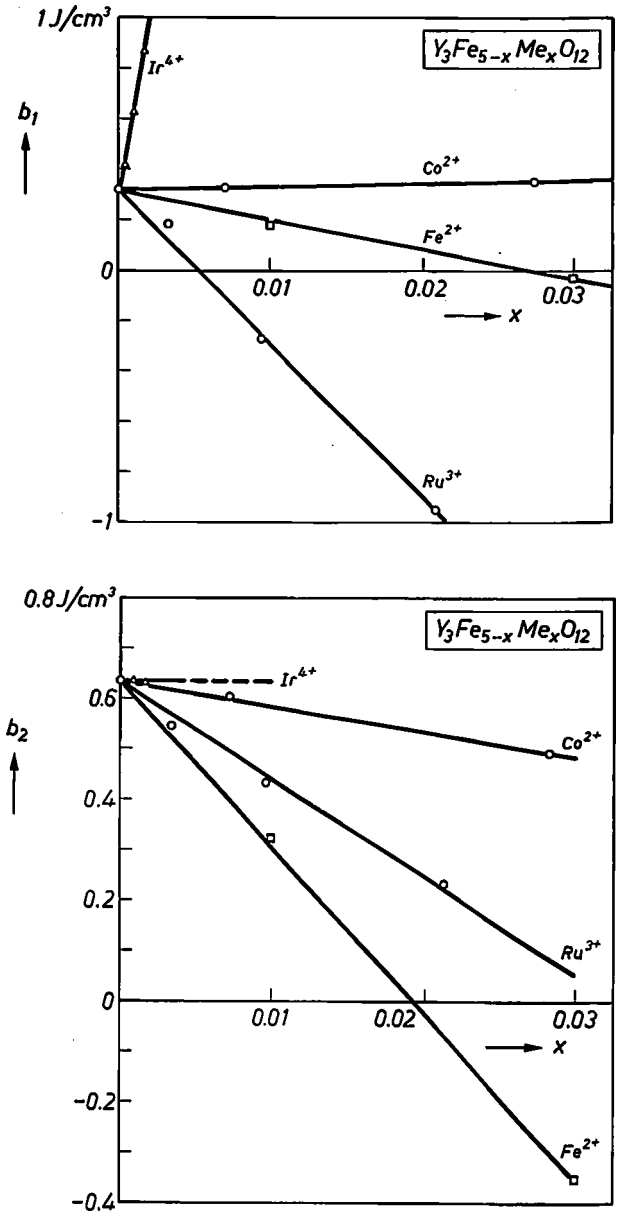


Fig. 4. The magnetoelastic constants b_1 and b_2 of $Y_3Fe_{5-x}Me_xO_{12}$ at room temperature as a function of the concentration of strongly anisotropic Me ions. Me stands for Fe^{2+} [14], Co^{2+} [17] Ru^{3+} [23] or Ir^{4+} [24]. The magnetostriction constants λ_{100} and λ_{111} are related to b_1 and b_2 by $\lambda_{100} = -0.41 \times 10^{-12} b_1$ and $\lambda_{111} = -0.44 \times 10^{-12} b_2$.

of the conventional 3d transition-metal ions, such as Fe^{2+} ($3d^6$) and Co^{2+} ($3d^7$). The plots in *fig. 3* and *fig. 4* show how K_1, K_2, b_1 and b_2 vary at room temperature for YIG doped with Fe^{2+} [13] [14], Co^{2+} [15-17], Ru^{3+} [18-23] and Ir^{4+} [19] [24]. The figures show clearly that even small concentrations of these ions can radically change the magnetostriction and anisotropy of YIG. It can also be seen that Ru^{3+} and Ir^{4+} have a greater effect than Fe^{2+} and Co^{2+}, a consequence of the stronger spin-orbit coupling. An important advantage of these ions, which we have observed, is that they produce hardly any side-effects, unlike the Fe^{2+} ions, which increase the electrical conductivity, and the Co^{2+} ions, which cause after-effects. The availability of a number of strongly anisotropic ions with different contributions to the magnetocrystalline energy and magnetostriction of the garnet makes it possible to control the anisotropy and magnetostriction more or less separately by suitably combined doping.

Ferromagnetic-resonance method for determining the anisotropy and the magnetostriction

The uniform magnetization of a magnetized single crystal can be excited in a uniform mode of precession, the ferromagnetic-resonance mode, by an r.f. magnetic field. The frequency of this resonance effect depends on the stiffness with which the magnetization is held to its equilibrium direction, and can therefore be expressed in terms of the second derivatives of the free energy in the equilibrium direction, as given by the generalized resonance equation (Smit's equation [25]):

$$\omega = \frac{\gamma}{M \sin\theta} \sqrt{\frac{\partial^2 F}{\partial \phi^2}\frac{\partial^2 F}{\partial \theta^2} - \left(\frac{\partial^2 F}{\partial \phi \partial \theta}\right)^2}. \qquad (6)$$

Here ω is the angular frequency, M the saturation magnetization and γ the gyromagnetic ratio. The z-axis is taken parallel to the equilibrium magnetization direction, while θ and ϕ are polar coordinates of α. Only the anisotropic terms of the free energy F contribute to the expression for the resonant frequency. These terms are the magnetostatic energy of M in the externally applied field, the anisotropy energy F_K and the magnetoelastic energy F_{me}. The demagnetization energy of an arbitrarily shaped anisotropic crystal is also anisotropic. Experiments are therefore usually carried out on crystals machined to a spherical shape, so that the demagnetization energy is isotropic.

Expressing F_K as a power series as in eq. (1) and retaining only the K_1 and K_2 terms, and considering for the present the unstrained case, we can calculate from eq. (6) the resonance conditions for the case where the

[13] T. S. Hartwick and J. Smit, J. appl. Phys. **40**, 3995, 1969.
[14] P. Hansen, W. Tolksdorf and J. Schuldt, J. appl. Phys. **43**, 4740, 1972.
[15] T. Okada, H. Sekizawa and S. Iida, J. Phys. Soc. Japan **18**, 981, 1963.
[16] M. D. Sturge, E. M. Gyorgy, R. C. LeCraw and J. P. Remeika, Phys. Rev. **180**, 413, 1969.
[17] G. F. Dionne and J. B. Goodenough, Mat. Res. Bull. **7**, 749, 1972.
[18] R. Krishnan, Phys. Stat. sol. (a) **1**, K17, 1970.
[19] P. Hansen, Philips Res. Repts. Suppl. 1970, No. 7.
[20] P. Hansen, Phys. Rev. B **3**, 862, 1971.
[21] P. Hansen, Phys. Stat. sol. (b) **47**, 565, 1971.
[22] P. Hansen, Phys. Rev. B **5**, 3737, 1972.
[23] P. Hansen, Phys. Rev. B **8**, 246, 1973.
[24] P. Hansen, J. Schuldt and W. Tolksdorf, Phys. Rev. B **8**, 4274, 1973.
[25] J. Smit and H. P. J. Wijn, Ferrites, Philips Technical Library, Eindhoven 1959.

external field is applied along one of the principal axes of a cubic crystal [25]:

$$\left.\begin{aligned}\frac{\omega}{\gamma} &= H_0[001] + 2\frac{K_1}{M},\\ \frac{\omega}{\gamma} &= H_0[111] - \frac{4}{3}\frac{K_1}{M} - \frac{4}{9}\frac{K_2}{M},\\ \frac{\omega}{\gamma} &= \sqrt{\left(H_0[110] - 2\frac{K_1}{M}\right)\left(H_0[110] + \frac{K_1}{M} + \frac{1}{2}\frac{K_2}{M}\right)}.\end{aligned}\right\} \quad (7)$$

The experiments are usually performed at a fixed frequency and the magnetic field is measured at which the resonance condition is satisfied (the resonance field). Using equations (7) we can then deduce the three unknowns, γ, K_1 and K_2, from the magnitude of the resonance field in each of the three principal directions.

The resonance method can also be used for determining the magnetostriction constants; a uniform stress then has to be applied to the sphere. The anisotropic part of F_{me} induced by this stress is given by eq. (2) and contributes to the resonance condition in accordance with eq. (6). In the experiment the resonance field is found to undergo a shift δH_0 proportional to the applied stress. A stress τ applied in the $[1\bar{1}0]$ direction causes a resonance-field shift $\delta H_0[hkl]$, measured with the magnetic field in each of the principal directions in the $(1\bar{1}0)$ plane, from which the magnetostriction constants can be determined by using the relations [26-28]

$$\left.\begin{aligned}\lambda_{100} &= -\frac{2}{3}\frac{M}{\tau}\delta H_0[001],\\ \lambda_{111} &= -\frac{2}{3}\frac{M}{\tau}\delta H_0[111].\end{aligned}\right\} \quad (8)$$

An important advantage of the resonance method is that it only requires very small single crystals. The diameter of the spheres used is usually about 0.7 mm. Ferrite and garnet single crystals of this size can readily be obtained.

Mechanism of the single-ion anisotropy of strongly anisotropic ions

The electron states of ions in a crystal

In an earlier section it was shown that the macroscopic anisotropy energy originates from the single-ion anisotropy of the magnetic ions. The key to the understanding of single-ion anisotropy lies in the electronic states of the ion, the spin S of the ion and the orbital momentum L, linked by the spin-orbit coupling. In a crystal the states of the electrons of an ion are determined by the electron configuration of the ion and the strength and symmetry of the electric fields due to the neighbouring ions [8]. The strength of these 'crystal fields' as experienced by a substituted ion in a crystal depends on the nature of the ions in the host lattice and on that of the substituted ion. Thus, for the 4d and 5d transition-metal ions the crystal field in garnet is much stronger than for the 3d ions. It is even strong enough to overcome the repulsion between the electrons, which normally forces them into states with parallel spins (the well known Hund's rule). The electrons of the 4d and 5d ions at octahedral sites in the garnet lattice therefore occupy states with opposite spins, which means that these ions have a 'low-spin' ground state (spin $S = 1/2$). A 3d ion like Fe^{2+} at the same site has a 'high-spin' ground state ($S = 2$).

The crystal field in the garnet lattice is mainly of cubic symmetry. A characteristic of the strongly anisotropic ions is that this crystal field does not completely quench the orbital momentum of the electrons, and consequently the spin-orbit coupling is able to make the ground state strongly anisotropic. The local deviation from cubic symmetry (see fig. 1) adds a low-symmetry component to the crystal field, which is about an order of magnitude smaller than the cubic component. We shall see presently that this plays a key role in the occurrence of single-ion anisotropy.

Simple physical explanation of the single-ion anisotropy of strongly anisotropic ions

To understand the anisotropic properties of a strongly anisotropic ion we must consider three interactions, as illustrated schematically in *fig. 5* [29]. The first is the interaction of the electrons with the local low-symmetry crystal field, which binds the orbital momentum L to the local axis. The second is the exchange interaction between the spin S of the ion and the spins of neighbouring magnetic ions. A model commonly used to describe this interaction is the molecular-field model, in which the interaction is represented by an effective magnetic field H_{exch}, acting on the spin S of the ion. This 'molecular field' is parallel to the magnetization direction α and proportional to the saturation magnetization M. The third interaction is the spin-orbit coupling between L and S.

Fig. 5 (upper diagram) presents the situation for the case where the spin-orbit coupling is much weaker than the other interactions. The spin S of the ion is then nearly parallel to the exchange field H_{exch}, and L is nearly parallel to the local axis, in one of the two directions. The figure shows these lowest two states of the ion; the state shown by solid lines is the one with the lowest energy. As a result of the spin-orbit coupling the energy of the ion depends on the direction ϕ between S and L in accordance with

$$E = \pm |\lambda| LS \cos\phi. \quad (9)$$

If the other interactions are sufficiently strong, ϕ will be nearly equal to θ, the angle between H_{exch} (the magnetization direction) and the local axis. It can be seen from fig. 5 that the energy of the lowest two states is then given approximately by

$$E \approx \pm |\lambda| LS \cos\theta. \quad (10)$$

At 0 K the ion will be in the state of lowest energy and therefore anisotropic, as given by

$$E_{ground} \approx -|\lambda| LS |\cos\theta|. \quad (11)$$

The situation outlined above is more or less applicable to the 3d ions Fe^{2+} [13] [14] and Co^{2+} [16] [30] at octahedral sites in YIG.

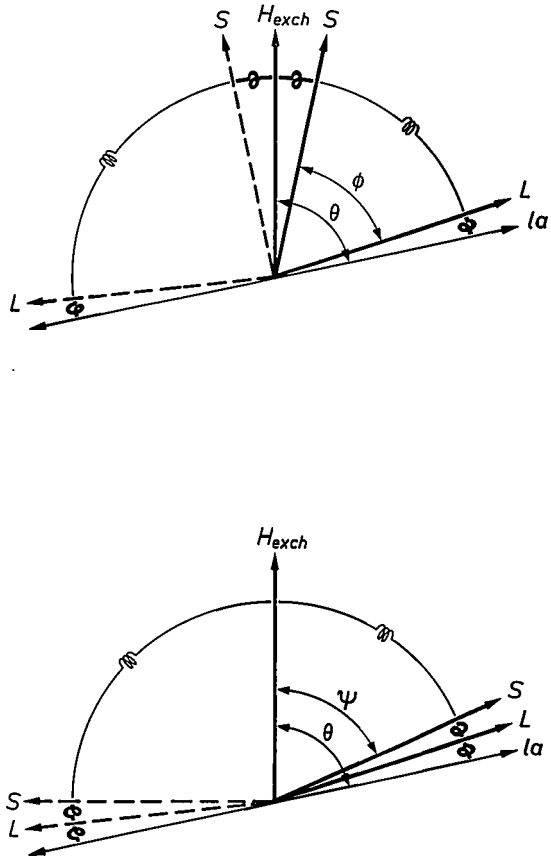


Fig. 5. Schematic representations of the interactions to which the spin momentum S and the orbital momentum L of a strongly anisotropic ion are subjected. The spin S is coupled to the spins of the neighbouring magnetic ions by exchange interaction. In the molecular-field model this interaction is represented by the exchange field H_{exch} parallel to the magnetization. The orbital momentum L of the ion is coupled by the local crystal field to the local axis *la* of the site occupied by the ion. The spin-orbit coupling links the spin with the orbital momentum. The *upper figure* represents the situation where the spin-orbit coupling is the weakest coupling, which is more or less the case for Co^{2+} and Fe^{2+} at octahedral sites in YIG. The spin S is then nearly parallel to H_{exch} and L is nearly parallel to the local axis, in one of the two directions. The state of lowest energy is drawn as a solid line; the energy level just above it is shown dashed. Since the spin-orbit coupling is equal to $-|\lambda|L\cdot S$, it is easy to see that the energy of the lowest state is given approximately by $-|\lambda| LS |\cos\theta|$, where θ is the angle between H_{exch} and the local axis. The *lower figure* is representative of the 4d ion Ru^{3+} and the 5d ion Ir^{4+}. The exchange interaction is the weakest coupling and the energy of the ground state is anisotropic in accordance with $-g\mu_B H_{exch}|\cos\theta|$.

This simple model shows that the anisotropy of the ground state of the ion is proportional to the spin-orbit-coupling parameter λ. The anisotropy of the 4d and 5d ions would therefore be expected to be much greater than that of the 3d ions, in accordance with the much greater λ.

It has already been stated that the crystal fields for these ions are also greater. It now appears that the situation for the 4d and 5d ions is such that both the spin-orbit coupling and the local crystal field are substantially greater than the exchange energy. This situation is illustrated in fig. 5 (lower diagram). As in the previous case, we find that the anisotropy of the ground state of these ions is proportional to the weakest interaction, in this case the exchange interaction. In the molecular-field model the exchange energy is equal to

$$E_{exch} = -g\mu_B H_{exch} S \cos\psi, \quad (12)$$

where g is the g factor, μ_B the Bohr magneton and ψ the angle between H_{exch} and S. By way of approximation we may again, as in the previous case, substitute θ for ψ provided the exchange energy is sufficiently small. The anisotropy of the lowest energy level is thus given by

$$E_{ground} \approx -g\mu_B H_{exch} S |\cos\theta|. \quad (13)$$

Summarizing, we find with this model that the ground state of a strongly anisotropic ion, assuming that the exchange interaction or the spin-orbit coupling is much weaker than the two other interactions, is twofold and that the anisotropy of these two states is proportional to the weakest interaction. The anisotropy of the lowest level follows approximately a $|\cos\theta|$ function with respect to the local axis.

Single-ion anisotropy of d^5 ions in the low-spin state

Theory

Now that the mechanism producing the anisotropy of the ground state of an ion has been described in general terms, this section will be confined to the presentation of some of the results of our quantum-mechanical calculations [19] of the anisotropy of the electronic states of 4d and 5d ions in garnet. For the d^5 ions in a low-spin state at an octahedral site in

[26] A. B. Smith and R. V. Jones, J. appl. Phys. 34, 1283, 1963.
[27] A. B. Smith, Rev. sci. Instr. 39, 378, 1968.
[28] G. F. Dionne, J. appl. Phys. 41, 2264, 1970.
[29] An introduction to the magnetic concepts is given in reference [25]; for a more recent treatment see S. Krupička, Physik der Ferrite und der verwandten magnetischen Oxide, Vieweg, Braunschweig 1973.
[30] J. C. Slonczewski, J. appl. Phys. 32, 253 S, 1961.

YIG it is found that the lowest two states are given by [19-22]

$$E_{1,2} = \pm \tfrac{1}{2}A\sqrt{B + C\cos^2\theta}, \qquad (14)$$

where A is proportional to H_{exch}. B and C depend on the ratio of the strength v of the attachment of $\boldsymbol{L}$ to the local axis, viathe low-symmetry crystal field, and the spin-orbit coupling λ. It is assumed that the exchange interaction is the weakest interaction. *Fig. 6* illustrates the arrangement of the energy levels as given by eq. (14). Here the magnetization (and hence $\boldsymbol{H}_{\text{exch}}$) is rotated in a (110) plane of the garnet. We saw in fig. 1 that the local axis of an octahedral site in garnet is parallel to one of the $\langle 111\rangle$ axes. In fig. 6 it is shown how the energy levels of ions at two types of site, with the [111] and [11$\bar{1}$] directions as their local axes, depend on the direction of $\boldsymbol{H}_{\text{exch}}$. It can be seen from eq. (14) that for a relatively small B the energy levels approximately follow the $|\cos\theta|$ function that was deduced earlier from the simple model.

A brief description of the procedure we used for calculating the electronic energy levels of the ions [19-22] will now be given. As stated earlier, the 4d and 5d ions have a low-spin ground state as a result of the high value of the crystal field. In this ground state only the single-electron states t_{2g} are occupied, corresponding to the sixfold degenerate $^2T_{2g}$ level. Because of the considerable distance to the next energy level, in which one of the single-electron states e_g is occupied, only the $^2T_{2g}$ level need be considered in order to explain the anisotropic properties of the ions. The threefold orbital degeneracy of this level is removed by the action of the low-symmetry crystal field and the spin-orbit coupling, which are of about the same strength (see *fig. 7*). The remaining twofold spin degeneracy is removed by the exchange interaction, finally resulting in three magnetic doublets.

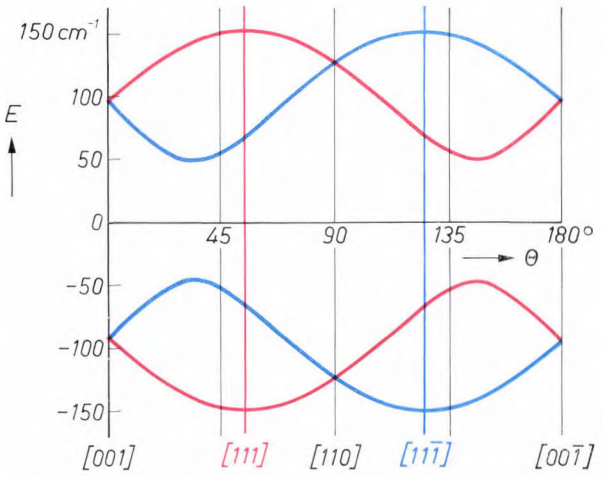


Fig. 6. Anisotropy of the ground-state energies of a low-spin d^5 ion in YIG. The energies were calculated from eq. (14) with the parameter values applicable to the Ru^{3+} ion. The energies are shown for Ru^{3+} at two octahedral sites: the red lines relate to an Ru^{3+} ion at an octahedral site with local [111] axis, the blue lines to Ru^{3+} at a site with local [11$\bar{1}$] axis. The angle Θ is the angle between the magnetization $\boldsymbol{M}$ in the plane through the two axes (the (1$\bar{1}$0) plane) and the [001] direction.

To calculate the splitting of the $^2T_{2g}$ level, and in particular the anisotropy of the lowest levels, the Schrödinger equation,

$$H\psi = E\psi, \qquad (15)$$

must be solved. For our problem the hamiltonian H has the form

$$H = V_t(\boldsymbol{r}) + \lambda\boldsymbol{L}\cdot\boldsymbol{S} + g\mu_B\boldsymbol{H}_{\text{exch}}\cdot\boldsymbol{S} + \delta V(\boldsymbol{r},\varepsilon). \qquad (16)$$

The first term describes the effect of the local crystal field. The second term represents the spin-orbit coupling. The third term expresses the exchange interaction in the molecular field approximation. Finally, the term $\delta V(\boldsymbol{r},\varepsilon)$ is the change in crystal-field

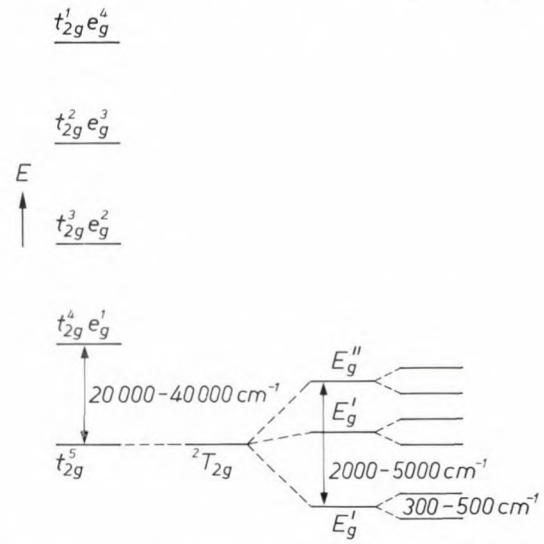


Fig. 7. Energy-level diagram for a low-spin d^5 ion at an octahedral site. Because of the interaction with the cubic part of the crystal field a $^2T_{2g}$ state becomes the ground state, where the orbital degeneracy is threefold and the spin degeneracy is twofold. Under the influence of the local field with threefold symmetry and the spin-orbit coupling, this state is split into three twofold degenerate states. These three doublets are in their turn split by the exchange interaction with surrounding magnetic ions. The energy differences are not shown to scale.

energy for a strain ε. This term is responsible for the magnetostrictive effects. We applied this hamiltonian to a suitably chosen set of wave functions. Because of the local symmetry we chose the d-wave functions, quantized along the local axis. The resultant anisotropic ground state found in this way is given by eq. (14).

At 0 K the ion will be in the state of lowest energy, whose anisotropic energy is given by

$$E_{\text{ground}} = -\tfrac{1}{2}A\sqrt{B + C\cos^2\theta}. \qquad (17)$$

At higher temperatures the single-ion anisotropy energy is obtained by calculating the free energy of the lowest two levels:

$$F_K(\theta) = -kT\ln\cosh\left(\frac{A}{2kT}\sqrt{B + C\cos^2\theta}\right). \qquad (18)$$

We are concerned, however, with energy levels of ions at the four sites with different local axes. For a given direction of the magnetization we must therefore average over the single-ion anisotropy of these four sites to find the single-ion anisotropy for each ion. In

doing so we assume that the ions are statistically distributed over all the octahedral sites, so that the various kinds of site are occupied by the same numbers of ions. The resultant expression for F_K has in general a rather complicated dependence on the direction of magnetization α, which is only poorly approximated by a power series like that of eq. (1). Nevertheless, an expression of this type, with only two constants, K_1 and K_2, can be used to describe the differences in energy when α lies along the different principal directions of the cubic crystal: [100], [111] and [110] [30] [31]. In this way we can find explicit expressions for the change ΔK_i in the anisotropy constants K_i as a consequence of substituting low-spin d^5 ions in YIG. These expressions completely describe the temperature dependence of the ΔK_i. For substitution at octahedral sites we find for the ΔK_i per cm^3 [21] [22]:

$$\left.\begin{aligned}\Delta K_1 &= 2kTN \times \\ &\times \ln\left[\frac{\cosh^2(t)}{\cosh(t\sqrt{1+\eta})\cosh(t\sqrt{1-\eta})}\right], \\ \Delta K_2 &= \tfrac{9}{4}kTN \times \\ &\times \ln\left[\frac{\cosh^8(t\sqrt{1+\eta})\cosh^8(t\sqrt{1-\eta})}{\cosh^4(t)\cosh^3(t\sqrt{1-2\eta})\cosh^9(t\sqrt{1+\tfrac{2}{3}\eta})}\right],\end{aligned}\right\} \quad (19)$$

where t and η are related to the constants A, B and C from eq. (17) by

$$t = \frac{A}{2kT}\sqrt{B + \tfrac{1}{3}C}, \qquad \eta = -\frac{C}{C+3B}, \quad (20)$$

and N is the number of substituted ions per cm^3.

At low temperatures, where the arguments of the cosh terms in equation (19) are much greater than 1, the ΔK_i are simply proportional to t and thus to A. Since A is proportional to the exchange field H_{exch}, and B and C depend only on v/λ, it can be seen that the ΔK_i at low temperatures are proportional to H_{exch}, as expected from the simple model. Dividing the ΔK_i by $Ng\mu_B H_{exch}$, we find that the anisotropy constants thus reduced depend only on v/λ. This dependence is plotted in *fig. 8*, which shows that large contributions to the anisotropy may be expected for positive v/λ, as found in the YIG lattice. The sign of ΔK_1 is found to be positive, while the sign of ΔK_2 is positive for negative v/λ and negative for positive v/λ. The contribution to the anisotropy can be calculated in the same way for d^5 ions at tetrahedral sites. The results for this case are also given in fig. 8. Note in particular that the sign of ΔK_1 is always negative, whereas it was positive at the octahedral site. Since ΔK_i is proportional to t at low temperatures, the ratio $\Delta K_2/\Delta K_1$ depends only on η. *Fig. 9* shows this dependence for d^5 ions at both octahedral and tetrahedral sites.

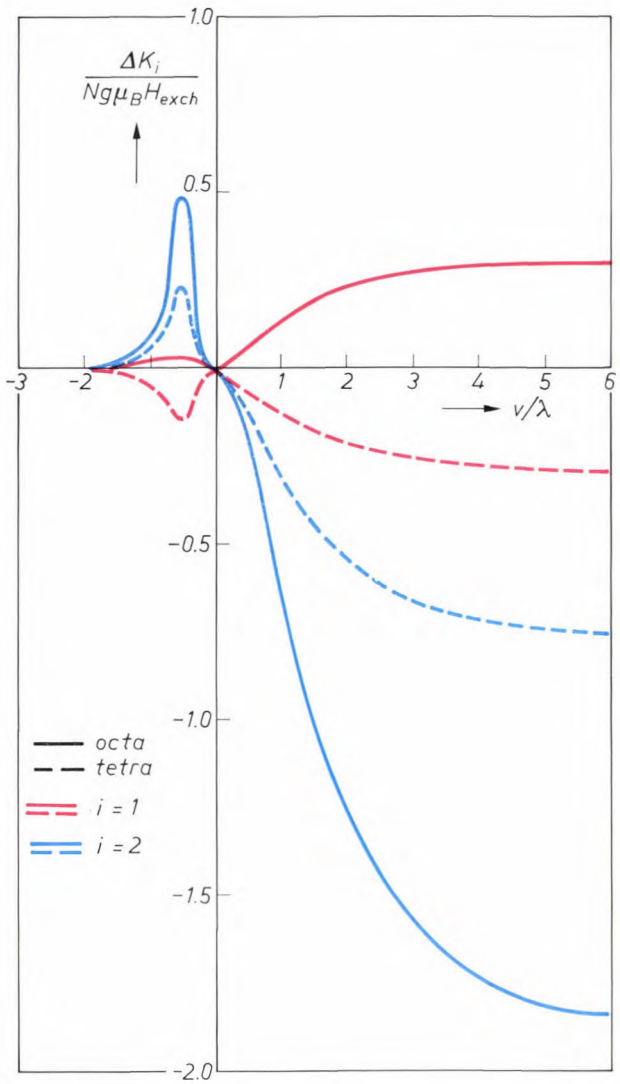


Fig. 8. Theoretical single-ion contributions to the low-temperature anisotropy constants of YIG of the low-spin d^5 ions at octahedral and tetrahedral sites as a function of the ratio of the strength v of the local crystal field and the spin-orbit coupling λ. The ΔK_i are proportional to the exchange energy. The reduced contributions are given, i.e. the ΔK_i divided by $Ng\mu_B H_{exch}$. It can be seen that an ion at an octahedral site always makes a positive contribution to K_1, whereas the contribution from an ion at a tetrahedral site is negative. The sign of the contribution to K_2 depends on the sign of the local crystal field.

The theory presented here for low-spin d^5 ions also holds to some extent for other ions. The theory led to eq. (14) for the twofold anisotropic ground state. The conclusions drawn from this, including eq. (19) for the ΔK_i and fig. 9, also apply of course for any ion with a ground state of this kind. In the simple model for the single-ion anisotropy we saw that the 3d ions Fe^{2+} and Co^{2+} have a ground state similar to that of the low-spin d^5 ions even though the exchange role is taken over by the spin-orbit coupling. From theoretical cal-

[31] J. Smit, F. K. Lotgering and R. P. van Stapele, J. Phys. Soc. Japan **17**, Suppl. B-I, 268, 1962.

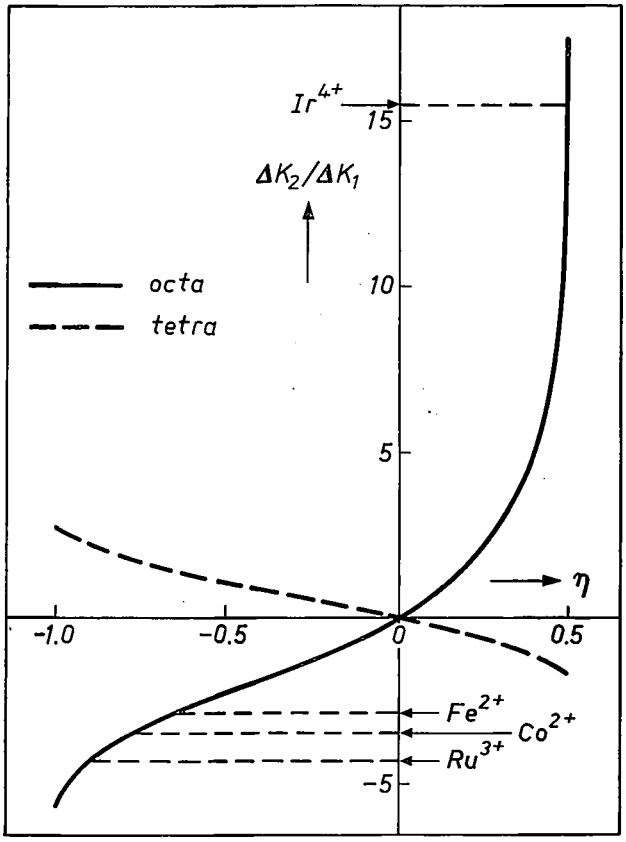


Fig. 9. Theoretical plot of $\Delta K_2/\Delta K_1$ at low temperature for an ion with a doublet ground state given by eq. (14), as a function of the parameter η. The experimentally determined values of $\Delta K_2/\Delta K_1$ for strongly anisotropic ions are indicated. The value of η for low-spin d^5 ions is related in a simple way to v/λ. The figure is also valid for any ion with a ground state as given by eq. (14).

culations [13] [14] [16] [30] it is indeed found that Fe^{2+} and Co^{2+} have a ground state as given by eq. (14), but that the physical parameters that determine the constants A, B and C, and hence t and η, are different.

Experiments

The single-ion theory showed that at low temperatures $\Delta K_2/\Delta K_1$ depends only on the parameter η, as illustrated in fig. 9. By plotting the experimental values obtained for $\Delta K_2/\Delta K_1$ by the ferromagnetic-resonance method (Fe^{2+} [14], Ru^{3+} [19] [20], Ir^{4+} [24]) and the torque method (Co^{2+} [16]), the corresponding value of η can be determined. The η values found for the Ru^{3+} and Ir^{4+} ions are related to v/λ; it is found experimentally that $v/\lambda = +1.0$ for Ru^{3+} and $v/\lambda = -0.5$ for Ir^{4+}. The correctness of the values found for these ions can be directly verified since the values of v/λ have also been measured by electron paramagnetic resonance (EPR experiment) on the nearly isomorphic yttrium gallium garnet doped with small amounts of these ions, yielding $v/\lambda = +1.06$ for Ru^{3+} [32] and $v/\lambda = +0.45$ for Ir^{4+} [33]. For the Ru^{3+} ions the agreement is good,

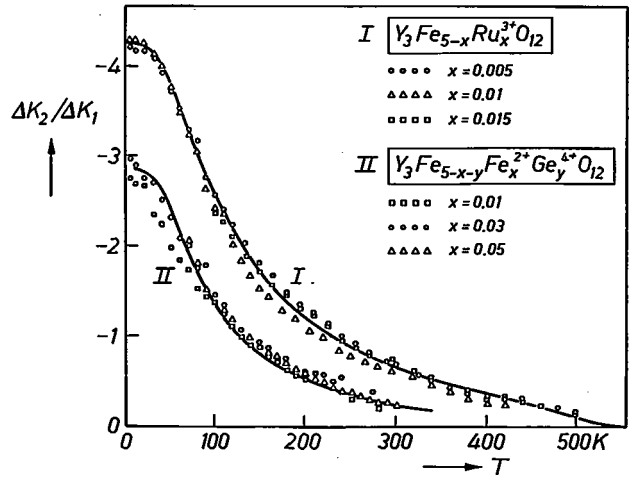


Fig. 10. Temperature dependence of $\Delta K_2/\Delta K_1$ for YIG doped with Fe^{2+} and Ru^{3+}. The curves are theoretical and were calculated from eq. (19) with η and t as parameters.

but for the Ir^{4+} ions the sign is different. A possible reason for this discrepancy is that the Ir ions do not all have the same valence or do not all occupy octahedral sites. Optical absorption measurements indicate that the Ir ions do in fact occupy tetrahedral sites as well and also occur in different valence states.

In the single-ion model the temperature dependence of $\Delta K_2/\Delta K_1$ is independent of the concentration. *Fig. 10* shows the experimental results obtained for YIG crystals doped with Fe^{2+} and Ru^{3+} [14] [22], which verify this statement. Similar results were found for gadolinium iron garnet doped with Ru^{3+} [22] and for YIG doped with Co^{2+} [34], whereas for Ir^{4+} a significant scatter is found, probably for the reasons mentioned above. The solid lines in fig. 10, representing the theoretical temperature dependence of $\Delta K_2/\Delta K_1$, were calculated from eq. (19) using the value of η deduced

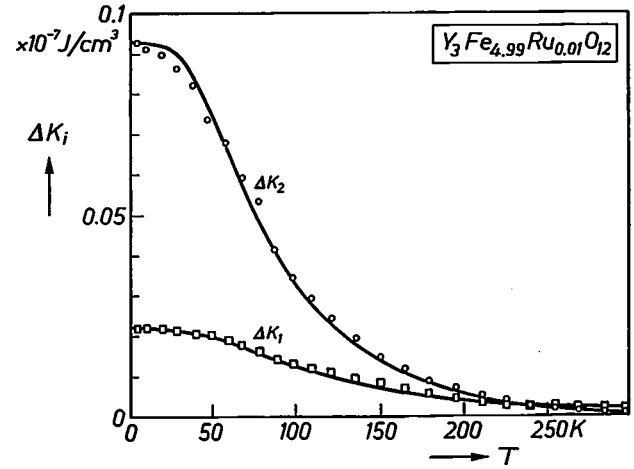


Fig. 11. Temperature dependence of the contribution ΔK_i to the anisotropy from Ru^{3+} ions in YIG. The curves show the theoretical relation for $v/\lambda = 1$ and $gH_{exch} = 9.5 \times 10^8$ A/m (1.2×10^7 Oe).

from $\Delta K_2/\Delta K_1$ at low temperatures and using t as an adjustable parameter. For Ru^{3+} the value of $t \times T$ is itself temperature-dependent since it is proportional to the exchange field and therefore proportional to the magnetization of the sublattice formed by the Fe^{3+} ions at tetrahedral sites. The absolute values of the ΔK_t for YIG doped with Ru^{3+} are shown in *fig. 11* together with the theoretical curves based on the parameters $v/\lambda = -0.95$ and $gH_{exch} = 9.5 \times 10^8$ A/m (1.2×10^7 Oe). From the comparison between the experimental results and the theory it appears that the single-ion theory gives a good description of the observed effects.

Single-ion magnetostriction of d^5 ions in the low-spin state

Theory

In the section on anisotropy we looked at the situation where the crystal field at octahedral sites has a large cubic component and a smaller uniaxial component with the local axis as the principal axis. This uniaxial crystal field plays a dominant role in the mechanism of single-ion anisotropy, since it binds the orbital momentum L to the local axis. The value of the crystal field used here so far is that of the unstrained lattice, corresponding to the definition of the anisotropy energy F_K for the unstrained lattice ($\varepsilon = 0$). If, however, a finite strain ε is present it affects the crystal field, so that the anisotropy of the ground state of the ion is slightly changed. This, briefly, is the single-ion mechanism for the production of magnetoelastic energy, as introduced in eq. (2) of the phenomenological treatment.

We have calculated the single-ion contribution to the magnetostriction [23] by quantum-mechanical methods [35] in which the change in the crystal field caused by a strain ε is treated as a relatively small perturbation. For low-spin d^5 ions the result has the form:

$$\Delta\lambda_{hkl} = n_{hkl} h_{hkl}(H_{exch}, v, \lambda, T), \quad hkl = 100 \text{ or } 111, \tag{21}$$

where the numbers n_{hkl} depend on the sensitivity of the crystal field to strains. The values of h_{hkl} at $T = 0$ K are given in *fig. 12* as a function of v/λ for various values of the ratio $g\mu_B H_{exch}/\lambda$. These values are such that considerable contributions to the magnetostriction constants are to be expected from d^5 ions in the low-spin state.

Experiments

Most of the experimental magnetostriction results reported in the literature have been obtained at room temperature. A comparison between experiment and theory is only useful, however, if low-temperature

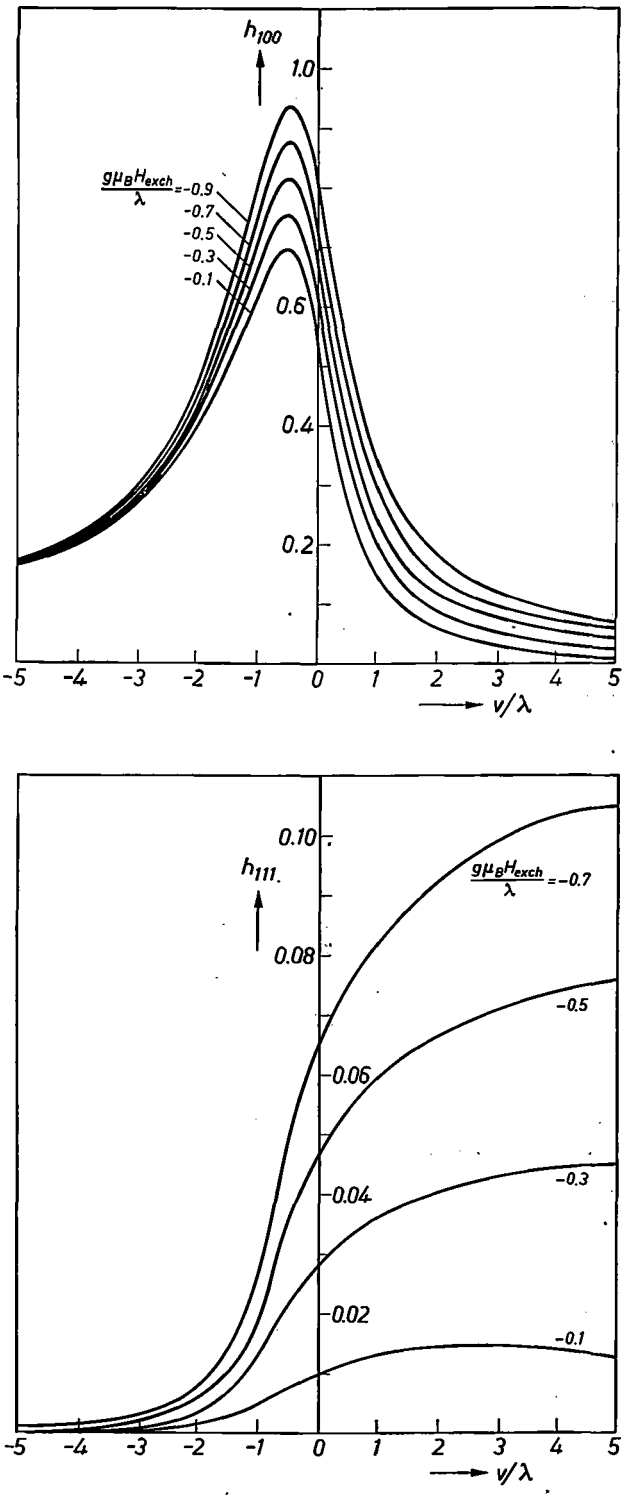


Fig. 12. Theoretical value of h_{100} and h_{111} for low-spin d^5 ions at octahedral sites at $T = 0$ K as a function of the ratio v/λ. The magnetostriction constants λ_{100} and λ_{111} are proportional to h_{100} and h_{111} respectively (see eq. 21).

[32] I. A. Miller and E. L. Offenbacher, Phys. Rev. 166, 269, 1968.
[33] E. L. Offenbacher and H. Waldman, Bull. Amer. Phys. Soc. 13, 435, 1968, and J. A. Hodges, Phys. Stat. sol. (b) 67, 715, 1975.
[34] P. Hansen, R. Krishnan and W. Tolksdorf, unpublished.
[35] J. C. Slonczewski, Phys. Rev. 122, 1367, 1961.
[36] L. M. Protopopova, G. A. Petrakovskii, E. M. Smokotin and R. A. Petrov, Sov. Phys. Solid State 15, 2192, 1974.

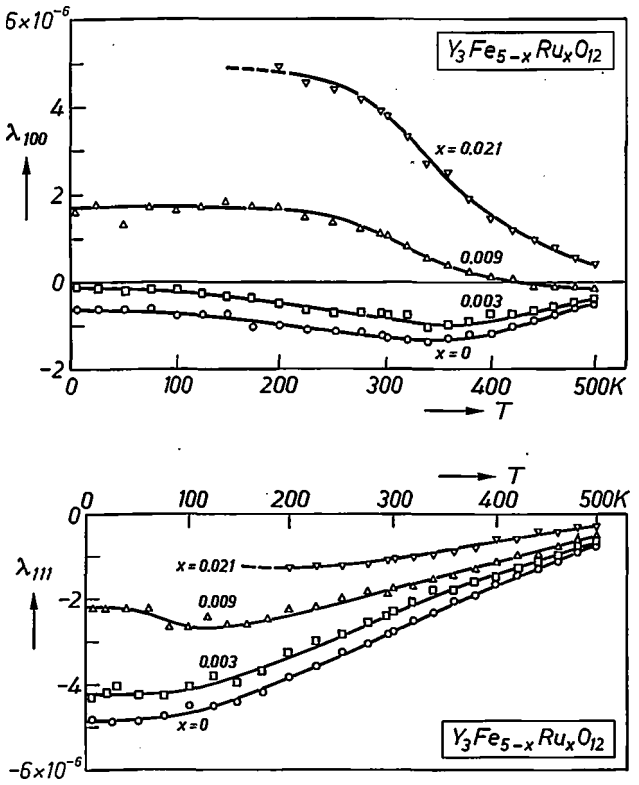


Fig. 13. Experimentally determined magnetostriction constants λ_{100} and λ_{111} for Ru-doped YIG as a function of temperature.

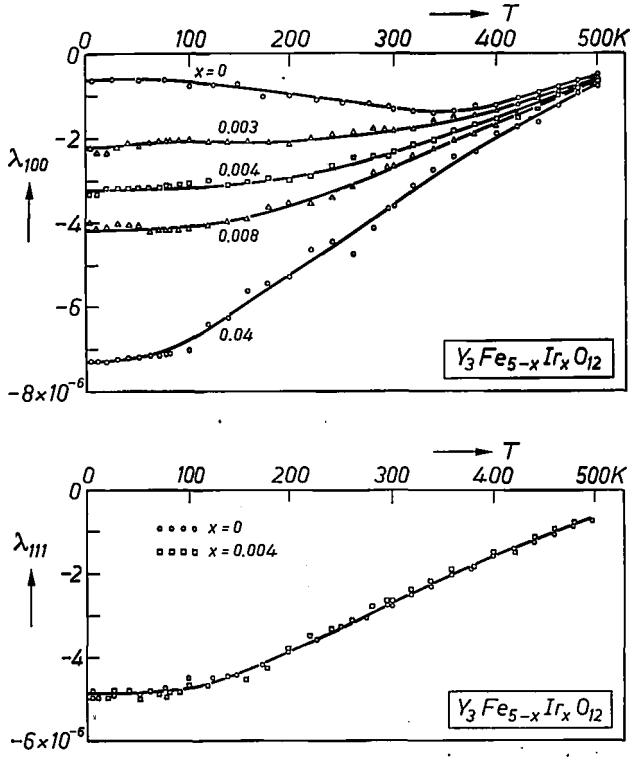


Fig. 14. Experimentally determined magnetostriction constants λ_{100} and λ_{111} of Ir-doped YIG as a function of temperature.

results are also available. Experiments at low temperatures have been carried out on YIG doped with Ru^{3+} [23], with Ir^{4+} [24] and with Co^{2+} [36], using the resonance method described earlier. The results for Co^{2+} are not quite clear, particularly for the concentration dependence, nor were the results compared with the single-ion theory. For Ru^{3+} and Ir^{4+} we determined $\Delta\lambda_{100}$ and $\Delta\lambda_{111}$ by the resonance method, and the results are given in *figs. 13* and *14* for various small concentrations in the temperature range from 4 to 500 K. The Ru ions induce a positive shift for both λ_{100} and λ_{111}. The concentration dependence at room temperature was shown in fig. 4: at 4 K it is somewhat greater. For the Ir^{4+} ions $\Delta\lambda_{100}$ was found to be negative, but λ_{111} remained unaffected. However, as mentioned in the section on anisotropy, this ion is assumed to be distributed over tetrahedral and octahedral sites, and is also assumed to be present in different valence states. This ion is not therefore suitable for a comparison with the single-ion theory. The only results that can serve for a comparison with the single-ion theory are those obtained for Ru^{3+}. In *fig. 15* the contributions $\Delta\lambda_{hkl} = \lambda_{hkl}(x) - \lambda_{hkl}(0)$ for $x = 0.009$ are plotted as a function of temperature. The theoretical curves were calculated from eq. (21) and are based on the values of the atomic parameters derived from the contribution of Ru^{3+} to the anisotropy. The absolute value of $\Delta\lambda_{hkl}$ at $T = 0$ K was fitted using the constants n_{hkl}. In general terms the temperature dependence of both $\Delta\lambda_{100}$ and $\Delta\lambda_{111}$ is well described by the theory, but the agreement between theory and experiment is not so good as in the case of anisotropy.

In the foregoing it has been shown that the anisotropic and magnetostrictive properties of strongly anisotropic ions, such as Ru^{3+} and Ir^{4+}, in yttrium iron garnet can be satisfactorily explained in terms of the single-ion theory. We have seen that, as expected, these ions are even more strongly anisotropic than the 3d transition-metal ions Fe^{2+} and Co^{2+}. The single-ion contribution to the anisotropy and magnetostriction of the strongly anisotropic ions Fe^{2+}, Co^{2+}, Ru^{3+} and Ir^{4+} in YIG are collected in *Table I*, together with the contributions of the far less strongly anisotropic ions Fe^{3+} and Cu^{2+} [37–39]. The contributions from weakly and strongly anisotropic ions differ by two or three

[37] R. Krishnan, J. appl. Phys. **40**, 1557, 1969.
[38] P. Novák, Int. J. Magn. **2**, 177, 1972.
[39] R. Krishnan and V. Cagan, Proc. Int. Conf. on Ferrites, Kyoto 1970, p. 57.
[40] See for example W. Tolksdorf, G. Bartels, G. P. Espinosa, P. Holst, D. Mateika and F. Welz, J. Crystal Growth **17**, 322, 1972, and J. A. Pistorius, J. M. Robertson and W. T. Stacy, Philips tech. Rev. 35, 1, 1975.

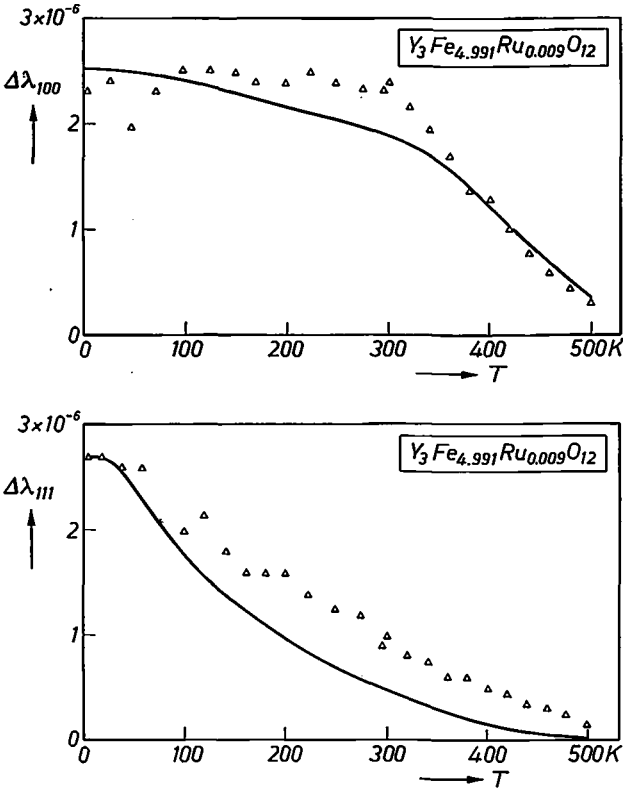


Fig. 15. Comparison of the experimentally determined contribution to the magnetostriction constants λ_{100} and λ_{111} for Ru^{3+} in Ru-doped YIG with the single-ion theory. The theoretical curves were calculated using the atomic parameters deduced from the fit of the anisotropy. The values of the magnetostriction constants at 0 K have been fitted by varying the parameters h_{100} and h_{111}.

orders of magnitude. A difficulty in the experimental determination of the single-ion anisotropy and magnetostriction of strongly anisotropic ions is that it is generally difficult to ascertain with any certainty the lattice sites at which very small concentrations of these ions occur and the valence of the ions. This applies particularly for Ir ions in YIG, which presumably occupy both octahedral and tetrahedral sites, possibly in both the trivalent and the tetravalent states. The tabulated data for this ion in particular are not therefore very accurate.

We have utilized the magnetostrictive properties of Ru^{3+} ions for controlling the anisotropy of garnet films in a magneto-optical memory device [2]. The magnitude of the uniaxial anisotropy of these films, which are grown epitaxially on a non-magnetic substrate [40], has to be given a specific value. The uniaxial anisotropy is due partly to the growth process and partly to the stress induced by the mismatch between film and substrate. The component due to growth can be reduced by subsequent annealing. The component due to induced stress depends on the degree of mismatch, i.e. the

Table I. The effect of transition-metal ions on the anisotropy and magnetostriction of yttrium iron garnet. The abbreviations oct and dod denote the octahedral and dodecahedral sites. The column headed 'ground state' indicates, with the usual symbols, the appropriate ground state in the cubic crystal field. The letter T denotes a threefold degenerate orbital state, the letter A a singlet orbital state.

Ion	Configuration	Ground state (site)	T(K)	Anisotropy		Ref.	Magnetostriction		Ref.
				K_1/N (cm^{-1})	K_2/N (cm^{-1})		$10^4\,\lambda_{100}/x$	$10^4\,\lambda_{111}/x$	
			4.2	16	260		−5.1	−0.2	
Ir^{4+}	$5d^5$	$^2T_{2g}$ (oct)	77	15	180	[24]	−4.8	−0.2	[24]
			295	2	1		−1.9	0	
			4.2	28	−120		2.7	2.9	
Ru^{3+}	$4d^5$	$^2T_{2g}$ (oct)	77	22	− 68	[20-22]	2.7	2.6	[23]
			295	2	− 3		2.5	0.8	
			4.2	29	−100		—	—	
Co^{2+}	$3d^7$	$^4T_{1g}$ (oct)	77	20	− 62	[16, 17]	—	—	[17]
			295	0.4	− 0.2		0	0.5	
			4.2	4.5	− 17		—	—	
Fe^{2+}	$3d^6$	$^5T_{2g}$ (oct)	77	2.3	− 4.1	[14]	2.7	6.3	[14]
			295	0.1	− 0.1		0.5	1.4	
			4.2	−0.4	—		−0.08	−0.09	
Cu^{2+}	$3d^9$	$^2T_{2g}$ (dod)	77	−0.3	—	[37, 38]	−0.05	−0.08	[39]
			295	−0.02	—		−0.07	−0.04	
			4.2	−0.005	0		−0.025	0.020	
Fe^{3+}	$3d^5$	$^6A_{1g}$ (oct)	77	−0.004	0	[10]	−0.011	0.020	[10]
			295	0.006	0		0.007	0.017	

magnitude of the difference in lattice constants, and on the magnetostrictive properties of the garnet film. In cases where the mismatch is fixed by requirements to be met by other properties, such as saturation magnetization and Faraday rotation, the stress-induced uniaxial anisotropy can be controlled by varying the magnetostrictive properties of the film with highly anisotropic ions [41]. *Fig. 16* demonstrates the observed linear relation between the uniaxial anisotropy, expressed as the uniaxial anisotropy constant K_u, and the magnetostriction constant λ_{111} of a garnet film. The value of K_u for $\lambda_{111} = 0$ is the growth-induced part of the anisotropy. The value of λ_{111} was varied by doping the garnets with small concentrations of Ru^{3+} ions. Since λ_{111} is proportional to the Ru^{3+} concentration, it is possible to plot this concentration linearly along the horizontal axis in fig. 16. This example shows how the properties of strongly anisotropic ions may be used for selectively controlling the anisotropic and magnetostrictive properties of garnets.

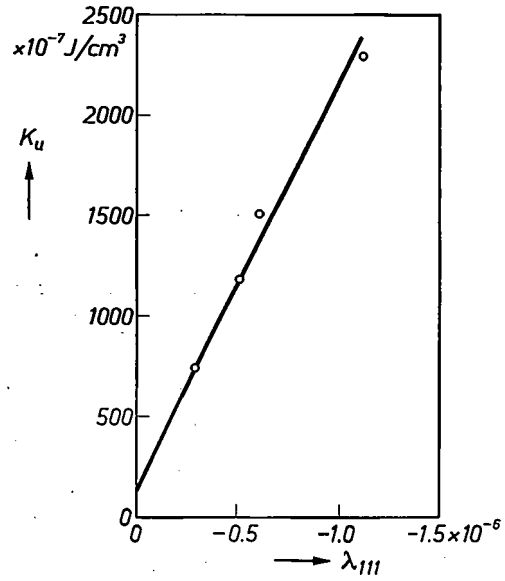


Fig. 16. In an epitaxial garnet film with [111] orientation the uniaxial anisotropy constant K_u depends linearly on the magnetostriction constant λ_{111}. The value of λ_{111} was varied by doping the garnet with Ru^{3+} ions. In this way the K_u of these films can be tailored to the requirements of the application.

[41] J.-P. Krumme, P. Hansen, G. Bartels and D. Mateika, J. appl. Phys. 46, 2801, 1975.

Summary. The source of the magnetic anisotropy and magnetostriction of cubic ferrites and garnets is the single-ion anisotropy and magnetostriction of the magnetic ions. The strongly anisotropic ions, such as Fe^{2+} and Co^{2+}, have a single-ion anisotropy and magnetostriction two orders of magnitude greater than that of the Fe^{3+} ions, often the only magnetic ions in the pure ferrite or garnet. By substituting very small concentrations of strongly anisotropic ions for the Fe^{3+} ions the anisotropy and magnetostriction of such a material can be selectively controlled, that is to say without significantly affecting other physical properties, such as the saturation magnetization.

This article gives an account of an investigation of the 4d transition-metal ion Ru^{3+} and the 5d transition-metal ion Ir^{4+}, both substituted in yttrium iron garnet. Both ions are an attractive alternative to Fe^{2+} and Co^{2+} ions because of their greater single-ion anisotropy and magnetostriction. For example, at room temperature their contribution to the first anisotropy constant K_1 is five times greater than that of Co^{2+}. After a description of the mechanism underlying the anisotropy and magnetostriction of these ions, results of theoretical calculations are presented.

The contribution of Ru^{3+} and Ir^{4+} ions to the anisotropy and magnetostriction of yttrium iron garnet was measured by the ferromagnetic-resonance method for single crystals of yttrium iron garnet doped with these ions. Owing to the uncertainty about the lattice site occupied by the Ir ion and about the valence state of the ion, the results with the Ir dopant cannot be used for a comparison with the single-ion theory. In the case of the Ru^{3+} ions, which occupy octahedral sites, good agreement is found between experiment and theory.

Investigation of crystal defects by cathodoluminescence

C. Schiller and M. Boulou

In the research on defects in crystalline material considerable progress has been made in the last fifteen years, in both theoretical and experimental aspects. One of the problems that has been studied — and contributed to by the authors of this article — is the relation between the dislocations and stacking faults in a semiconductor crystal and the electrical properties of this crystal. Means of detection of dislocations, preferably non-destructive, are essential in this work. The cathodoluminescence technique described in this article has been of great advantage here.

Introduction

The control of crystal defects is a very important factor in the production of semiconductor devices, because such defects in the crystal may have a direct or indirect effect on the electrical and optical characteristics of the device.

Point defects, such as vacancies and impurities, generally have a direct influence on the electrical behaviour owing to the changes they cause in the effective number of charge carriers. Electrical and optical methods are therefore generally used for studying point defects. Spatially extended defects, such as dislocations, stacking faults, grain boundaries and three-dimensional imperfections (such as segregations) may also have a direct influence on the electrical properties of materials. Indirectly they may affect the electrical and optical properties of semiconductors through an effect they may have — via local mechanical stresses inside the material — on the diffusion and etching processes that play such an important role in semiconductor technology.

The methods required for studying these spatially extended defects are entirely different from those used for studying point defects. X-ray diffraction methods, selective etching, chemical decoration and electron-microscope examinations are techniques widely used for this purpose.

X-ray diffraction methods have a limited resolution (10^5 dislocations/cm^2). The resolution of transmission electron microscopes, on the other hand, is very high (10^{10}/cm^2), but this form of microscopy is usually destructive and rather time-consuming, and the results are often difficult to interpret. Cathodoluminescence investigations using a scanning electron microscope, which will be the subject of this article, fill a definite gap here. The resolution is good, 10^7/cm^2, the technique itself is simple and in many cases it is non-destructive.

Fig. 1. Cathodoluminescence image of Te-doped GaAs. A number of dislocations are visible as black lines and dots, while the recombination of electron-hole pairs induced by electron bombardment occurs locally without radiation. The matrix appears somewhat grey, due to Te precipitation. In the vicinity of dislocations all recombinations radiate while all Te precipitates are concentrated locally on the dislocation (decoration of the dislocations). Magnification 145×.

In the method used a semiconductor specimen in a scanning electron microscope [1] is bombarded by an electron beam. The bombardment produces electron-hole pairs, each primary electron giving rise to a large

Dr C. Schiller and Ing. M. Boulou are with the Laboratoires d'Electronique et de Physique Appliquée, Limeil-Brévannes (Val-de-Marne), France.

[1] A description of a scanning electron microscope can be found in W. Kuypers and J. C. Tiemeijer, Philips tech. Rev. **35**, 153, 1975 (No. 6).

number of such pairs. In a perfect crystal these pairs may recombine in a process that gives rise to recombination emission. In an imperfect crystal there is a higher probability that non-radiative recombination will take place at the locations of irregularities in the crystal lattice. At these locations there will therefore be little if any luminescence. If the recombination emission, the cathodoluminescence, is detected with a photomultiplier tube and the output signal from the tube is used to produce a display, an image can be obtained of the dislocations and other irregularities in the structure of the specimen. An example of a typical cathodoluminescence image is shown in *fig. 1*.

To obtain such a cathodoluminescence image the specimen must of course be sufficiently transparent to the radiation emitted upon recombination. The band gap in the material must also be such that the radiation emitted falls within the sensitivity range of the detectors available. Two further factors are involved, and it is important to have a good quantitative understanding of these factors to interpret the image correctly. First of all it is important to know where, that is to say at what depth below the surface, the interaction takes place between the incoming electrons and the solid. The other factor to be taken into account is the diffusion of the minority carriers between the moment at which they are formed and the moment at which they recombine. The majority carriers are always present in a considerable excess and need not therefore be considered here.

We shall now first consider these two factors separately, and then we shall present some of the results of our investigation.

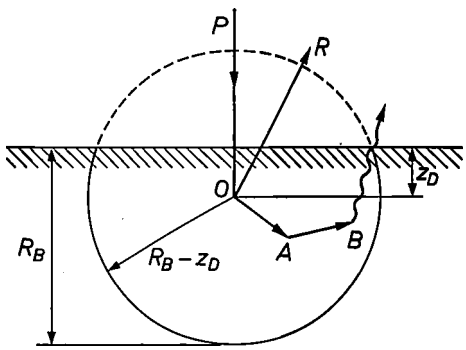


Fig. 2. Theoretical model of the processes that lead to cathodoluminescence. Primary electrons penetrate the material to a depth z_D with no energy loss. Subsequent diffusion transports the electrons to random positions within a sphere of radius $R_B - z_D$, where they give rise to electron-hole pair creation (R_B is the maximum depth of penetration). The minority carriers resulting from this process at *A* may diffuse over some distance before radiative or non-radiative recombination occurs at *B*. Electrons that have penetrated the solid to a depth z_D may diffuse back to the surface and leave the solid as 'reflected' electrons (*R*).

How a cathodoluminescence image is formed

Creation of electron-hole pairs

Electrons fired into a solid lose their energy owing to interaction with the solid. There are various models that describe this interaction [2]. The model we shall use for our present purposes is as follows. The electrons penetrate without loss of energy to a depth z_D in the solid and are then scattered in all directions, their energy decreasing exponentially through the creation of electron-hole pairs. In this way the electrons lose their energy within a sphere of radius $R_B - z_D$, where R_B is the maximum penetration depth as given by H. Bethe [3]. These parameters of the model are illustrated in *fig. 2*. For gallium arsenide, the material on which much of our investigations were carried out, $z_D = 0.24$ μm and $R_B = 2.06$ μm, at a primary electron energy of 15 keV.

On the basis of the model outlined above the energy loss per electron and per unit length travelled can be described as a function of the depth z by the equation

$$\frac{dE}{dz} = \frac{\mu E_0}{4}(1 + \mu|z - z_D|)\exp(-\mu|z - z_D|).$$

Experimental results confirm the following relations [2]:

$$\mu = 880\,\varrho E_0^{-1.65},$$
$$\mu R_B = 11,$$
$$z_D = \frac{4R_B}{Z+4}.$$

Here ϱ is the density of the solid (in mg/cm³) and Z is the (average) atomic number; the initial energy E_0 is expressed in keV and the energy-absorption coefficient μ in μm⁻¹.

Because of the dependence of z_D and R_B on the atomic number and density of the solid, the volume in which the interaction with incident electrons occurs may take various forms. The shape and dimensions of the interaction volume for Si and GaAs are given in *fig. 3*.

Experimental evidence for the model can be found in the literature [4]. We have also confirmed the model by measuring the fraction of the incident electrons that leave the solid after interaction. The fractions found from these measurements may be compared with those calculated with the model [5].

The creation of electron-hole pairs is now proportional at every point in the solid to the local electron current and to the differential energy decrease, i.e. the energy loss of these electrons per unit of distance travelled. The proportionality constant found in this case comprises the band gap of the semiconductor.

These data enable us to calculate the spatial distribution of the creation of charge carriers as a function of depth below the surface. The result (see *fig. 4*) depends very much on the energy of the incident electrons.

Diffusion of charge carriers

The charge carriers generated in the spatial distribution described above will be subject to diffusion before recombination takes place. Recombination may give rise to emission of radiation. The three factors of generation, diffusion and recombination determine the three-dimensional distribution of the minority carriers. Assuming that the bombardment lasts long enough for a steady state to be reached and that the diffusion may be treated as predominantly one-dimensional, the differential equation giving this spatial distribution for a *P*-type semiconductor is:

$$D\frac{d^2n(z)}{dz^2} - \frac{n(z)}{\tau} + g(z) = 0. \qquad (1)$$

Here $n(z)$ is the minority-carrier excess at a depth z, while D is the diffusion constant of these carriers, τ their lifetime and $g(z)$ the generation function mentioned above, which describes the generation of minority carriers as a function of depth. The terms in the equation give successively the three factors that determine the spatial distribution of minority carriers: diffusion, recombination and generation.

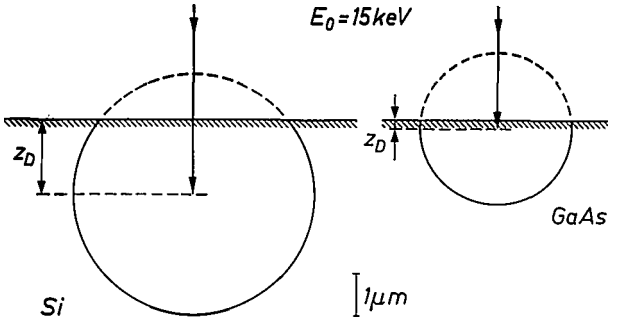


Fig. 3. Shape of the interaction volume for silicon and GaAs at the same primary-electron energy E_0.

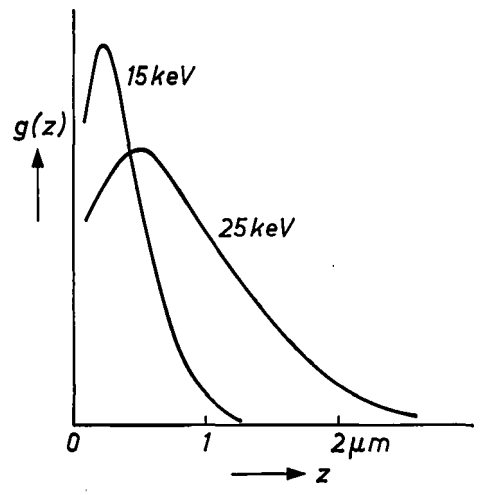


Fig. 4. The function $g(z)$ giving the creation of a minority-carrier excess as a function of the depth z below the surface for two values of the primary-electron energy E_0.

There are various boundary conditions to be taken into account when solving this equation. One of them is the recombination rate at the surface. The solution contains a characteristic length L, called the diffusion length, which is the average distance through which charge carriers are displaced by diffusion before recombination occurs. We can write for this quantity:

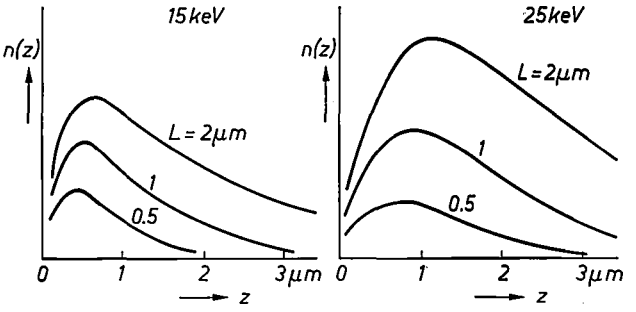


Fig. 5. Minority-carrier excess $n(z)$ as a function of the depth z below the surface; this distribution results from the generation function $g(z)$ given in fig. 4, taking diffusion and recombination into account. The concentration $n(z)$ was calculated at two values of the primary electron energy E_0 and for three values of the diffusion length L. Comparison with fig. 4 shows the effect of diffusion and the high surface-recombination rate.

$L^2 = D\tau$. The solution of equation (1) is given in *fig. 5* for two values of the energy of the incident electrons and at each energy value for three values of the diffusion length L. A reduction of the diffusion length immediately gives rise to a marked decrease in the number of minority carriers. Comparison with fig. 4 indicates that the recombination rate at the surface is high.

Intensity of the cathodoluminescence

The intensity of the luminescence resulting from recombination is proportional to the concentration of the minority carriers and inversely proportional to the lifetime for a radiative recombination τ_R. The observed intensity I is given by:

$$I(z) = \Omega\frac{n(z)}{\tau_R},$$

where Ω is the efficiency of the observation instrument. The curves in fig. 5 thus give the luminescence intensity, multiplied by a constant factor, as a function of depth.

The curves show that the intensity is weak in a layer immediately below the surface. The layer is thicker the

[2] C. A. Klein, in: P. L. Kelley, B. Lax and P. E. Tannenwald (ed.), Physics of quantum electronics, McGraw-Hill, New York 1966, p. 424, and C. A. Klein, Appl. Optics 5, 1922, 1966.
[3] H. Bethe, Ann. Physik (5) 5, 325, 1930.
[4] H. E. Bishop, in: D. B. Holt, M. D. Muir, P. R. Grant and I. M. Boswarva (ed.), Quantitative scanning electron microscopy, Academic Press, London 1974, p. 41.
[5] C. Schiller and D. Bois, Rev. Phys. appl. 9, 361, 1974.

greater the primary energy and the diffusion length. For a diffusion length of 1 μm the thickness of the layer increases from 0.5 to 1 μm when the primary energy is raised from 15 to 25 keV.

The contrast caused by imperfections in the crystal, such as lattice defects, dislocations or segregations, depends partly on the diffusion length and on the extent of the defect region in the lattice where non-radiative transitions take place. The other factor affecting the contrast is the local concentration of minority carriers. The contrast is greatest near the maxima in the curves shown in fig. 5.

The total luminescence yield can be calculated by integrating the intensity $I(z)$ over a depth in the material equal to the excitation depth plus the diffusion length. The optical absorption in the semiconductor must also be included in the calculation.

Measurement of this intensity as a function of the electron energy at a constant power yields data that are used to determine the diffusion length. The results of such measurements also give information on the surface recombination and the effect of absorption in the semiconductor.

Fig. 6 gives the calculated luminescence yield as a function of the primary energy, with the diffusion length L as parameter. The calculations were carried out both for no absorption and for an absorption coefficient equal to 4×10^3 cm^{-1}. The experimental results are found to correspond to a low absorption coefficient and high surface-recombination rates. It is important to have a good knowledge of parameters of a particular material and to adapt the experimental conditions for each sample accordingly.

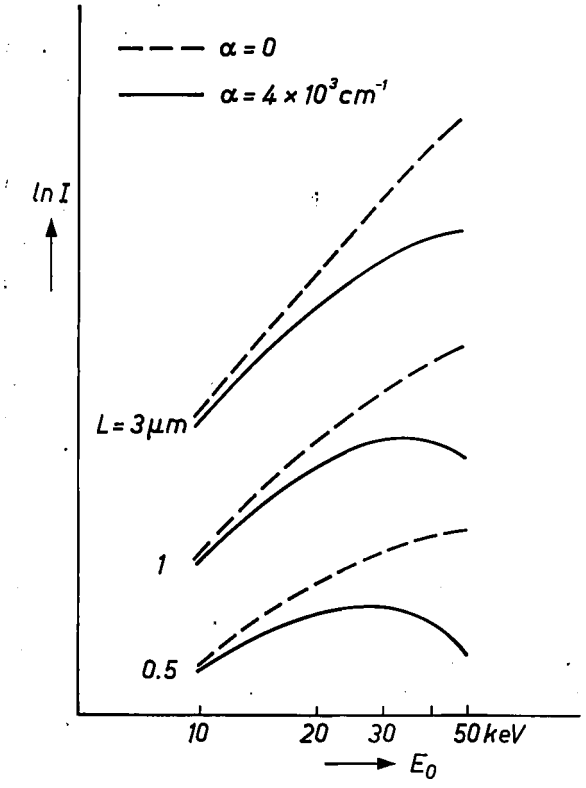


Fig. 6. Calculated total luminescence intensity I as a function of primary-electron energy. Calculations were made for three values of the diffusion length L and in each case for both zero and finite values of the absorption coefficient α for recombination radiation.

Apart from the visualization of crystal defects, which will be dealt with in the next section, it is possible in some cases to use the luminescence yield, because of its dependence on the diffusion length L, as a measure of certain material properties.

Experimental results

The luminescence images are easily obtained with a scanning electron microscope by placing a photomultiplier tube at a distance of a few centimetres from the specimen. The arrangement we used for this is shown in *fig. 7*. To increase the efficiency of detection of the luminescence, we surrounded the multiplier tube by a reflecting screen. The resultant increase in the aperture angle of the detector improved the signal by a factor of four. It is of course of great importance that the wavelength of the luminescence and the spectral sensitivity of the photocathode are well matched. The cathode we used was of the type S1, with a maximum sensitivity at 0.8 μm, decreasing to zero above 1.1 μm.

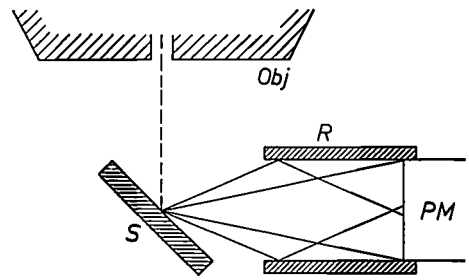


Fig. 7. The arrangement for cathodoluminescence experiments. *Obj* pole piece of the final lens of the scanning electron microscope. *S* sample. *R* reflecting screen. *PM* photomultiplier tube detecting the radiation resulting from electron bombardment of the sample.

Most semiconductors at present being developed, such as GaAs for infrared applications and GaAlAs and GaAsP for applications in the visible region, can be investigated in this way. The S1 photocathode gives a fairly high noise output at room temperature and therefore has to be cooled to −25 °C, which reduces the noise by a factor of about 1000. If the highly sensitive S20 cathode is used, the small amount of light from the electron gun in the scanning microscope is sufficient to produce a background signal that may be as much as 90 to 95% of the total.

The model we have given to describe the creation of luminescence radiation has general validity, but quantitative measurements are only possible on materials whose essential parameters are known and controlled — as they are in the case of most modern semiconductors.

We shall now make use of cathodoluminescence images obtained with a scanning electron microscope for discussion of some of our results on various semiconductor materials.

Measurements on GaAs

Fig. 8 shows pictures of a GaAs specimen taken with two primary electron energies: 17 and 35 keV. Two dislocations in these images are of interest; they both appear to be curved. The dislocation D_1 lies with its convex side towards the surface; it is visible in its entirety at low energies, but at higher energies the middle of the line has practically disappeared. The reverse applies to dislocation D_2, whose concave side is turned towards the surface. This reveals the influence of the surface layer, where the concentration of minority carriers is small as a result of surface recombination; this layer becomes thicker as the primary energy is increased.

Fig. 9 relates to a more heavily doped material. The result of this higher doping level is that dislocations are decorated by segregations. Unlike the previous case, this has the effect of making the contrast between a dislocation and its environment stronger than with stacking faults (*SF*), which produce pairs of dots in the image. Difference in contrast between the stacking faults are due to differences in the depth at which these faults are located.

The zone in which interaction between a dislocation and the surrounding lattice is noticeable depends on the dopant. If the material is doped with tellurium (fig. 9) a zone extending to ten microns around the dislocation is cleared of segregations. If the dopant is silicon (*fig. 10*) the zone is only two microns broad and segregations in the rest of the material are scarcely perceptible. The very small modulation that segregations cause here is due to the fact that they are too small to be made separately visible at the primary energy of 20 keV required to obtain an image. In the case of fig. 9 an X-ray examination showed that the Te concentration near the dislocation was about 10^{-3}. This corresponds to segregations with dimensions of about 0.1 μm, which is the order of magnitude also found in transmission

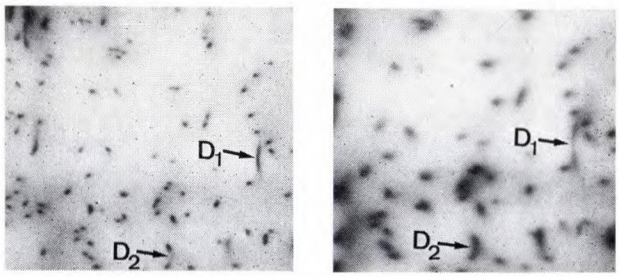


Fig. 8. Luminescence images of a GaAs sample at two values of primary-electron energy. *a*) 17 keV. *b*) 35 keV. The dislocation D_1 is clearly and completely visible at low energy, whereas at high energy only the two ends of the dislocation can be seen. The dislocation must therefore be curved, with the convex side towards the specimen surface. Similar arguments show that dislocation D_2 is also curved, but with the concave side towards the surface. Magnification 160×.

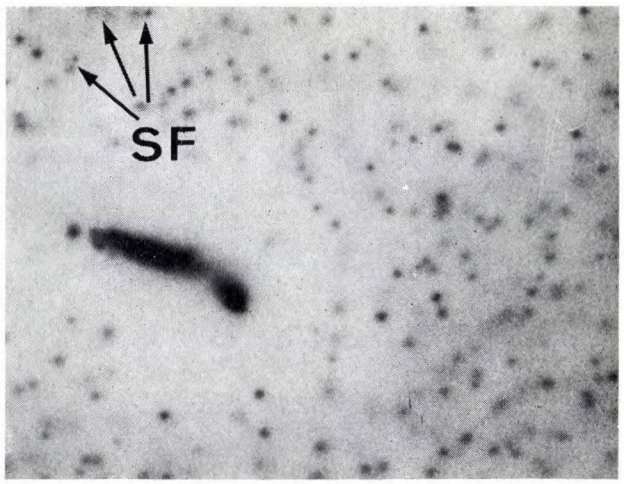


Fig. 9. Luminescence image of a GaAs sample more strongly doped than the sample of fig. 8. This high doping level has given rise to segregations, resulting in a darker background. Segregation occurs preferentially at dislocations (decoration), so that the region around a dislocation is relatively free from segregations. Stacking faults (*SF*), which appear as pairs of dots, are not decorated and consequently show a less marked contrast with the background and no segregation-free border. Magnification 900×.

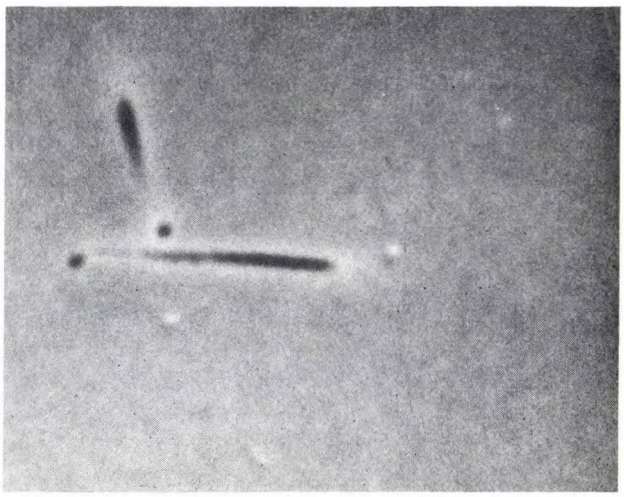
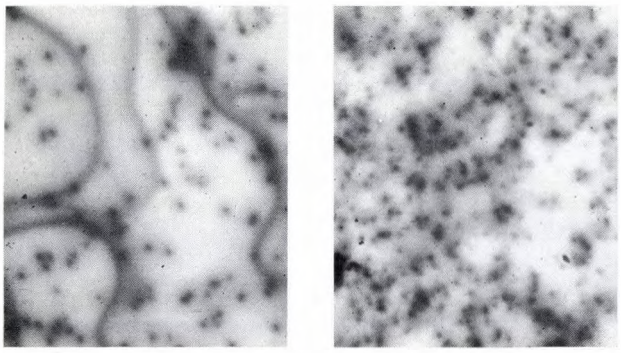


Fig. 10. Luminescence image of an Si-doped GaAs sample. Here again the dislocations are decorated. The segregation-free border is narrower than with GaAs; segregations are much smaller in size and are consequently not visible individually. Magnification 900×.



a *b*

Fig. 11. Luminescence images of two nitrogen-doped GaP samples at widely different dopant concentrations. The wavy lines in *a*) are due to surface steps. Magnification 270×.

electron microscopy. Semiquantitative measurements of the contrast in fig. 8*a* and fig. 9 show that an undecorated dislocation perturbs the lattice over almost a micron, whereas the corresponding area for the decorated dislocation in fig. 8 is considerably larger, more than 4 μm. In the latter case it was also found that the centre of the dislocation emitted no luminescence at all.

Observations on GaP

Luminescence images were also obtained for materials with an indirect band-to-band transition, such as GaP. *Fig. 11* shows images of this type obtained from two GaP samples with about the same charge-carrier density but with very different nitrogen contents. Fig. 11*a* shows various types of defect. The wavy lines are attributable to surface recombination at steps at the surface, like those sometimes found in epitaxial growth from the liquid phase. The spots that can be seen relate to dislocations, and their number corresponds to the density of dislocations in the substrate. The pairs of points relate to stacking faults. In fig. 11*b* the density of crystal defects is five or six times greater, and the variations observed in the contrast justify the conclusion that these are images of segregations. What the segregations consist of cannot be determined by this method. Cathodoluminescence is used here purely and simply for determining the number of defects; in the case of fig. 11*b* there were approximately 10^{10} segregations per cm^3.

Weakly doped and semi-insulating materials

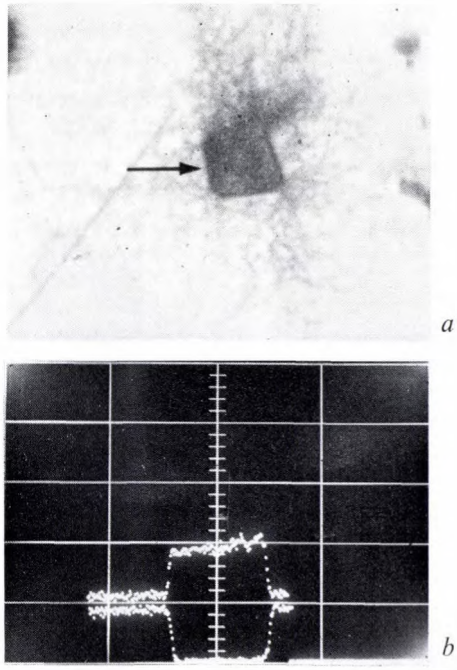
Semi-insulating and weakly doped materials are more difficult to study than the materials mentioned so far. The luminescence yield is low. The energy and also the current density in the primary beam must be high and both of these requirements entail a lower resolution. It is nevertheless possible to obtain luminescence images from these materials, as can be seen from the image of CdTe shown in *fig. 12a*.

Depending on the conditions in which the material is made, non-luminescent zones are found which vary in size from 5 to 50 μm or more. Electron-microprobe measurements showed that the zones consist of pure tellurium (fig. 12*b*). The luminescence image reveals that these tellurium inclusions are surrounded by a dense network of dislocations, which may extend over distances of 100 μm or more.

GaAsP layers

Apart from its value in determining intrinsic material properties by the measurements we have described, cathodoluminescence has also proved to be very useful as a means of monitoring epitaxial growth and diffusion treatments [5].

This is particularly useful in the preparation of layers that consist of mixtures of GaAs and GaP, materials in which the band gap depends on the ratio in which the two components are mixed. These materials are used in the manufacture of light-emitting diodes. The GaAsP is obtained by depositing an epitaxial layer on a GaAs substrate and then, starting from GaAs, slowly increasing the phosphorus content to the required value. Since the lattice constants of the substrate and the ultimate layer differ by 1.4 per cent, dislocations will occur locally to bridge this gap. The manner in which such dis-



a *b*

Fig. 12. *a*) Luminescence image of a CdTe sample. As the luminescence yield is low in this material, primary-electron energy and beam-current density must be relatively high, with a consequent loss in resolving power. The non-luminescent region in the centre turned out to be pure Te upon X-ray microprobe analysis. Magnification 150×.
b) Oscilloscope traces of X-ray microprobe signals obtained along a horizontal line at the level of the arrow in (*a*). The upper trace gives the TeLα intensity, the lower one the CdLα intensity.

locations are distributed has a very important bearing on the quality of the ultimate layer. This is affected by the dislocations in the substrate material, as they may propagate into the epitaxial layer.

Luminescence analyses performed on cleavage planes perpendicular to the growth planes enabled us to determine the conditions in which a GaAsP layer of the required structure can be deposited.

The luminescence images in *fig. 13* show what such layers may look like when grown under non-optimum conditions. Fig. 13*a*, relating to a cleavage plane perpendicular to an epitaxial layer, clearly shows the accumulation of dislocations in the transition zone and also their apparently complete absence in the final layer where composition is homogeneous. To facilitate the study of such growth processes a cross-section can be ground at an angle of 2° to the growth planes. Fig. 13*b* shows a section ground in this way, on which the details can be seen much more clearly than in the previous figure. Two zones can now be observed in the epitaxial layer and a plateau in the composition 15 μm away from the original GaAs substrate surface.

Fig. 14 shows part of a transition zone. The density of the dislocation network here is virtually in agreement with what might be theoretically expected from the phosphorus concentration gradient. With a not unduly high primary energy a resolution of approximately 0.3 μm can be achieved here. This is probably the best that can be expected bearing in mind the size of the area where the interaction takes place and of the diffusion length, which is less than 0.4 μm in this case.

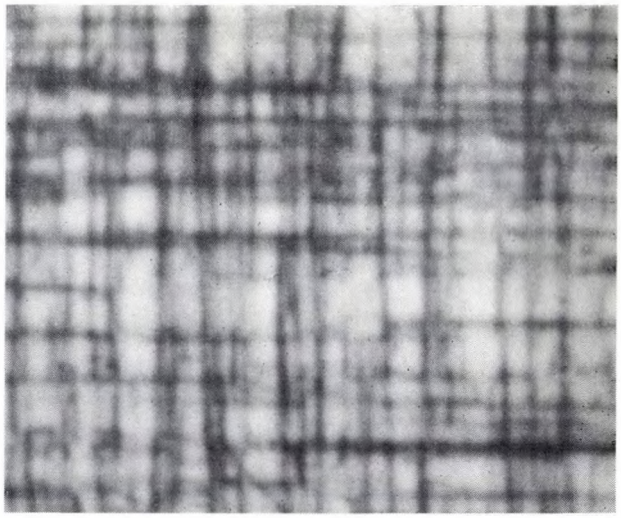


Fig. 14. Luminescence image of the dislocation network at the transition from GaAs to GaAsP. The dislocation density corresponds to values expected from the phosphorus-concentration gradient. Magnification 1900×.

Fig. 15 shows the surface of an epitaxial layer, where the contrast is constant and the dislocation density low. This explains why these dislocations are no longer observable in a cleavage plane.

The examples given above show that cathodoluminescence measurements permit the perfection of single-crystal material to be evaluated via a non-destructive method. The non-destructive nature of the investigation allows the method to be used to verify the separate steps in production processes in semiconductor technology. The results obtained are often in very

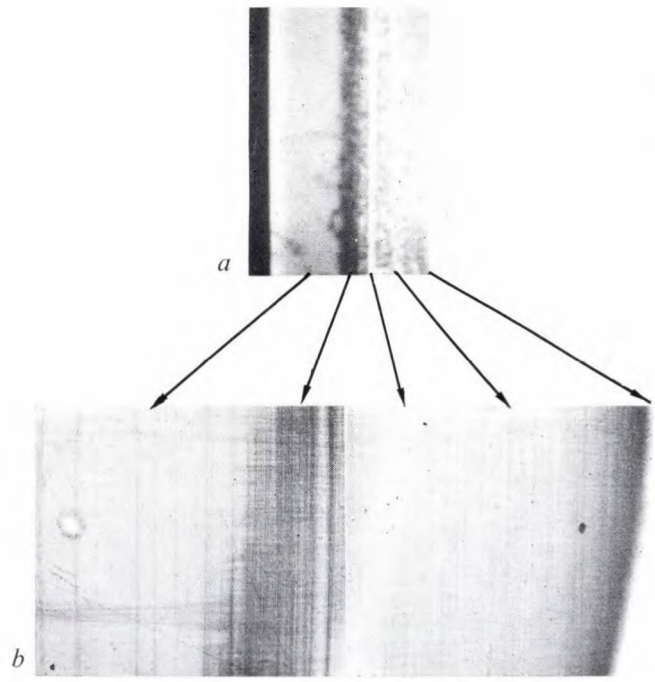


Fig. 13. Luminescence images of an epitaxial GaAsP layer grown on a GaAs substrate. *a*) A cleavage plane perpendicular to the growth plane of the layer. At the centre there is a high concentration of dislocations due to the 1.4 per cent mismatch in lattice constants between GaAs and GaAsP. *b*) Image of a surface ground at an angle of 2° to the growth plane. The variations in dislocation density can clearly be seen here. Magnification 540×.

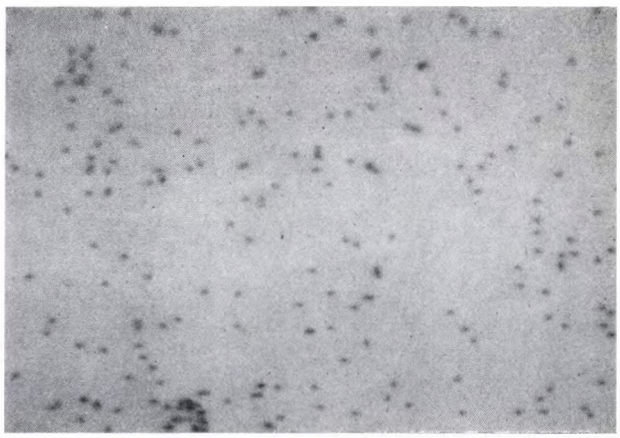


Fig. 15. Luminescence image of the surface of an epitaxial GaAsP-layer on a GaAs substrate. A low dislocation density has been obtained by careful adjustment of the gradients in the transition zone. Magnification 160×.

good agreement with results obtained by scanning electron microscopy. This is particularly so with electron-beam-induced current images [6].

Our results confirm the validity of the model used for the interaction between primary electrons and the solid under investigation. The optimum spatial resolution predicted from this model can, however, only be realized in materials that show a good luminescence yield at relatively low primary energies. A resolution of about 1 μm is readily obtained for materials where the band gap corresponds to the wavelengths at which the modern photocathodes are most sensitive.

[6] M. R. Burgess, H. B. Haanstra and C. Schiller, Acta Electronica **18**, 15, 1975 (No. 1).

Summary. Scanning electron microscopy in which images are obtained from cathodoluminescence is a rapid and non-destructive method for investigating crystal defects in a number of materials. Electron bombardment creates electron-hole pairs. If the band gap in the solid has a suitable width, recombination will result in luminescence. At the site of crystal defects (e.g. dislocations, stacking faults, or segregations) non-radiative recombination occurs and these defects will appear as black regions in the scanning microscope image. The various steps in the process can be incorporated in a model that also allows for diffusion of minority carriers between creation and recombination. A number of results are shown that illustrate the usefulness of this mainly non-destructive test method, particularly in the investigation of semiconducting III-V compounds for optical applications.

Pyroelectric materials based on triglycine sulphate (TGS) for infrared detection

E. T. Keve

Since the introduction of the ferroelectric triglycine sulphate (TGS) as a pyroelectric material most of the work on pyroelectric detectors has been based on it. A fundamental difficulty with the original material is its tendency to depole. However, it was discovered a few years ago that this tendency could be inhibited by partially substituting asymmetrical l-alanine molecules for symmetrical glycine molecules. The doped crystals are permanently polarized and consist of a single polarization domain. This article reviews recent work by Mullard Research Laboratories (Redhill, Surrey) and Mullard Southampton on pyroelectric detectors based on l-alanine-doped triglycine sulphate and selenate. An explanation is given of the inhibiting action of l-alanine on the depoling process in these crystals.

Introduction: pyroelectric effect

The pyroelectric effect is the development of an electrical polarization in a crystal due to a temperature change. The effect can in theory appear in all crystals whose symmetry belongs to one of the ten polar point groups. These crystals consist of ions or molecules that form groups of electric dipoles arranged in such a way that they have a non-zero resultant along the polar axis. This state of electrical asymmetry, or polarization, is not generally observable, because under equilibrium conditions it is compensated by the presence of free carrier charges that have reached the surface of the crystal by conduction through the crystal and from outside (*fig. 1a*). The magnitude of the polarization, and hence of the compensating charge, depends on the temperature. If the temperature of the crystal is changed so that the supply of compensating charges cannot keep pace with the variation in polarizing charge, then the crystal surfaces acquire an observable charge. If the crystal is made into a capacitor by providing the surfaces perpendicular to the polar axis with electrodes, then the redistribution of the compensating charges causes a current to flow in an external circuit connecting the electrodes (fig. *b*). For a small temperature change dT, the change dP in the polarization of a crystal (and hence in the compensating charge per unit area of polar crystal surface) is proportional to the temperature change: $dP = \lambda dT$, or $\lambda = dP/dT$. The proportionality constant λ is called the pyroelectric coefficient of the crystal.

If the temperature change of the crystal is caused by the absorption of infrared radiation, the observation of the pyroelectric effect is equivalent to the detection of the incident radiation. The crystal can therefore be used as an infrared (radiation) detector. Such a detector

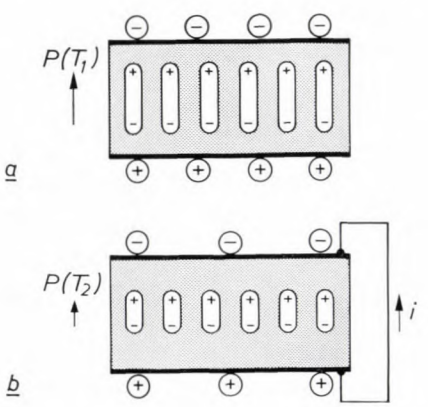


Fig. 1. *a*) A pyroelectric crystal in equilibrium at a temperature T_1. The crystal contains dipoles that all have the same orientation and together produce a polarization P (dipole moment per unit volume). The polarization charge on the crystal surfaces perpendicular to the direction of polarization (P units of surface-charge density) is compensated by free charge carriers in the equilibrium state. *b*) For a temperature rise from T_1 to T_2, by absorption of radiation, the polarization changes from $P(T_1)$ to $P(T_2)$ (indicated by the shorter dipoles). The establishment of a new state of equilibrium at the surface is associated with a current i in the external circuit.

Dr E. T. Keve is with Mullard Research Laboratories, Redhill, Surrey, England.

is not sensitive to static temperatures; it reacts only to temperature *changes*. The current i in the external circuit is proportional to the area of the radiation-receiving electrode A and also to the rate at which the polarization charge changes:

$$i = A\frac{dP}{dT}\cdot\frac{dT}{dt} = A\lambda\frac{dT}{dt}.$$

Application of the pyroelectric effect to infrared detection was suggested more than 30 years ago [1]. In practice, however, the performance of such devices was poor, partly because of the limitations of the pyroelectric materials then available. Only recently, after the introduction of triglycine-sulphate crystals, was there any significant improvement in the prospects for pyroelectric devices. Improved understanding of the properties of this material has contributed to the emergence of a new class of infrared detectors. Some of the most important developments were the improvements that could be obtained by doping triglycine sulphate with other substances. The pyroelectric instruments now available for infrared detection give an excellent performance and are convenient and versatile [*]. They can be applied wherever other thermal detectors have been used, in laboratory applications for general problems such as power measurements and checking optical alignment, in spectroscopy and work with lasers, in intruder alarms and in satellite instrumentation, in particular for horizon sensing and meteorological measurements. Pyroelectric materials can also be used in thermal imaging, either incorporated in linear arrays of detectors, or as a target in a vidicon-type imaging tube.

This article includes a short survey of the work at Mullard Research Laboratories and Mullard Southampton that has led to the development of the improved materials mentioned above. Their properties and their physical basis are then discussed, and brief accounts are given of crystal preparation, device construction and the characteristics of three experimental infrared detectors. In the first section, however, the relation between device performance and the properties of the material is discussed. Since the performance of a device also depends on other factors, such as amplifier characteristics, frequency and geometry of the device, the quality of performance of particular materials can only be assessed with a particular type of application in mind.

The noise equivalent power of pyroelectric detectors

Before considering recently developed pyroelectric materials in detail, it is perhaps appropriate to consider the characteristics of a pyroelectric detector. Such a detector consists of a pyroelectric crystal with electrodes on the surfaces perpendicular to the polar axis, and an associated amplifier. To obtain a signal from a constant radiation flux, the incident radiation is periodically interrupted ('chopped') at a frequency that can vary from a few Hz to more than 10^4 Hz. The amplifier then gives a constant a.c. voltage at this frequency.

The usefulness of such a device in detecting small signals is expressed by the *n*oise *e*quivalent *p*ower (NEP), for which the signal is equal to the noise from the detector. Noise arises from several sources in the detector: temperature or radiation noise, Johnson noise (sometimes called thermal noise) in the crystal, and two components in the amplifier: current noise and voltage noise. The NEP is the sum of the contributions from these four noise sources. Consideration of the equivalent circuit of the detector device (*fig. 2*) leads to the following expression for the NEP per unit bandwidth [2] [3]:

$$NEP = \frac{1}{\eta}\sqrt{4kT\left\{TG_T + \frac{H_T^2 + (G_T/\omega)^2}{\lambda^2 A^2}\times\left[G_N + G_E + R_N(G_E^2 + \omega^2 C_E^2)\right]\right\}}. \qquad (1)$$

Here ω is the central frequency of the interval of unit bandwidth considered, η the emissivity of the element, k Boltzmann's constant, T the absolute temperature of the detector, G_T the thermal conductance between the element and its surroundings, H_T the thermal capacity of the element, A the electrode area exposed to radiation, G_E the electrical conductance of the element, C_E the total capacitance across the element (including strays), G_N the noise conductance of the amplifier and R_N the noise resistance of the amplifier. ($G_N = 2\pi i_N^2/4kT\Delta\omega$ and $R_N = 2\pi V_N^2/4kT\Delta\omega$, where i_N is the noise current and V_N the noise voltage.)

It is usually valid to neglect thermal-noise contributions, and to consider only frequencies greater than

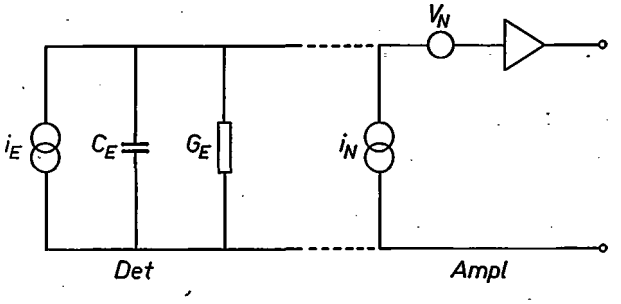


Fig. 2. Equivalent circuit of pyroelectric detector (*Det*) and amplifier (*Ampl*). The generator of the pyroelectric current is represented by i_E, the total capacitance of the detector element by C_E and its electrical conductance by G_E. The generators i_N and V_N represent the amplifier current noise and the amplifier voltage noise respectively.

G_T/H_T. The expression for the NEP then simplifies to:

$$NEP = \frac{dc'}{\eta\lambda}\sqrt{4kT\left\{G_N + G_E + R_N(G_E^2 + \omega^2 C_E^2)\right\}}. \quad (2)$$

(1) (2) (3)

where d is the thickness of the element and c' its specific heat per unit volume. The three labelled parts of the expression refer to the contributions from the amplifier current noise (*1*), the Johnson noise in the element (*2*), and the amplifier voltage noise (*3*). The schematic indication in *fig. 3* of the variation with frequency and area A of the NEP shows the successive dominance of each of these contributions to the noise.

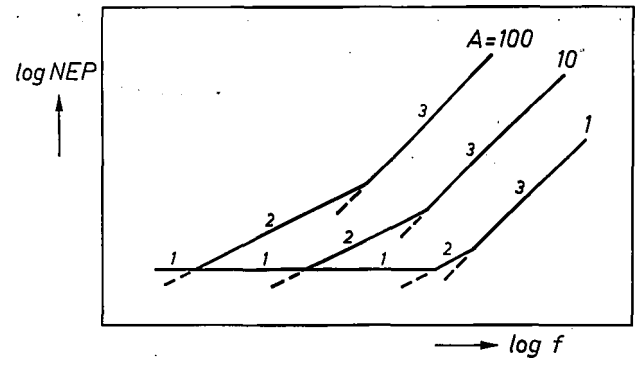


Fig. 3. Variation (schematic) of noise equivalent power *NEP* with frequency f for three values of the irradiated electrode area A. In region *1* the amplifier current noise is dominant; in *2* the Johnson (thermal) noise in the pyroelectric crystal and in *3* the amplifier voltage noise.

The properties required of the pyroelectric material to minimize the NEP clearly depend on the operating conditions under which the detector is used. Three different kinds of situation can be identified from fig. 3, and for each a figure of merit can be deduced from expression (2): λ/c' for low frequencies and small areas (*1* in fig. 3), $\lambda\varrho^{\frac{1}{2}}/c'$ for intermediate frequencies and areas (*2* in fig. 3), and $\lambda/c'\varepsilon$ for high frequencies and large areas (*3* in fig. 3), where ϱ is the electrical resistivity and ε the dielectric constant of the pyroelectric material at the frequency of interest.

These figures of merit can serve as a guide in evaluating materials, but it must be remembered that the change from one regime to the next is itself a variable depending on material and amplifier parameters.

The two parameters that are important under all conditions are λ and c'. As c' has about the same value for different materials it is λ, the pyroelectric coefficient, that is required to be as large as possible in all applications.

Pyroelectric materials

Of the large number of crystals in which the pyroelectric effect appears, crystals of the relatively small group of ferroelectric materials have the largest pyroelectric coefficients. These ferroelectrics are distinguished from other pyroelectrics by the fact that they have a spontaneous polarization, which can be reversed by applying an external electric field. Moreover, the polarization disappears as soon as a particular characteristic temperature of the material — the Curie temperature — is reached. At this temperature there is a phase transition from the polar to a non-polar crystal phase. In general at temperatures below the Curie temperature pyroelectric coefficients can be observed that are 10 to 100 times higher than for a typical non-ferroelectric pyroelectric material.

Ferroelectrics do however have some disadvantages. They generally have high, temperature-dependent dielectric constants, which have their highest values at temperatures close to the Curie temperature. Since the polarization is reversible, the crystals will generally consist of a number of domains with opposing polarizations. This means that an electric field has to be applied before use, to give the crystal a unique sense of polarization. In itself this is no problem, but since in a detector the polarizing field must be removed to achieve low-noise performance, the possibility exists that domains of the opposite polarization may arise later. This 'depoling' causes a reduction of the detector signal.

The pyroelectric coefficients of ferroelectrics are however so high that other materials are hardly ever used in pyroelectric devices.[*] Until recently the best material for a wide range of areas and frequencies was ferroelectric triglycine sulphate (TGS). The new materials that have been developed at Mullard Southampton [4] are chemically closely related to triglycine sulphate, but do not depole. As an introduction to the discussion of these materials triglycine sulphate must first be described.

Triglycine sulphate

Triglycine-sulphate crystals are monoclinic with space group $P2_1$ below the Curie temperature (49 °C) and space group $P2_1/m$ above it. Single crystals of TGS are grown from aqueous solutions. The crystals have a well defined cleavage plane perpendicular to the polar

[1] Yeou Ta, C.R. Acad. Sci. 207, 1042, 1938.
[2] E. H. Putley, in: R. K. Willardson and A. C. Beer (eds), Semiconductors and semimetals, Vol. 5, Academic Press, New York 1970, p. 259.
[3] G. Baker, D. E. Charlton and P. J. Lock, Proc. Joint Conf. on Infra-red techniques, Reading 1971 (I.E.R.E. Conf. Proc. No. 22), p. 1. See also The Radio and Electronic Engineer 42, 260, 1972.
[4] P. J. Lock, Appl. Phys. Lett. 19, 390, 1971.
[*] These detectors are commercially available as a product of Mullard Ltd., England.

axis. The crystal structure is relatively complex [5] (*fig. 4*), with three independent glycine molecules per unit cell below the Curie temperature. The glycine molecules and the sulphate ions (shown schematically as tetrahedra in fig. 4) are held together by a network of hydrogen bonds. The polar axis is *b*. When the crystal is heated above the Curie temperature the polarity of the *b*-axis is lost, and a new symmetry element appears in the form of a set of mirror planes perpendicular to the *b*-axis.

The temperature dependence of the polarization P, the pyroelectric coefficient λ and the dielectric constant ε of a typical TGS crystal are shown in *fig. 5*. The polarization is seen to go to zero at the Curie temperature; the pyroelectric coefficient peaks at this temperature but is zero above it. Consequently pyroelectric devices are limited in their operation to below the Curie temperature.

TGS appeared to be a very suitable pyroelectric material for devices to be used at about 30 °C below the Curie temperature. The best compromise between performance, resistance to depoling and a reasonable temperature range of operation is then achieved. Numerous pyroelectric devices have in fact been fabricated from TGS crystals in the late 1960s. Now, however, they have been surpassed by devices based on doped varieties of TGS. These materials will be discussed in the following sections.

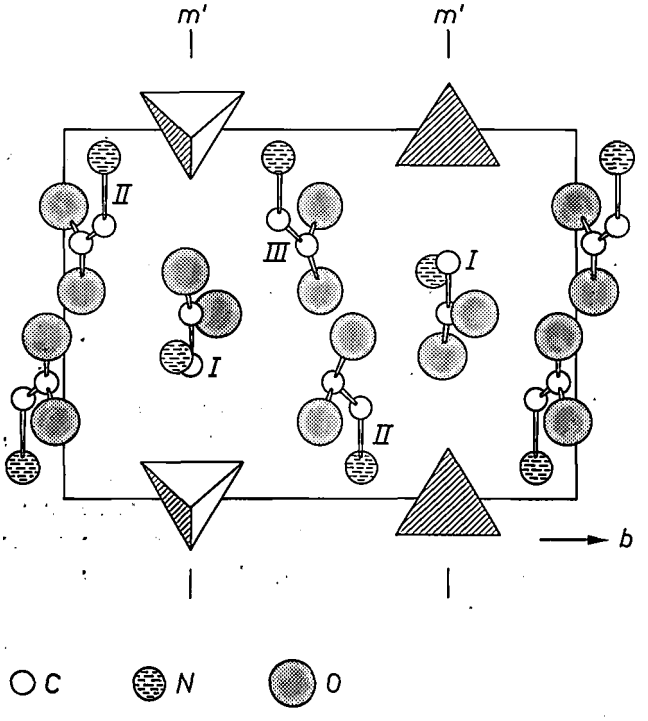


Fig. 4. Schematic structure of a triglycine-sulphate (TGS) crystal, seen along the *c*-axis. Glycine molecules occur at three different positions (indicated by *I*, *II* and *III*). The sulphate ions are represented schematically by tetrahedra. All the hydrogens have been omitted. *m'* represents the sets of pseudo mirror planes in which glycine-*I* molecules are inverted on ferroelectric switching.

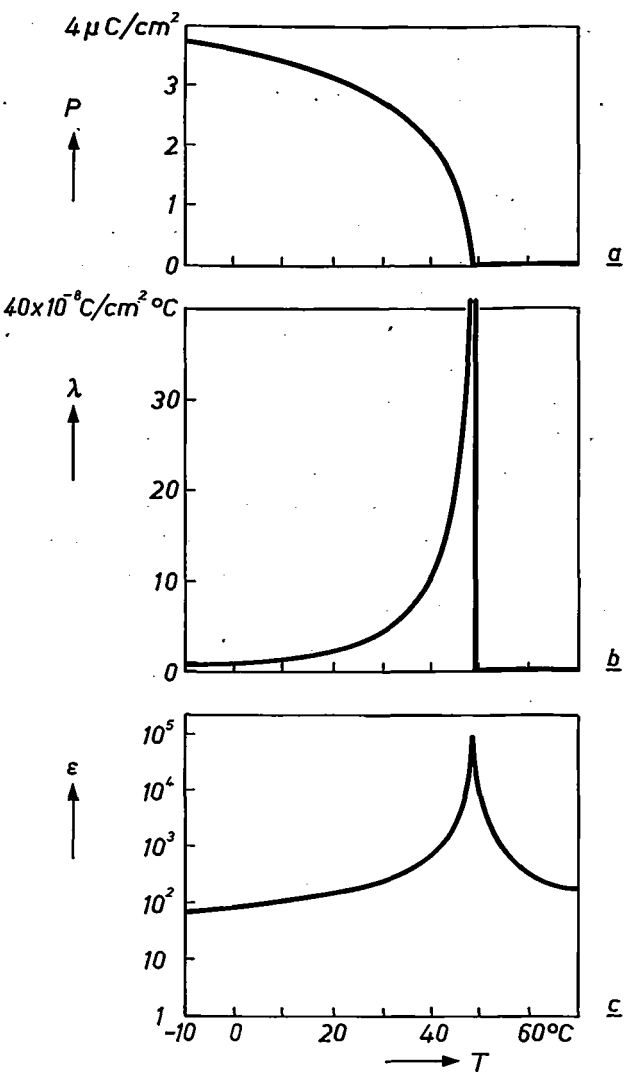


Fig. 5. The temperature dependence of the polarization P, the pyroelectric coefficient λ and the dielectric constant ε for a crystal of pure TGS ($\lambda = dP/dT$).

Alanine-doped triglycine sulphate (LATGS)

When crystals are grown from a solution of TGS to which the related material l-alanine has been added, some of the l-alanine is taken up by the TGS crystal lattice. The crystals obtained by this method also have a number of special properties that are highly relevant for pyroelectric applications [4]. The difference between crystals of pure TGS and of LATGS is most conveniently illustrated in terms of the dielectric hysteresis loops. The loop for a pure TGS crystal (*fig. 6a*) is typical of an ordinary ferroelectric: it indicates the existence of two equally possible (opposite) orientations of the spontaneous polarization at zero field. Fig. 6*b* shows that the loop for a doped crystal is displaced along the field axis by an amount E_b. This behaviour has also been observed in other ferroelectrics, e.g. colemanite [6], and is believed to arise from the pres-

ence of a permanent internal biasing field E_b [7]. The presence of an internal biasing field is of primary importance because it makes the crystal permanently single-domain. This overcomes the major disadvantage of the depoling of ferroelectric materials in pyroelectric devices. The internal biasing field affects other properties of the crystal, but this is either directly advantageous, or can be made so, as will be shown later.

Comparison of fig. 5 and *fig. 7* indicates that the polarization no longer goes directly to zero at the Curie temperature for a TGS crystal containing 0.2% of l-alanine. The peaks in the curves for the pyroelectric and dielectric constants are both depressed. The value of the pyroelectric coefficient of l-alanine-doped TGS at room temperature is slightly higher than that of pure TGS. This can give better response and lower values of the noise equivalent power. Another significant finding is that the a.c. conductance in LATGS crystals is considerably lower than in crystals of pure TGS. This leads directly to improved device performance in regime (*2*) of fig. 3.

The properties of doped crystals vary with the alanine content [8]. This can be seen for the dielectric constant and polarization in *fig. 8*, where these quantities are shown as functions of the temperature for crystals obtained from solutions of TGS with various values for the l-alanine content (0, 10, 20, 40 and 60 mol% for the curves *1* to *5*) [*]. From fig. 8*a* it can be seen that with increasing l-alanine content the maximum value for the dielectric constant decreases considerably (only 120 in curve *5*) and the temperature at which this maximum occurs is increased. In addition, fig. 8*b* shows that the transition from the polar to the non-polar phase occurs in a wider temperature range. This broadened transition has a significant effect on pyroelectric-device application, since highly doped LATGS crystals remain polar well above the Curie temperature of pure TGS crystals (49 °C) and also above the temperature at which the maximum of the dielectric constant occurs. *Fig. 9* shows the pyroelectric a.c. current output as a function of temperature from 20 to 90 °C from a crystal with the highest alanine content (corresponding to curve *5* in fig. 8), with an internal field of about 90 kV/cm. The temperature of the crystal was raised gradually while the current was being measured; the

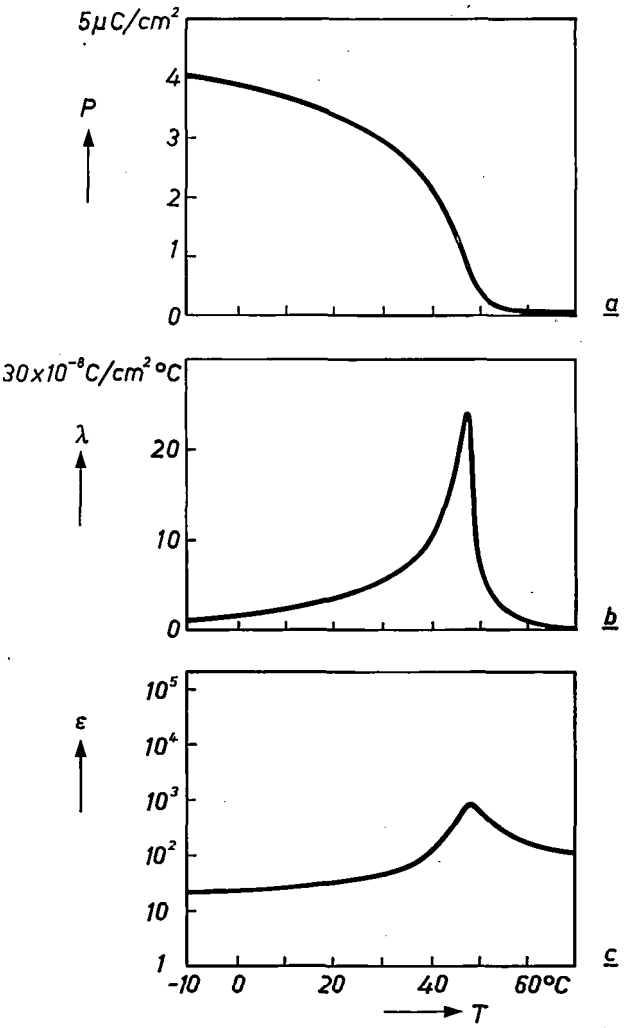


Fig. 7. The temperature dependence of the polarization P, the pyroelectric coefficient λ and the dielectric constant ε for a crystal of l-alanine-doped TGS (LATGS). The polarization P does not go directly to zero at the Curie temperature (49 °C) and the peak in the ε curve is much lower than in the case of TGS (see fig. 5).

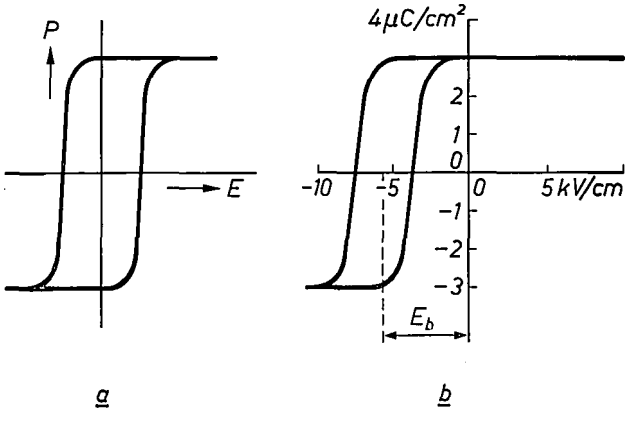


Fig. 6. Hysteresis loop for crystals (polarization plotted against electric field-strength) of pure TGS (*a*) and of l-alanine doped TGS (LATGS) (*b*). For LATGS the hysteresis loop has been displaced by an amount E_b due to the presence of an internal biasing field. The effect of this field is that the crystal is in its maximum polarization state at zero applied field; the crystal contains only one polarization domain. In highly doped crystals E_b can be of the order of 100 kV/cm.

[5] S. R. Fletcher, A. C. Skapski and E. T. Keve, J. Physics C 4, L 255, 1971, and Ferroelectrics 8, 479, 1974.
[6] H. H. Wieder, J. appl. Phys. 30, 1010, 1959.
[7] E. T. Keve, K. L. Bye, P. W. Whipps and A. D. Annis, Ferroelectrics 3, 39, 1971.
[8] K. L. Bye, P. W. Whipps and E. T. Keve, Ferroelectrics 4, 253, 1972.
[*] The l-alanine content of the crystals is 20 to 30 times lower than that of the solution.

incident infrared radiation was chopped at a convenient frequency. Even at a temperature of 90 °C — 40 °C above the Curie temperature of pure TGS — the crystal still gives an appreciable pyroelectric current.

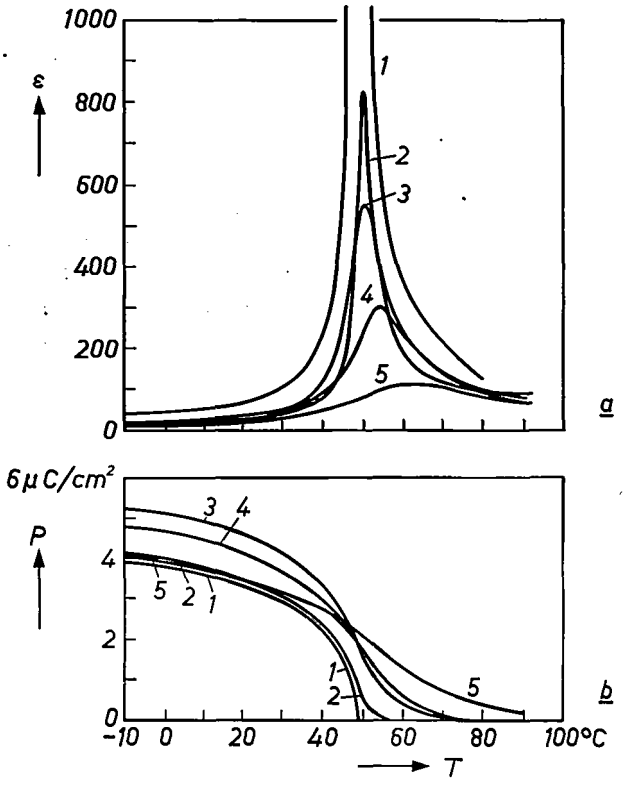


Fig. 8. The temperature dependence of the dielectric constant ε and the polarization P for crystals of pure TGS (curve *1*) and of alanine-doped TGS crystals grown from solutions containing 10 (curve *2*), 20 (curve *3*), 40 (curve *4*) and 60 (curve *5*) mol% of l-alanine respectively.

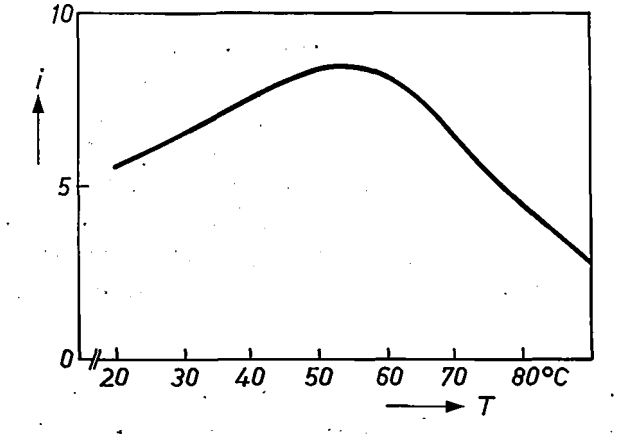


Fig. 9. The temperature dependence of the pyroelectric a.c. current i for an l-alanine-doped crystal (LATGS) grown from a solution containing 60 mol% of l-alanine. During the measurements the temperature of the crystal of the detector element was gradually raised, while the incident infrared radiation was chopped at a convenient frequency.

Substitution mechanism in alanine-doped TGS

Before proceeding to examine further the effects of alanine doping it is of interest to consider the underlying mechanism.

Alanine is an amino acid like glycine (*fig. 10a*), but with a CH_3 group replacing one of the α-hydrogen atoms. The model proposed [7] to explain the doping effects of alanine in TGS postulates that the similarity is sufficient for alanine to be substituted at glycine sites in doped crystals.

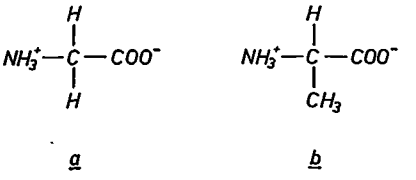


Fig. 10. Formulae of glycine (*a*) and alanine (*b*).

Fig. 11a again shows the glycine molecule labelled *1* in the schematic crystal structure of fig. 4. When the polarity of the crystal switches, this molecule must switch to its mirror image shown in fig. 11*c*, passing through the intermediate stage of mirror symmetry with respect to a plane (fig. 11*b*) normal to the pyroelectric axis *b*. Since the free glycine molecule has mirror-plane symmetry, the orientations of fig. 11*a* and *c* in the crystal lattice are equally possible. It is the possibility of transition from one orientation to the other that decides the bistable polarization state of TGS.

The crucial difference between glycine and alanine is that the alanine molecules do not have mirror-plane symmetry. Two forms exist, d- and l-alanine, each the mirror image of the other. Consequently there is only one possible orientation of the l-alanine molecules in the TGS crystal lattice, and only one possible orientation for the polarization. Inversion of the polarization would only occur if l-alanine could be transformed into its mirror image, d-alanine, but chemical bonds would then have to be broken.

The argument given above only holds for single alanine molecules, of course. However, the presence of a relatively small concentration of alanine molecules is sufficient to inhibit the reversal of polarization of a whole crystal. The explanation for this phenomenon may be as follows. Local strains must be introduced in the crystal lattice where glycine is replaced by a larger molecule. In the case of alanine, the strain field must be asymmetrical (i.e. polar) due to the asymmetry of the deviation of alanine from glycine. Furthermore, the polarity of the strain field cannot be inverted as the substituent molecule cannot change into its mirror image. The alanine molecule and its associated lattice distortion are equivalent to a dipole with relatively long-

range influence. This dipole is both irreversible and permanent as long as the substituent molecule remains intact. If the LATGS crystal contains only molecules of one of the optically active forms of the additive, e.g. l-alanine, all the alanine molecules must lie in the same sense with respect to the pyroelectric axis if the stereochemistry and the hydrogen bonding are to be preserved. All the permanent dipoles, therefore, have the same polar sense. A sufficiently high concentration of them will result in a self bias on the crystal, which can polarize the intervening matrix of pure TGS. A spontaneously and permanently poled single crystal is then produced.

If a 'racemic' mixture of equal quantities of d- and l-alanine is incorporated in TGS crystals, then crystals with large permanently poled domains are formed below the Curie temperature, some with one polarization, others with the opposite one.

and are symmetrical in the free state, but are almost certainly asymmetrical in the crystal (fig. 12*c*).

So far, of all the substituents tested alanine has proved to be the most convenient. It is used in a single optically active form (in practice l-alanine) since a unique sense of the pyroelectric or polar axis is clearly preferred in crystals for use in pyroelectric devices.

Mixed crystals of triglycine sulphate and selenate, doped with alanine

In an earlier section it was pointed out that although pyroelectric coefficients of TGS were very high at temperatures close to the Curie temperature, this could not be put to use in pyroelectric devices because of the tendency to depoling and the high values of the dielectric constants at such temperatures. For LATGS crystals with a high level of alanine doping, however, these considerations are no longer important, since such crystals cannot depole and the lower dielectric constant near the Curie temperature is not a limiting factor in detector performance. In fact, highly doped alanine crystals give optimum performance in the region of 40-50 °C. However, it is generally undesirable to heat the pyroelectric crystal in a device to these temperatures. The need for this heating can be avoided if instead of triglycine sulphate, crystals are used in which some, or all, of the sulphate ions are replaced by selenate ions.

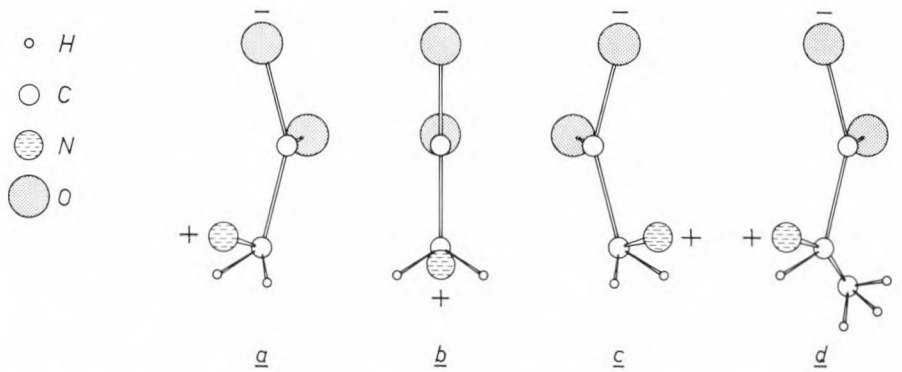


Fig. 11. Schematic diagram of the glycine molecule that takes up position *I* in fig. 4, now shown with its hydrogen atoms. *a*) One of the two polarization states. *b*) The non-polarized state. *c*) The polarization state opposite to that of (*a*). *d*) An l-alanine molecule of fixed polarization state.

The above model suggests that any molecule that can replace the glycine molecule labelled *I* in the crystal structure shown in fig. 4 and, like alanine, does not have a mirror plane of symmetry will behave similarly to l-alanine in LATGS cyrstals. It has indeed been observed that substituents such as serine (*fig. 12a*) and α-amino-n-butyric acid (fig. 12*b*) give results similar to alanine. It has also been observed that the same is true for molecules such as sarcosine that can replace alanine

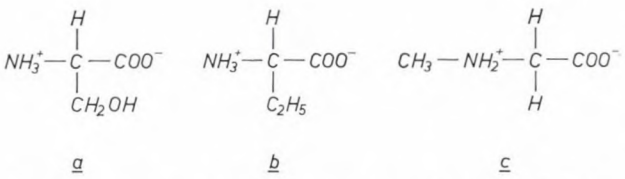


Fig. 12. Formulae of serine (*a*), α-amino-n-butyric acid (*b*) and sarcosine (*c*). These are other possible substituents for a glycine molecule (*I* in fig. 4) in TGS.

Triglycine selenate and sulphate have the same crystal structure. The selenate is a ferroelectric with a Curie temperature of 22 °C. Mixed crystals can be grown with any composition ratio, and by choosing the composition ratio appropriately crystals can be 'tailored' to have Curie temperatures anywhere between 22° and 49 °C. Moreover, these mixed crystals can be per-

Fig. 14. Various types of pyroelectric infrared detector made by Mullard Southampton. All these detectors contain crystals of alanine-doped TGS. *Top picture:* the most usual configuration (see fig. 13). *Left:* a version in which the crystal is mounted on the end of a short post, for a special application. *Immediately above:* an experimental miniature detector in a TO5 package. *Below:* an experimental array for a thermal imaging system, consisting of 128 detectors on one crystal, each detector with its own amplifier; the connector strips can be seen in front of the housing and behind it.

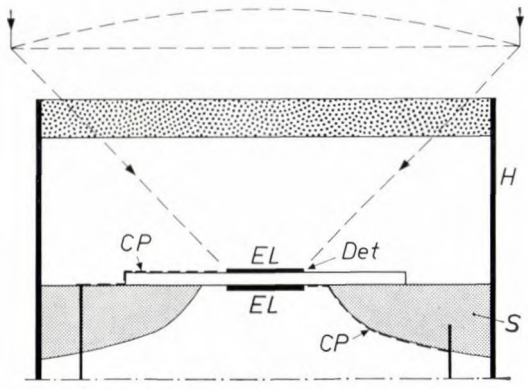


Fig. 13. Schematic cross-section of the optical unit of a pyroelectric infrared detector. *Det* wafer of a pyroelectric crystal. *EL* evaporated metal electrodes. *S* ceramic substrate, with aperture a little larger than the electrodes. *H* housing. *W* window. The electrodes are connected via strips of conducting paint *CP* to the input leads of an amplifier stage. The detector is usually used in combination with an optical system, e.g. a lens (as shown here) or an array of slits.

manently poled by doping with l-alanine [9]. The crystal composition can therefore be adjusted to give the most desirable combination of properties at ambient temperatures. With this approach it is possible to design devices that give their highest performance at room temperature. The reduced temperature range of operation of such devices compared with those in which doped LATGS crystals are used may be a disadvantage in some applications. This temperature range, however, increases with increased l-alanine content in the material.

Crystal growth; construction and performance of pyroelectric detectors

Single crystals of the TGS family of materials are grown from a solution in water of sulphuric acid, glycine and the selected dopant. Seed crystals are mounted on platforms in containers of the solution (10 to 20 litres). Each platform is made to rotate or oscillate to stir the solution, which is cooled at a rate of a few tenths of a degree per day. The crystals produced in this way have a mass of about 100 g. They have a characteristic habit (external shape) and are of excellent optical quality. The crystals are cut perpendicular to the polar axis and the plates thus obtained are polished to thicknesses in the range 5-20 μm. Electrodes are evaporated on to the (010) surfaces of the plates, usually with a thick deposit of nichrome for the back electrode and a semi-transparent electrode on the front surface. This front electrode does not have to be blackened since the crystals themselves are excellent absorbers in the infrared. The crystal is then mounted on a ceramic substrate in which there is an opening slightly larger than the electrodes. The detector assembly is placed behind a window inside a housing in which there is also a high-input-impedance amplifier stage (*fig. 13*). The final shape, the question of hermetic sealing or evacuation, the choice of window material and of detector area all depend on the application. *Fig. 14* shows some examples of the many types that have been constructed.

Results of measurements made with three devices having detector crystals of different materials of the TGS family illustrate some aspects of the interrelationships of performance and material. The first detector has a TGS crystal doped with alanine; this gives the best low-frequency performance measured. It has been designed for use at room temperature, but it can also be used to about 50 °C. The second detector has a similar crystal with a stronger internal field, and can therefore be used to about 70 °C. The third detector is made from alanine-doped TGSe. This detector gives results inferior in comparison to those of the first detector, but demonstrates that the material is potentially the best for low-frequency performance at room temperature.

Performance of thermal detectors is often expressed in terms of the normalized or specific detectivity D^*, i.e. the root of the area A of the electrode exposed to the radiation, divided by the noise equivalent power: $D^* = A^{\frac{1}{2}}/NEP$.

The best detectors based on alanine-doped TGS can give NEP values as low as 4×10^{-11} $WHz^{-\frac{1}{2}}$ at low frequencies with a corresponding D^* of 5×10^{9} cm $Hz^{\frac{1}{2}}W^{-1}$. The variation of both of these quantities as a function of frequency is shown for an LATGS detector in *fig. 15*.

The curves for D^* in fig. 15 are put into a wider context in *fig. 16* by a comparison with thermal infrared detectors of other types. It can be seen that the per-

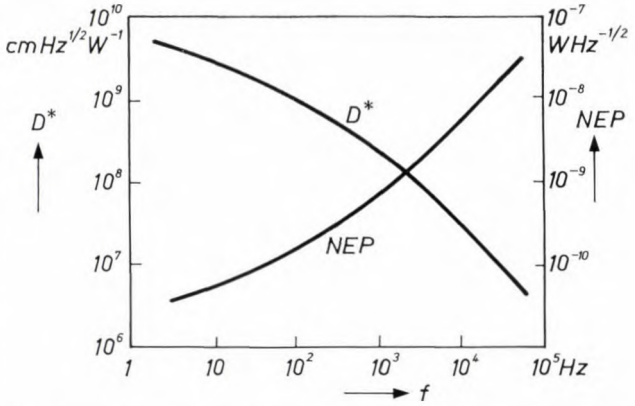


Fig. 15. The frequency dependence of the noise equivalent power NEP and the detectivity D^* for a detector with an LATGS crystal as pyroelectric element. ($D^* = A^{\frac{1}{2}}/NEP$; A is the area of electrode exposed to radiation.)

[9] K. L. Bye, P. W. Whipps, E. T. Keve and M. R. Josey, Ferroelectrics 7, 179, 1974.

formance already achieved with pyroelectric instruments surpasses the other types in D^* value at all frequencies. Further improvements in pyroelectric performance are still to be expected from the use of the more recently developed mixed crystals of triglycine sulphate and selenate.

As mentioned above, detectors made from TGS crystals with an average alanine content give no signal above 50 °C, but if the alanine content is high they are still pyroelectric at 90 °C (see fig. 9). *Fig. 17*, showing *NEP* plotted against temperature for this latter case, demonstrates that such a detector actually retains good performance right up to 70 °C. *Fig. 18* shows that the NEP is almost constant between 10 and 30 °C for a mixed crystal with 70% sulphate and 30% selenate. The minimum value of *NEP* for this detector

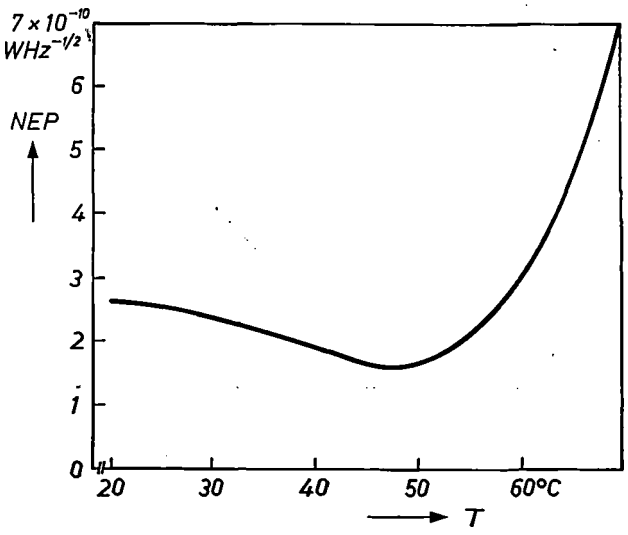


Fig. 17. The temperature dependence of the noise equivalent power *NEP* for a pyroelectric infrared detector with a highly l-alanine-doped TGS crystal, measured at 125 Hz with $A = 0.25$ mm². The lowest *NEP* values are found at about 50 °C.

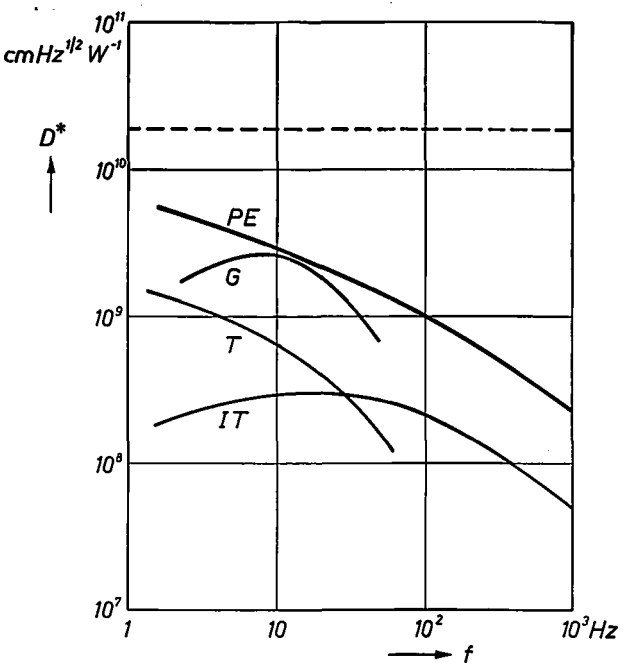


Fig. 16. The detectivity D^* (as a function of frequency) for various types of infrared detector: the pyroelectric element (LATGS crystal) of fig. 15 (*PE*), a Golay cell (*G*), a thermopile (*T*) and an immersed thermistor (*IT*). The dashed line indicates the theoretical upper limit for a thermal detector.

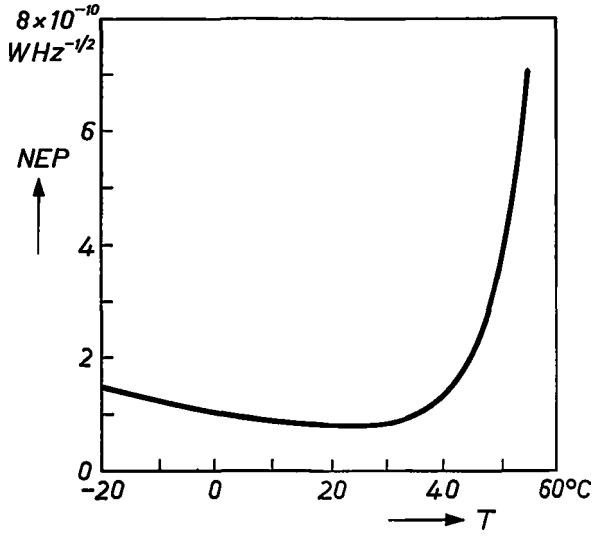


Fig. 18. The temperature dependence of the noise equivalent power *NEP* for a pyroelectric infrared detector with an l-alanine-doped mixed triglycine sulphate/selenate crystal: 70% sulphate, 30% selenate. Electrode area 0.25 mm², measurements made at 12.5 Hz.

is about half of that for a similarly sized crystal of LATGS. The best performance is to be expected from triglycine selenate. Measured detector performances at room temperatures are shown in *fig. 19*. These represent an improvement of approximately 2.5 times on an LATGS crystal of the same size. (Figs. 15 and 18 can be compared directly since the crystals used were not not of the same size.)

The current state of the technology for the new pyroelectric detectors can be summarized as follows. Sensitivity values of about 2.5×10^9 cm Hz$^{\frac{1}{2}}$/W^{-1} at 10 Hz have been achieved, i.e. better than for all other types of thermal detector (see fig. 16). The upper limit of the operating range has been pushed up to about 70 °C. The detectors do not depole, even near the Curie temperature and recover their performance automatically after heating to above the Curie temperature. It seems extremely likely that even better detectors can be made from the latest materials. The new pyroelectric detectors are both sturdy and stable. They are not only more sensitive, but much faster than other types of thermal

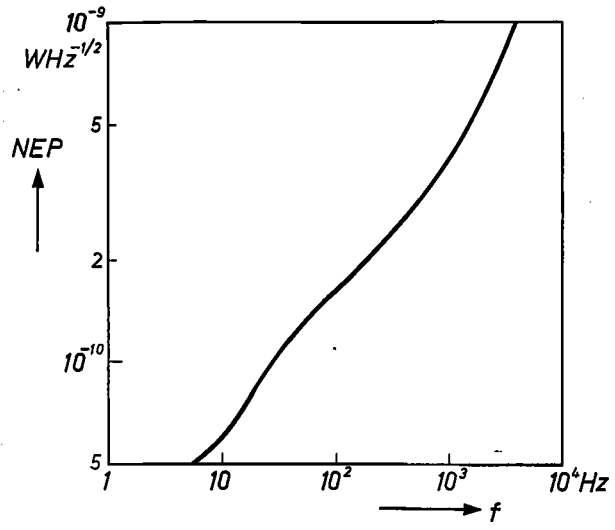


Fig. 19. The frequency dependence of the noise equivalent power *NEP* for a pyroelectric infrared detector with an l-alanine-doped triglycine-selenate crystal; $A = 0.25$ mm².

detectors, such as the Golay cell (based on a change in gas pressure with temperature), thermistors, thermocouples and thermopiles.

This work has been carried out with the support of the Procurement Executive of the British Ministry of Defence.

Summary. Ferroelectric materials have been developed that do not depole: crystals of l-alanine-doped triglycine sulphate (LATGS). Even in the absence of an external field these materials give displaced dielectric hysteresis loops below the Curie temperature, due to an internal field. The sulphate can also be completely or partly replaced by selenate. Detectors with minimum values of the noise equivalent power (NEP) at temperatures from about 20 °C up to about 50 °C have been made. Such detectors give better performance than all other types of thermal infrared detector. They are sturdy, stable, sensitive and fast. The l-alanine prevents depoling because the molecules do not have mirror-plane symmetry like glycine molecules, and cannot therefore change over to an orientation corresponding to the opposite polarization.

Recent scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands *E*
Mullard Research Laboratories, Redhill, Surrey, England *M*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium *B*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.) *N*

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, The Netherlands.

G. A. Acket, W. Nijman & H. 't Lam: Electron lifetime and diffusion constant in germanium-doped gallium arsenide.
J. appl. Phys. **45**, 3033-3040, 1974 (No. 7). *E*

C. Belouet & M. Monnier: Décoration des hétérogénéités de conductivité électrique dans KDP et DKDP. Corrélation avec les défauts cristallins.
Acta Electronica **18**, 143-153, 1975 (No. 2). *L*

H. Bex: Die Hochfrequenzfelder der vormagnetisierten Ferritkugel.
Frequenz **28**, 305-310, 1974 (No. 11). *A*

R. N. Bhargava & P. C. Mürau: Efficient red GaP LED's with compensated *p* layers.
J. appl. Phys. **45**, 3541-3546, 1974 (No. 8). *N*

P. Blood: Planar channeling in ion-implanted silicon.
Phys. Stat. sol. (a) **25**, K 151-154, 1974 (No. 2). *M*

K. Board, J. Basterfield, M. J. Josh & A. Gill: Alloyed heat-sinking of transferred electron devices.
Metal-semiconductor contacts, Proc. Conf. Manchester 1974 (Inst. Phys. Conf. Ser. No. 22), pp. 123-128. *M*

H. C. Brinkhoff: Substitution of aluminium by gallium in β-alumina.
J. Phys. Chem. Solids **35**, 1225-1229, 1974 (No. 9). *E*

A. Broese van Groenou, J. H. N. Creyghton & J. G. M. de Lau: Trivalent cobalt in nickel-zinc ferrites, II. Magnetic relaxations.
J. Phys. Chem. Solids **35**, 1081-1088, 1974 (No. 9). *E*

K. H. J. Buschow: Calcium-nickel intermetallic compounds.
J. less-common Met. **38**, 95-98, 1974 (No. 1). *E*

K. H. J. Buschow, A. M. van Diepen & H. W. de Wijn (State University of Utrecht): Crystal-field anisotropy of Sm^{3+} in $SmCo_5$.
Solid State Comm. **15**, 903-906, 1974 (No. 5). *E*

K. H. J. Buschow, A. R. Miedema & M. Brouha: Thermal expansion anomalies of rare earth – cobalt compounds.
J. less-common Met. **38**, 9-13, 1974 (No. 1). *E*

T. Claasen & L. Kristiansson (Royal Institute of Technology, Stockholm): Improvement of overflow behaviour of 2nd-order digital filters by means of error feedback.
Electronics Letters **10**, 240-241, 1974 (No. 12). *E*

T. Claasen, W. F. G. Mecklenbräuker & J. B. H. Peek: A comparison between the stability of second-order digital filters with various arithmetics.
1974 Eur. Conf. on Circuit theory and design, London (IEE Conf. Publn. No. 116), pp. 354-358. *E*

P. Delsarte: The association schemes of coding theory.
Math. Centre Tracts No. 55, 139-157, 1974. *B*

P. A. Devijver: On a generalization of the Bayesian distance.
2nd Int. Joint Conf. on Pattern recognition, Copenhagen 1974, pp. 52-55. *B*

J. Donjon, R. Le Pape, G. Marie & B. Monod: Le tube Titus. Application à la projection d'images sur grand écran.
Acta Electronica **18**, 103-116, 1975 (No. 2). *L*

O. P. van Driel, H. J. Prins & G. W. Veltkamp (Eindhoven University of Technology): Estimating the parameters of the factor analysis model without the usual constraints of positive definiteness.
Proc. Symp. on Computational statistics, Vienna 1974, pp. 255-265. *E*

C. Ducot: New optical relays and their potential applications.
Acta Electronica **18**, 88 *(in English)*, 89 *(in French)*, 1975 (No. 2). *L*

F. Dumont & D. Rossier: Phototitus: un relais optique convertisseur d'images.
Acta Electronica **18**, 131-141, 1975 (No. 2). *L*

J. B. A. A. Elemans (Kamerlingh Onnes Laboratorium, Leiden) **& K. H. J. Buschow:** Crystal and magnetic structure of intermetallic compounds of the type $Th(Co_xFe_{1-x})_5$.
Phys. Stat. sol. (a) **24**, 393-399, 1974 (No. 2). *E*

F. J. M. Frankort: Vibration and sound radiation of loudspeaker cones.
Thesis, Delft 1975. (Philips Res. Repts. Suppl. 1975, No. 2.) *E*

W. Funk & W. Schilz: Thick film techniques for hybrid integrated microwave circuits.
Radio & electronic Engr. **44**, 504-508, 1974 (No. 9). *H*

C. J. Gerritsma, C. Z. van Doorn & P. van Zanten: Transient effects in the electrically controlled light transmission of a twisted nematic layer.
Physics Letters **48A**, 263-264, 1974 (No. 4). *E*

J. Graf: Multiplicateurs d'électrons à microcanaux. Des composants miniatures pour des images plus intenses.
Phase Zéro No. 12, pp. 20-23, 1974. *L*

B. K. Herbert: Accurate current integrator for ion implanting.
SERT J. **8**, 209-210, 1974 (No. 9). *M*

C. Hily, J. J. Hunzinger & J. Michel: Problèmes de l'optimisation de l'optique de projection associée à Titus.
Acta Electronica **18**, 117-130, 1975 (No. 2). *L*

W. H. de Jeu & Th. W. Lathouwers: Nematic phenyl benzoates in electric fields: I. Static and dynamic properties of the dielectric permittivity, II. Instabilities around the frequency of dielectric isotropy.
Mol. Cryst. liq. Cryst. **26**, 225-234, 235-243, 1974 (No. 3/4). *E*

H. Kalis, M. Klinck, G. Landvogt, J. Lemmrich & G. Schröder: Frequenzanaloges Prozeßführungssystem.
Elektronik **23**, 361-364, 1974 (No. 10). *H*

J. P. Klozenberg: Numerical computation of scattering phase shifts for a screened Coulomb potential.
J. Physics A **7**, 1840-1846, 1974 (No. 15). *M*

W. L. Konijnendijk: The structure of borosilicate glasses.
Thesis, Eindhoven 1975. (Philips Res. Repts. Suppl. 1975, No. 1.) *E*

M. Laguës: Band bending effect on the thermal dependence of photoemission from Li activated Si.
Surface Sci. **45**, 432-440, 1974 (No. 2). *L*

P. K. Larsen & J. M. Robertson: Electrical and optical properties of thin films of Pb^{2+}- and Si^{4+}-doped YIG produced by liquid phase epitaxy.
J. appl. Phys. **45**, 2867-2873, 1974 (No. 7). *E*

J. G. M. de Lau, R. M. van den Heuvel & A. Broese van Groenou: Trivalent cobalt in nickel-zinc ferrites, I. Chemical problems.
J. Phys. Chem. Solids **35**, 1073-1080, 1974 (No. 9). *E*

F. H. de Leeuw: Domain wall motion in Ga:YIG films.
J. appl. Phys. **45**, 3106-3108, 1974 (No. 7). *E*

M. H. van Maaren & E. E. Havinga: Variation of the electron-phonon interaction strength in superconducting $LaIn_{3-x}Sn_x$.
Low temperature physics — LT 13, editors K. D. Timmerhaus, W. J. O'Sullivan & E. F. Hammel, Plenum Press, New York 1974, Vol. 3, pp. 392-394. *E*

G. Marie: Les relais optiques d'images utilisant l'effet Pockels. Rappel historique.
Acta Electronica **18**, 91-102, 1975 (No. 2). *L*

G. M. Martin, P. Bach & P. Guétin: Transport properties in HgI_2.
Appl. Phys. Letters **25**, 286-288, 1974 (No. 5). *L*

F. Möllers, H. J. Tolle & R. Memming: On the origin of the photocatalytic deposition of noble metals on TiO_2.
J. Electrochem. Soc. **121**, 1160-1167, 1974 (No. 9). *H*

B. J. Mulder: The opening up of insoluble oxides (tantalum, niobium, chromium, aluminium and others) with liquid selenium dioxide.
Anal. chim. Acta **72**, 220-222, 1974 (No. 1). *E*

G. F. Neumark: Decay kinetics of the red luminescence of GaP(Zn, O) for quasi-thermal-equilibrium conditions.
Phys. Rev. B **10**, 1574-1584, 1974 (No. 4). *N*

P. Piret: Good block codes derived from cyclic codes.
Electronics Letters **10**, 391-392, 1974 (No. 18). *B*

R. G. Pratt, G. Simpson & C. W. O'Connor: Acoustic-surface-wave properties of $Bi_{12}SiO_{20}$ and related compounds.
Electronics Letters **10**, 369-371, 1974 (No. 18). *M*

H. Rau: Defect equilibria in silver sulphide.
J. Phys. Chem. Solids **35**, 1553-1559, 1974 (No. 11). *A*

J. A. J. Roufs (Institute for Perception Research, Eindhoven): Dynamic properties of vision: IV. Thresholds of decremental flashes, incremental flashes and doublets in relation to flicker fusion, V. Perception lag and reaction time in relation to flicker and flash thresholds, VI. Stochastic threshold fluctuations and their effect on flash-to-flicker sensitivity ratio.
Vision Res. **14**, 831-851, 853-869, 871-888, 1974 (No. 9).

B. Schiek: Hybrid branchline couplers — a useful new class of directional couplers.
IEEE Trans. **MTT-22**, 864-869, 1974 (No. 10). *H*

J. M. S. Schofield: The strike and extinction voltages of gas discharge display cells.
3rd Int. Conf. on Gas discharges, London 1974 (IEE Conf. Publn. No. 118), pp. 329-333. *M*

W. Schreiber & R. Orlowski: Die Regelparameter von Kryostaten mit Kontaktgaskoppelung im Temperaturbereich 4,2 . . . 300 K.
Klima- + Kälte-Ing. **2**, 375-378, 1974 (No. 9). *H*

M. J. Sparnaay: Ellipsometry.
Vakuum-Technik **23**, 141-151, 1974 (No. 5). *E*

H. van Tongeren: Positive column of cesium- and sodium-noble-gas discharges.
Thesis, Eindhoven 1975. (Philips Res. Repts. Suppl. 1975, No. 3.) *E*

M. J. Underhill & P. A. Lewis: Automatic tuning of antennae.
SERT J. **8**, 183-184, 1974 (No. 8). *M*

A. Venema: Ultra high vacuum technology in relation with surface physics studies.
Electrónica y Fis. apl. **17**, 9-15, 1974 (No. 1/2). *E*

J. F. Verwey & B. J. de Maagt: Avalanche-injected electron currents in SiO_2 at high injection densities.
Solid-State Electronics **17**, 963-971, 1974 (No. 9). *E*

L. Vriens: Raman scattering cross sections for In and Tl atoms and multiphoton processes in Sr.
Optics Comm. **11**, 396-401, 1974 (No. 4). *E*

L. Vriens & M. Adriaansz: Depolarization ratios for anti-Stokes Raman scattering from In and Tl.
Optics Comm. **11**, 402-404, 1974 (No. 4). *E*

H. W. Werner & A. J. Linssen: A study on the maximum tolerable pressure in a residual gas analyser.
J. Vac. Sci. Technol. **11**, 843-847, 1974 (No. 4). *E*

H. W. de Wijn (State University of Utrecht), **A. M. van Diepen & K. H. J. Buschow:** Influence of crystal fields on the magnetic form factor of samarium.
Solid State Comm. **15**, 583-585, 1974 (No. 3). *E*

Contents of Philips Telecommunication Review **33**, No. 2, 1975:

R. Bodart: TDM telegraph and data transmission system 3TR 1500 (pp. 53-67).
J. M. Fripiat: Synchronous modems for data transmission on a group band circuit (pp. 68-85).
H. Bouwman & R. Papeschi: A new VF telegraph and data carrier system (pp. 86-96).
J. S. Visser: Using Bode's return loss integral in filter design (pp. 97-101).

Contents of Electronic Applications Bulletin **33**, No. 1, 1975:

A. C. Smaal: TV receivers as domestic video accessories (pp. 1-11).
G. H. Haenen: Evaluation of mains-borne harmonics due to phase-controlled switching (pp. 13-28).
Switched-mode power supplies: symposium report (pp. 29-30).
W. M. Dorn, W. Th. Hetterscheid, C. W. J. v. d. Pol & W. Smulders: SOAR — the basis for reliable power circuit design, Part II: Using published SOAR data (pp. 31-42).
Nomogram for determining heatsink sizes for power transistors (pp. 42-43).

Contents of Mullard Technical Communications **13**, No. 127, 1975:

L. M. White: Television switched-mode power supply using the TDA2640 (pp. 258-279).
J. A. Tijou: High-quality 25W audio amplifier (pp. 280-288).
B. G. Starr: An opto-isolated oscilloscope trigger circuit (pp. 289-290).
GZF1202 encoder and decoder for data error detection (pp. 291-292).

Contents of Valvo Berichte **19**, No. 4, 1974:

F. Spitzer: Ursprung und Ziel moderner Herstellverfahren der Halbleitertechnik (pp. 133-138).
N. Hoffmann: Entwicklungstendenzen bei Herstellverfahren für monolithisch integrierte Analogschaltungen (pp. 139-145).
H. Schaumburg: Ein Verfahren zur schnellen Bestimmung zweidimensionaler Dotierungsverteilungen auf Halbleiterscheiben (pp. 146-150).
Titan-Gold-Metallisierung für HF-Leistungstransistoren (pp. 151-158).
J. H. M. Uijlings: Rauschen in Kernstrahlungsspektrometern mit Halbleiterdetektor (pp. 159-170).

Volume 35, 1975, No. 9 pages 225-260 Published 2nd January 1976

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 10

Structural testing of digital circuits

W. G. J. Kreuwels

Without the aid of a computer it would be almost impossible to develop large logic networks like LSI circuits. Designing the circuit, testing the design, designing and manufacturing the masks, testing during manufacture, all these stages require so much computing work that only computer-aided design can provide the answer. At the same time it is necessary to have good software, otherwise the computing time may become impractically long, even for a fast computer. This article deals with the functional test, whose purpose is to check whether a new design of circuit does in fact perform the required logic function. If the test were to be carried out in full on a network with 64 inputs — not an unusual number for an LSI circuit — it would require 2^{64} experiments. Even with test equipment specially built for the purpose and checking at a rate of once per nanosecond, the test would take more than 300 years. Such a complete test, however, would cover very many more possibilities than those of any practical interest. The programs dealt with in this article, which are based on the structure of a network to be tested, can be used in setting up a much shorter test for checking the functional specifications. Depending on the type of network, this procedure requires only a few minutes to a few hours of computer time, while the actual test takes only a few seconds during manufacture.

Introduction

In the manufacture of logic circuits test procedures have to be carried out during both the design phase and manufacture. The most extensive of these is the 'functional' test, which determines whether a circuit meets the functional specifications, in other words whether the circuit performs the desired logic function correctly.

The functional test for small circuits with few gates (AND, OR, NAND and NOR elements) can be carried out quite simply by successively applying all possible combinations of the logic values '1' and '0' to the inputs. The logic values at the outputs must then correspond to the response of a fault-free circuit. In the design phase this procedure can be simulated on a computer. In a production test the various test patterns can be generated by a counter circuit and fed to the network under test and to a calibration circuit: the two responses are then compared (*fig. 1*).

Ir W. G. J. Kreuwels is with Philips Research Laboratories, Eindhoven.

Such a procedure, even using a very fast computer for generating the test patterns, is not feasible for large circuits that have several hundred gates and scores of inputs and outputs. In a network with n inputs the number of different combinations of input signals is

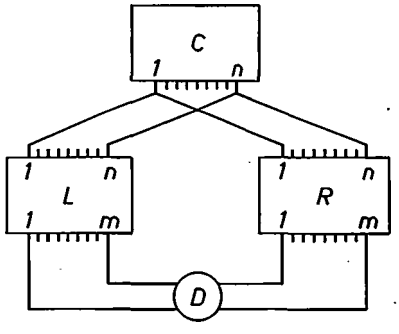


Fig. 1. Testing a simple logic circuit with n inputs and m outputs. A counter C generates the test patterns and simultaneously feeds them to the inputs of the circuit under test L and to a calibration circuit R. The output signals are fed to a comparator circuit D, which delivers a signal when the two responses are not identical.

2^n, and 2^n test patterns therefore have to be applied. Even if such patterns are generated at a rate of one per nanosecond, such a test takes far too long. For a network with 64 inputs it would take no less than 300 years. A further complication is that the response is not entirely determined in all the circuits by the instantaneous combination of input signals. This is only the case in *combinational* networks. Usually, however, large networks are *sequential*, and in such circuits the response is dependent not only on the instantaneous inputs but also on a number of input patterns preceding them and on the order in which they are presented. For circuits of this type the computer would have to generate different series of test patterns, which would make the test last even longer. For large networks, therefore, a very much shorter functional-test procedure is needed.

Even with a shorter test procedure it is only the input and output signals that are examined. Indeed, this is the only way of testing large-scale integrated (LSI) circuits, since the separate components are so small as to be inaccessible: but even with networks on printed-wiring panels it is much simpler to make the measurements only at the input and output connectors. We have therefore tried to find ways of arriving at a group of test patterns that is smaller, yet still sufficient for the functional testing of digital networks.

The existence of such a small group of test patterns may be demonstrated as follows. The starting point of the complete functional test is the truth table; this gives the functional specification of a combinational network, this is to say the relation between all the combinations of input and output signals (a sequential network is described by a series of these truth tables). A large truth table may not only correspond to a single network, it may also correspond to many other kinds of network of widely different structure. Since all these networks have their own characteristic types of fault the complete functional-test procedure is capable of detecting all possible faults in all these various networks. This is much more than is required: all we want to know is whether the particular design or network being tested meets the specifications. It is obvious that a much smaller group of test patterns will suffice for this. These patterns should not be derived from the truth table but from the structural design data. To describe the abbreviated functional test, which only uses this small group of test patterns, we have therefore intruduced the term *structural* test.

This article will give an account of the computer programs that generate the patterns for the structural test. The programs are meant to be used in the design phase of a network, and we shall see that certain design faults can be directly brought to light by such a program even at the stage when the test patterns are being designed. The principle we use for deriving the test patterns was introduced many years ago [1], but has not yet been used in practical programs.

Once the test patterns have been derived, the actual structural test during manufacture takes very little time. Of course the number of test patterns required will depend to a great extent on the type of network, but a circuit with a thousand gates will require a few hundred to a few thousand patterns. If these patterns are generated during the manufacturing test at a speed of one per microsecond, the test will take no more than a few milliseconds. It should be realized when applying the test patterns that the method aims simply at answering the question of whether the network functions correctly or not: the test patterns are therefore primarily meant for fault *detection* and not for fault *location* [2].

Before dealing with the procedure for deriving the test patterns, the scope of the structural test will be indicated by describing the faults that can occur in the various stages of designing and manufacturing a network.

The faults in a digital circuit

The design of a digital network can be subdivided into three phases, each with its own distinct types of fault.

1) In the first phase the time variation of signals in small circuits consisting of a few dozen transistors at most is calculated for various values of the physical parameters that determine the characteristics of the circuit. This circuit analysis gives a general circuit description from which the digital behaviour can be deduced by the introduction of the logic levels and time sampling. This yields descriptions of gate circuits with various specifications for the transistor parameters and for the tolerances ('margins') within which other parameters such as supply voltages, load, temperature, etc. must remain. If one or more of these specifications is not met, the result may be that a signal does not reach the required logic value but a value somewhere in between the two. Such an out-of-tolerance value is referred to as a '½' value (*figs. 2a* and *b*).

Faults of this type are detected during the manufacture of an integrated circuit by measuring the parameters of a special test transistor, which is formed on the same chip as the circuit and is regarded as being representative of the transistors in the circuit. 'Marginal tests' are also carried out to verify whether the circuit is unaffected by variations of the other parameters within the specified limits.

2) In the second phase the gate circuits are combined to form a network. If all the specifications mentioned

under (1) are now taken into account, only the time factor need be investigated. In a large network the delay times of the gate circuits are additive over various paths, and as a result the signals at various places in the circuit can get out of step. A signal may then reach the required level too late and thus be unknown at the sampling time: this case is also referred to as a '½' value (fig. 2*c*).

These dynamic errors in a design are found by simulating the dynamic behaviour of the network on a computer. The programs required for this are very much simpler than those for circuit analysis but they have to be applied to a much larger network. The amount of computation can be substantially reduced, however, if the designer indicates where difficulties in the design are likely to occur. The dynamic investigation of a design produces a number of test patterns, which can also be used for testing the dynamic behaviour of the circuits during production.

3) Even if all faults of the first type are eliminated, so that all signals reach a logic level at the instants of sampling, design or production faults can still cause a wrong response in the network. These faults are detected by the functional test or by its abbreviated form, the structural test. A fault analysis shows that at this stage we need only consider situations in which the signal has a *fixed* value ('0' or '1') at one particular point in the circuit and does not respond to the other signals in the desired manner. These are referred to as 'stuck at' faults: either 'stuck at zero', abbreviated to s-a-0, or 'stuck at one', abbreviated to s-a-1. Production faults that can cause s.a. faults are broken connections and short circuits. Possible design faults are of course difficult to indicate, but by analogy with open- and short-circuits it is possible to think of a necessary connection that might be missing or an unnecessary one whose inclusion could introduce a fault.

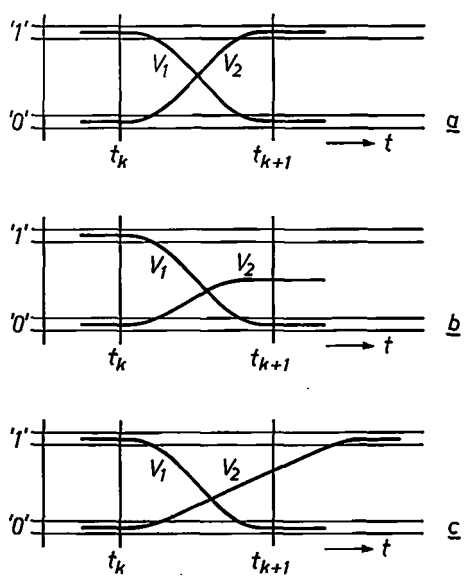


Fig. 2. Illustrating the signal faults in a digital circuit. *a*) Behaviour of two signals V_1 and V_2. Two regions are defined in which the logic values '0' and '1' are assigned to the signals. Only the values at the times t_k; t_{k+1}, etc. are important; in a *synchronous* circuit these times are determined by an external clock signal, in an *asynchronous* circuit by a signal from the circuit itself. *b*) Signal V_2 does not reach level '1', so that at the time t_{k+1} the value of V_2 is indeterminate. It is then said to have a value '½'. *c*) The signal V_2 reaches level '1' but it does so too late. When this dynamic fault occurs, the signal is also said to have a '½' value.

The fault analysis is performed in two phases. First the possible physical faults in the circuit are converted into s.a. faults (other types of fault have already been detected in the other tests). Some physical faults correspond to one s.a. fault, others such as short-circuits may result in a group of s.a. faults at various points in the circuit. In the second part of the analysis it is found that all these physical faults can be detected by a test program for single s.a. faults. It has been found that this program can also detect all groups of s.a. faults.

It will be clear from what has been said above that the design of a large digital network is only possible with the aid of a computer: it is an example of computer-aided design. The designer generally has a large number of existing programs available (sometimes called 'tools') for the computing work. Examples at Philips are the PHILPAC programs for circuit analysis, PHILSIM for simulating the dynamic behaviour of networks, and the TESTCC and TESTSC programs, described in this article, for deriving the test patterns for the structural test.

Test patterns for a combinational network

The procedure used in deriving the test patterns for the simple case of a combinational network will first be described. The computer program developed for this has been given the name TESTCC.

Structure input

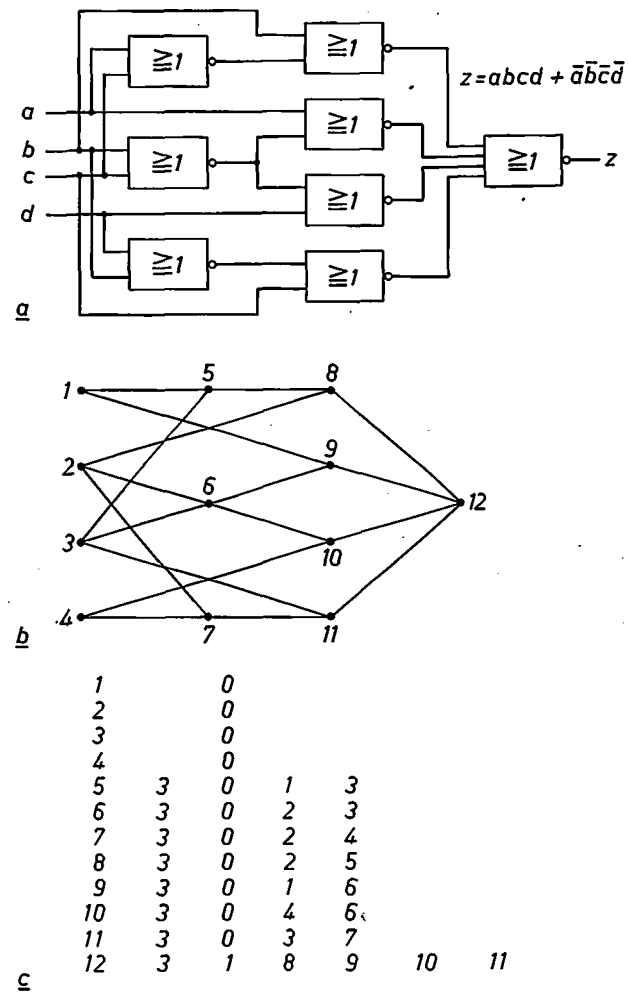
The computer that works out the test patterns is supplied with the structural data for the network in the form of a table; *fig. 3* shows how this is arrived at for a combinational network. The network (*a*) is presented in the form of a 'directed graph' (*b*), in which all the connections are shown but the components are reduced to nodal points. Inputs and nodes are numbered in such a way that the information in the circuit is always

[1] R. D. Eldred, Test routines based on symbolic logical statements, J. Ass. Computing Mach. 6, 33-36, 1959.
D. B. Armstrong, On finding a nearly minimal set of fault detection tests for combinational logic nets, IEEE Trans. EC-15, 66-73, 1966.

[2] A preliminary investigation with a simple network has revealed that the structural-test program gave almost the same amount of information about the location of the faults in this network as the complete functional test. We therefore intend to design fault-location programs on the basis of the structural test.

passed on from a lower to a higher number. Each line in the table (*c*) corresponds to an input or a node point: the first column gives the number of this point, the second column gives a code for the type of component represented by the node (1, 2, 3, 4 for OR, AND, NOR and NAND gates). The third column contains a '1' if the node is 'observable' in other words if it represents an output gate of the network (in fig. 3 only the last NOR gate), otherwise the column contains a '0'. The other columns give the numbers of the other nodes that are connected with the inputs of the component. In this way the whole network is described in a form that can easily be fed to the computer, e.g. on punched cards. The truth tables for the various types of component still have to be added to this table, of course.

In addition to the structural data for the network, the computer also receives information about the faults to be detected by the test patterns. For the structural test these are only the 'stuck at' faults, which can occur on any line connecting two nodes. If the network branches at a particular node, the fault need not always affect all the lines branching out from that point; it is possible that the fault will only appear on one of the lines beyond the node, and hence only at the input of one of the next nodal points. It is then called a 'gate-input fault'.



1		0				
2		0				
3		0				
4		0				
5	3	0	1	3		
6	3	0	2	3		
7	3	0	2	4		
8	3	0	2	5		
9	3	0	1	6		
10	3	0	4	6		
11	3	0	3	7		
12	3	1	8	9	10	11

Fig. 3. Coding of the structure of the network for feeding the structural information to a computer. *a*) Example of a combinational network. *b*) Simplified diagram of the network (directed graph) in which the inputs and components are represented as nodes. *c*) Table containing all the data relating to the structure of the network. The first column contains the numbers assigned to the nodes, the second column gives a code number for the component represented by the node (a NOR gate in all cases here), the third column contains a 1 where an output gate is referred to, otherwise a 0, and the other columns give the numbers of the other nodes that are connected to the inputs of the component.

The critical path

The procedure for deriving the test patterns is based on finding *critical paths*. A critical path is a path through the network that makes the effect of an s.a. fault on a particular line in the network visible at an output. To each critical path there corresponds a pattern of input signals, which is therefore the test pattern for the particular fault. The term 'critical path' is used to indicate that all the signals on that path are 'critical': any change in the value of one of the signals propagates to the output. It therefore follows that the corresponding test pattern shows up not only the original s.a. fault but also all the other faults that can occur along that path and hence bring out the same change of value at the output. It does not matter which output gives the error signal because the test checks the total response of the network.

For some faults several critical paths will be possible, and therefore several test patterns. However, since we wish to derive the smallest possible group of test patterns, the search process is stopped as soon as a critical path is found. Of course, a check is made for each test pattern discovered, to see if any other faults can be detected with it at the same time. These need not necessarily be faults occurring along the critical path: in networks with more than one output faults can appear at the same time at other outputs.

The signal under test will be indicated by the symbol k or $\bar{k}$ so that we can follow it through the circuit. If we are looking for a test pattern that detects an s-a-0 fault we call the signal k, if we want to detect an s-a-1 fault we call the signal $\bar{k}$. We can now use the rule that the fault is *not* present if k is 1 but *is* present if k is '0'. *Fig. 4* shows part of a critical path in a network. It can be seen from this figure that the other input signals of the components forming the critical path have to meet a number of conditions, referred to as 'imperative implications'. Of course, these conditions may lead to contradictions in other parts of the network, in which case the path cannot be used and another must be looked for. It will be evident that isolating a path in which no contradictions occur, and doing this for all

s.a. faults that can occur in a network, is a very considerable task. We cannot therefore discuss the complete TESTCC program here but will have to confine ourselves to the algorithm used by the computer.

Non-testable faults

When a critical path for a particular fault has been found, this shows in itself that the fault is capable of being tested, and at the same time the test pattern to be used has been found. The question that now arises, of course, is whether a critical path can in fact be found for every s.a. fault. We know that a critical path exists for all faults that are detected by the *complete* functional test, since this test also makes the faults observable at an output. Thus when all the possible critical paths have been explored systematically in the derivation of test patterns for the structural test, test patterns have to be found for all these faults. There are also faults, however, that are *not* discovered by the complete test. There is no critical path for a fault of this kind and therefore no test pattern can be found for it. This relates mainly to faults in parts of a network that are redundant in the functional logic. In some cases such parts of the circuit are added for technical reasons, but the discovery of a non-testable fault in the design phase of a network more usually points to a design fault. The ability to reveal such faults at an early stage is therefore a valuable feature of the structural procedure.

Generation of test patterns

Let us suppose we wish to find a test pattern that will show whether there is an s-a-0 fault on a connection line j in a network. The algorithm for looking for test patterns consists of a number of procedures: the first one, called FOR (for forward) is started by assigning the value '2' to all lines in the network. Expressed in words, this means that the values of the signals are unknown. For the component that has the line j as an output procedure FOR makes use of the truth tables to look for a combination of input signals for which this line has the value '1'. If there are various ways of doing this, FOR remembers them. In accordance with the first possibility investigated, the values '2' at the inputs of the component are now replaced by a '0' or a '1'. The signal on the line j acquires the value k. FOR then starts to look for a critical path along which k appears at the output. The various possible paths are investigated one by one, and in each case the values at the inputs of the other components are changed from '2' to '1' or '0', in the manner shown in fig. 4. If contradictions are found, FOR tries the next possible path. If the filling in of these 'imperative implications' leads to contradictions in all paths, FOR carries out the same investigation for the other combinations of input signals of the component that has j as its output.

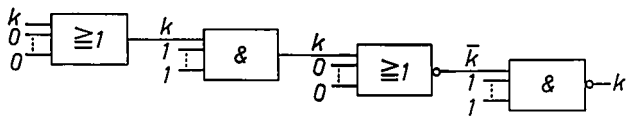


Fig. 4. Example of a chain of components forming part of a *critical path*. At one of the inputs of the component on the far left there is a signal k; the level '0' or '1' of k indicates whether or not a 'stuck at' fault is present somewhere on the left of this component in the network from which this chain has been taken. (A 'stuck at one' or 'stuck at zero' fault means that the signal at a particular point remains fixed at the value '1' or '0', irrespective of the input signals applied to the network.) For the signal k to be transferred to the right, the other inputs of the OR and the NOR gate must all be '0'and those of the AND and NAND gate must all be '1'.

In the search for a critical path there are two alternatives: either FOR encounters contradictions in all possible paths, or FOR finds a critical path along which an output can acquire the value k (of course a path can also be used if the output acquires the value $\bar{k}$). In the first case there is no test for the fault investigated; FOR then receives the instruction to look for a critical path for another fault. If FOR does find a critical path it remains available, as there may be more possibilities for the line j and there is not as yet any certainty about the usefulness of the path that has been found. Contradictions may yet arise during the introduction of the 'non-imperative' implications. Before there is any question of a test pattern it is, of course, also necessary to substitute '0' or '1' for the value '2' on all lines in the network that are *not* directly connected with the critical path, and these values must also not be contradictory, either with one another or with the values already assigned. These other signals are filled in by a procedure called BACK, which begins at the output side of the network and works through it in the opposite direction to FOR. In general BACK will also encounter various possible choices, which must be dealt with in succession.

When the BACK procedure encounters no contradictions, a test pattern has been found. If inputs of the network remain with the value '2', this means that the signals here may be chosen arbitrarily. If BACK continuously comes up against contradictions the path found by FOR does not after all lead to a usable test. FOR must then look for the next possibility, which in turn is investigated by BACK. In this way the interplay between FOR and BACK leads either to a test pattern or to the conclusion that there is no test for this particular fault. The combination of FOR and BACK is called LOOK.

When a test pattern has been found for an s.a. fault on a particular line in the network, this same test pattern will also reveal faults on a number of other lines, as we saw earlier. To prevent the whole LOOK procedure from being repeated for these faults as well, the

algorithm contains a procedure MORE. This procedure examines each test pattern to find out which other faults it will also discover. A procedure BOOK keeps a tally of the faults for which a test pattern has already been found in this way, and on the basis of this information a procedure CAND determines the next 'candidate', in other words the particular fault for which a test pattern must now be sought. The interplay of BOOK and CAND in fact 'directs' the generation of the test patterns and is therefore called DIRECT. The DIRECT, LOOK and MORE procedures together constitute a 'fault-detection strategy' STRAT, which can solve the whole problem.

Fig. 5 illustrates schematically the interaction of the various procedures that together constitute STRAT. Of course, depending on the actual execution of the procedures FOR, BACK, etc. there are various possible STRAT procedures. Fig. 5 shows only one example, the procedure followed by the TESTCC program, which derives the test patterns for a combinational network.

Test patterns for a sequential network

As already mentioned above, the response of a sequential network to a particular input pattern depends not only on this one pattern but also on a number of preceding patterns and on the sequence in which they are presented. This means that *series* of test patterns are necessary for the structural testing of a sequential network. The next section will give an account of the TESTSC program used for deriving these 'test sequences'. This program is a collection of a number of strategies that work in much the same way as STRAT. First, however, we shall look at the way in which the structure of such a circuit can be fed to the computer.

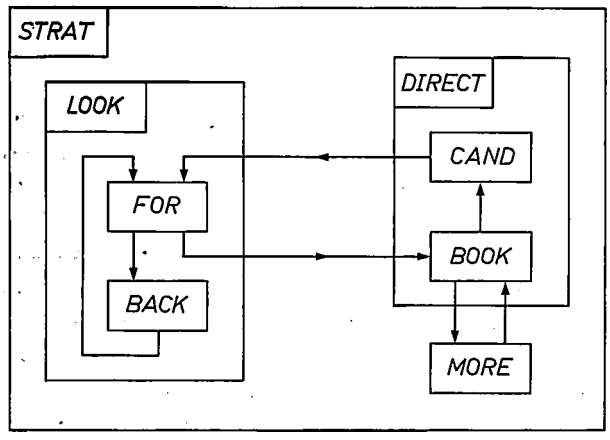


Fig. 5. Diagram of the procedures that together form the strategy STRAT for generating the test patterns for a combinational network. The procedures FOR (forward) and BACK, together forming LOOK, generate the test patterns; MORE investigates the total fault-detection capacity of the test patterns found; BOOK and CAND, together forming DIRECT, determine which new faults will require a search for a test pattern.

Structure of a sequential network

The characteristic feature of the structure of a sequential network is the occurrence of feedback loops in the information flow through the network. In describing this structure by the method illustrated in fig. 3 it is not therefore possible to number the nodal points in the directed graph in such a way that the information flows only from a low to a high number. At certain places information will inevitably flow back from a high to a low number. Since the FOR and BACK procedures that look for a critical path are unable to handle such feedback loops, we use a special approach in which a sequential network is represented as a series of combinational networks.

In a sequential network all the internal feedback loops can be considered as being broken and taken to the outside (see *fig. 6a*). This produces a number of new inputs and outputs, denoted by Y, which differ from the existing inputs and outputs, denoted respectively by X and Z, because they are not accessible in the actual network. The signals on these lines cannot therefore be directly affected or measured. The feedback lines in this model are produced by connecting the Y outputs located outside the circuit to the corresponding Y inputs. The resultant network — the inside block in fig. 6*a* — contains no further feedback lines and is therefore combinational. It can be described by a table of the type shown in fig. 3*c*: it is, however, necessary to indicate beside each input or output whether it relates to an X, Y or Z line.

Our next step is to replace the feedback lines in the model shown in fig. 6*a* by 'feed-forward' lines to copies of the same network. This yields the model in fig. 6*b*, which shows the sequential circuit as a series of identical combinational circuits, interconnected by the Y lines. This model gives a good picture of the actual situation in a sequential network: at the time t_0 to t_n the input patterns X_0 to X_n are successively presented to the network, giving the responses Z_0 to Z_n. The output pattern Z_n depends on the input signal X_n and on the values on the Y lines at the time t_n. If we indicate this functional relation by f_Z, then $Z_n = f_Z(X_n, Y_n)$. Similarly, Y_n is a function of the preceding input pattern and of the Y values at that moment of time: $Y_n = f_Y(X_{n-1}, Y_{n-1})$. Repeated substitution of this expression in the expression for Z_n yields the relation illustrated in fig. 6*b* (the figure must then be read from right to left). The signals X_0 to X_n actually presented in the network at the times t_0 to t_n are fed in this model to the copies 0 to n at the same times.

Generation of the test sequences

The principle of the derivation of a test sequence is illustrated in fig. 6. A search is made with the LOOK procedure for a critical path for a particular 'stuck at' fault in the combinational network from the model in fig. 6*a*. We thus look for a test pattern that makes the fault visible at an output (we shall return below to the initial values of Y_i assigned to the Y inputs for this purpose). If the fault signal appears via this path at a Z output, this test pattern alone is all that is necessary for detecting the fault. If, however, k appears at a Y output, the fault can then no longer be observed, since the Y lines are not observable. We again look for a critical path in the model given in fig. 6*a*, but now starting from the values Y_{i+1} on the Y lines obtained with the first test pattern. In fig. 6*b* this procedure means that the first pattern is looked for in copy i and the second pattern in copy $i + 1$. In this copy as well k may again appear at a Z or at a Y output; if it appears at a Y output we repeat the procedure in the next succeeding copy. If different test patterns are found in a copy with k appearing at a Y output, we must then make a further examination of all these different Y patterns in the next copy. As soon as the fault appears at a Z output in one of the copies, a critical path has been found and the search is stopped. If we call this copy j, then the critical path runs through copies i to j (the dashed line in fig. 6*b*): the test patterns X_i to X_j calculated in these copies then form a test sequence for this particular s.a. fault.

The initial values Y_i that we fed in when calculating the path in copy i are the result of 'initialization'. This is done by means of a series of patterns X_0 to X_{i-1}, which starts from an arbitrary state of the network in which the values on the Y lines are completely unknown, and applies a known pattern Y_i to the Y lines. All the subsequent Y and Z responses on input patterns X can be calculated from these initial values. The initialization sequence X_0 to X_{i-1} should be specified by the network designer and must precede every test sequence. The complete test for an s.a. fault thus consists of the series of patterns $X_0, \ldots X_{i-1}, X_i, \ldots X_j$.

When calculating a critical path we must bear in mind that the fault to be detected is of course present *in each copy*, i.e. including the copies 0 to $i - 1$ that are traversed during the initialization. Apart from the result of the initialization for the fault-free circuit, we must therefore also calculate the pattern Y_i for every possible s.a. fault. This means that in deriving a critical path for a particular fault we must always proceed from the initialization result Y_i computed for this fault.

During the calculation of the critical path the transition to a subsequent copy is always commanded by the clock signal that generates the times t_0 to t_n. The model

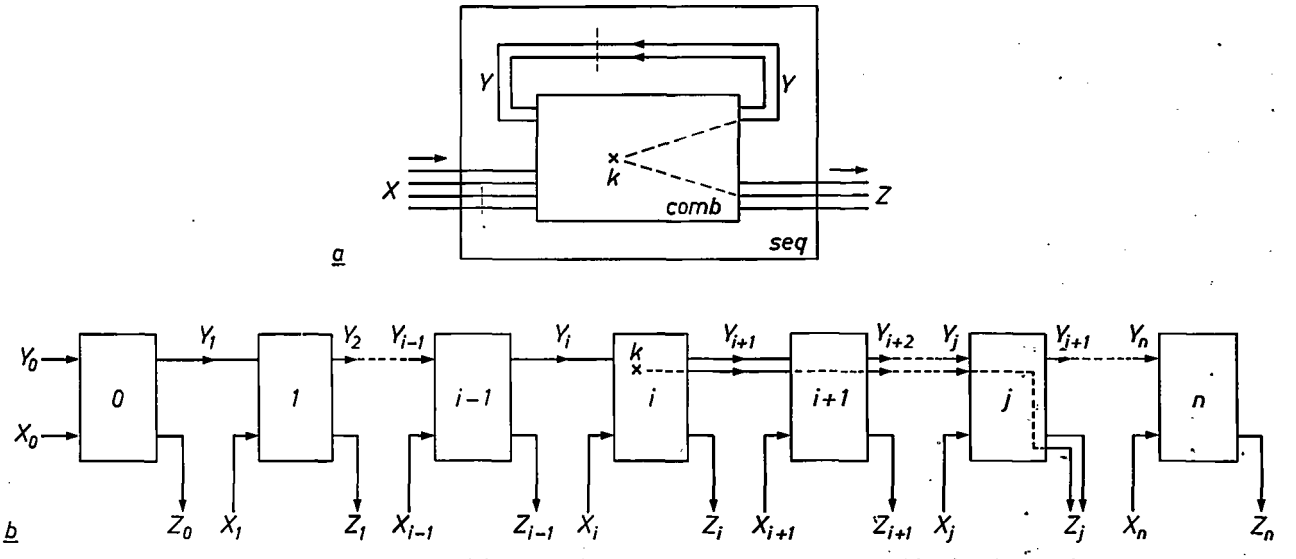


Fig. 6. *a*) Model of a sequential network in which it is represented as a combinational network with external feedback lines Y. The response Z in this model depends not only on the instantaneous input pattern X but also on the values on the Y lines. These in their turn are determined by what has been presented earlier on the X lines. For an s.a. fault k a search for a critical path can be made in the combinational model. If k appears at a Z output, one test pattern is sufficient for detection; if k appears at a Y output, a series of patterns must be derived. *b*) Extension of (*a*) in which the feedback lines are replaced by connecting lines to copies of the same network. (The various lines that together transport an X, Y or Z pattern are all represented here as a single line.) A critical path in copy i that ends at a Y output is continued in copy $i + 1$, starting from the calculated values Y_{i+1}. By continuing until the fault appears at a Z output we find a test sequence X_i to X_j. The critical path is schematically represented as a dashed line.

in fig. 6*b* assigns this signal to the input pattern X; the X lines corresponding to it are called clock lines and the others are called functional X lines. If we vary only the values on the functional lines and keep the clock signal constant, the response Z_i changes but the subscript i does not. This is the situation in which we look for a critical path in copy i. Changing the clock signal causes i to go to $i + 1$, so that the search can be continued in the following copy.

The TESTSC program

The TESTSC program begins with a START procedure for feeding the data to the computer (*fig. 7*). Information on the structure of the network is read in by the procedure READ and processed by the procedure BOOK. The latter procedure derives from the structure the list of candidates, i.e. the list of possible 'stuck at' faults for which test sequences have to be sought. The initial information X_0 to X_{i-1} is presented by the INIT procedure. After this has been entered, a MORE procedure calculates for each of the fault candidates the initial Y values Y_i from which a critical path is to be sought. This also reveals the faults that have already been detected by the initialization sequence: it is of course possible that certain faults will appear at the Z output when this sequence is applied. The results of MORE are similarly stored by BOOK.

A number of faults that the designer knows to be undetectable are 'masked' in a sub-procedure of START called MASK; no sequence will be sought for these faults. These might typically be faults in parts of the network that are technically necessary but logically redundant.

The strategy STRAT

The principal fault-detection strategy STRAT (*fig. 8*) generates the test sequences in the manner described above. Since we always look for the critical path in a combinational network, this strategy is virtually identical with that in fig. 5. The only difference is a procedure CLOCK, interpolated between LOOK and DIRECT. As long as this procedure is not active, a search is made for a critical path in one particular copy in the manner described in fig. 5. The CLOCK procedure always ensures the transition to the next copy. For each fault selected by the CAND procedure the search is started in copy i, always proceeding from the Y_i values relating to the fault. Under the control of CLOCK a series of copies i to j is then run through for each fault and a test pattern is calculated in each copy. The sequences for the various faults will generally be of different length. The MORE procedure investigates to find the other faults detected by each sequence.

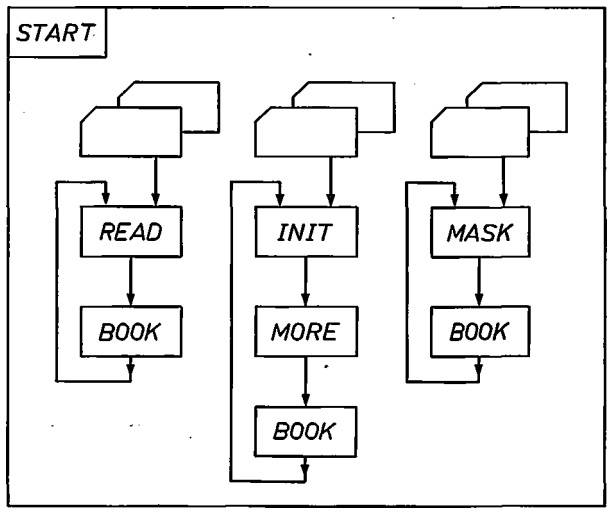


Fig. 7. START procedure for entering the data into the computer. The information sources are symbolically represented by punched-card packages. The structural information of the network is read in by the READ procedure and stored in BOOK. The initialization sequence is read in by INIT; the procedure MORE then calculates the starting pattern Y_i (see fig. 6) for the various possible faults and at the same time investigates to find out which faults have already been detected during the initialization. The MASK procedure receives data relating to faults which the designer knows to be undetectable. All this information is collected in BOOK.

BOOK keeps a continuous tally of particular faults found by the different sequences.

The fact that STRAT fails to find a test sequence for a particular fault does not permit the conclusion that tests cannot be made for the fault. The strategy STRAT does not investigate *all* detection possibilities but confines itself to the most promising patterns — those that contain the fault signal k. If there are too many of these in a particular copy, even these may not all be used. To restrict the computer time required a limit is

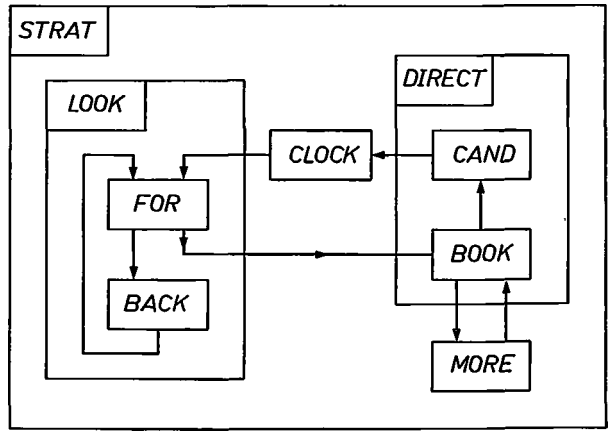


Fig. 8. A fault-detection strategy STRAT for sequential networks. For the various faults indicated by the procedure START test sequences are derived by generating test patterns in a series of copies i to j (see fig. 6*b*), for each fault starting from the input pattern Y_i for the given fault. The test pattern in a particular copy is generated in the same way as illustrated in fig. 5; the CLOCK procedure is responsible for the transition to the next copy. Here again a MORE procedure investigates to find out the total fault-detection capacity for the test sequence found.

also set to the length of a test sequence, so that the search procedure is stopped after a certain number of copies. In these latter points the program is very flexible: the user may indicate how far he wishes the search to be continued.

The strategies STRATZ and STRATYZ

In a well designed network the strategy STRAT will generally be sufficient for computing the whole test program. If faults remain for which STRAT finds no sequence, it is very important for the designer to know whether these faults are in fact non-testable, or whether a test may yet be found for other Y values. To investigate this we added two additional strategies to the program TESTSC. Our reasoning was as below.

If a fault is testable a 'last' copy j exists in which the fault appears at a Z output. There is then also a pattern Y_j which, together with a pattern X_j, gives rise to a critical path to a Z output. It must be possible to find this path by means of the LOOK procedure provided we allow *all* the possible values at the Y inputs. Having found one or more Y_j patterns in this way, we proceed to derive the test sequence. We do this by working through the model in fig. 6*b* from right to left until we find a Y pattern that corresponds to the initialization result Y_i. *Fig. 9* illustrates schematically the strategy STRATZ that carries out this search; since *all* the Y patterns now have to be investigated in each copy, this strategy requires a great deal more computer time than the STRAT strategy, in which only specific Y values were involved.

If the STRATZ strategy also fails to find a test for particular faults there still remains one possibility to be investigated. Depending on the places where the loops in the model shown in fig. 6*a* happen to have been broken, faults will exist that cannot appear directly at a Z output. These faults — *if* they are detectable — first have to emerge via a Y output of the copy before last. This means that such faults are only detectable if the value k appears at one of the Y inputs in the last copy. The FOR procedure only propagates k values to the right, and BACK cannot introduce any new k values at all. If we want to have k values at Y inputs in the last copy, we must therefore obtain them with a LOOK procedure in the penultimate copy. For each of the faults remaining after STRATZ we must thus investigate whether a Y pattern exists that will cause the fault to appear at a Y output and in the next succeeding copy at a Z output. This is investigated by the strategy STRATYZ (*fig. 10*). Here again *all* the Y values in a copy must be investigated, and therefore this strategy also requires a great deal of computer time. STRATYZ is consequently used only if faults without a test sequence remain after STRATZ. Only when neither

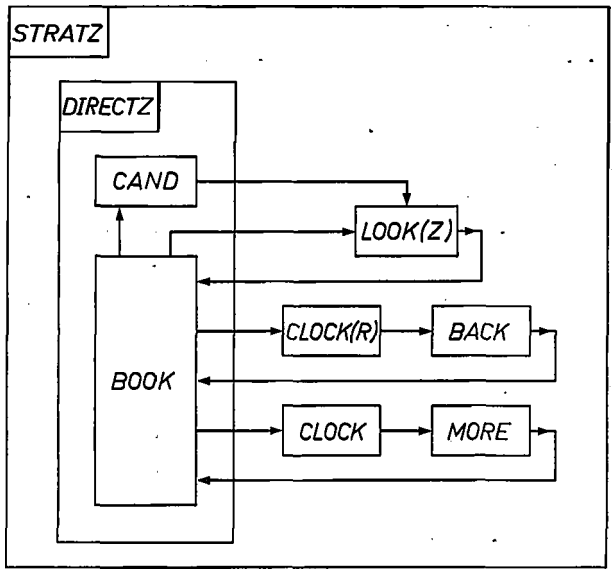


Fig. 9. Strategy STRATZ for investigating the testability of faults for which STRAT has not found any test sequence. For each of these faults, indicated by CAND, the procedure LOOK(Z) searches for patterns Y_j with which the fault appears at a Z output. The existence of one or more such patterns indicates that the fault is testable. To investigate which of the patterns may emerge from the initialization result Y_i via a particular test sequence, the CLOCK(R) procedure starts the BACK procedure in the preceding copy and calculates in this the corresponding input patterns Y_{j-1}. In this way the model in fig. 6*b* is computed from right to left (CLOCK(R) is a CLOCK procedure that works in reverse) until a Y pattern is found that is identical with Y_i, thus yielding the X patterns that form the test sequence. Finally the CLOCK and MORE procedures investigate whether this sequence is capable of detecting other faults as well.

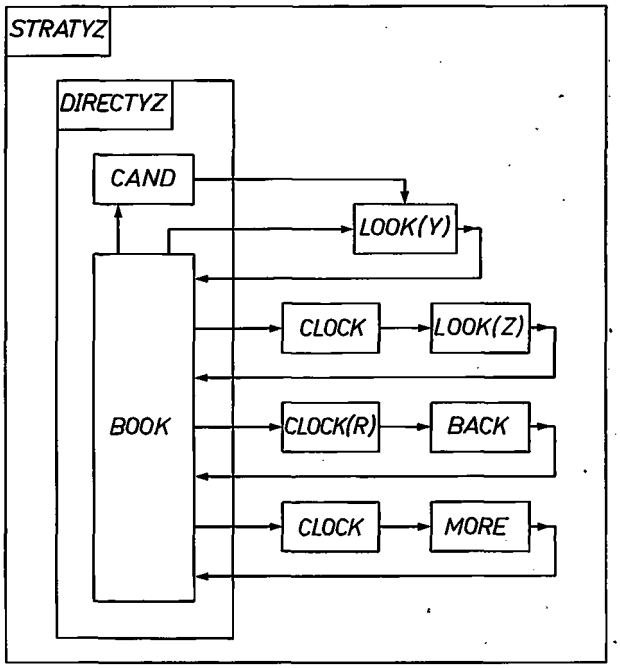


Fig. 10. The strategy STRATYZ used for investigating whether a fault in the penultimate copy appears at a Y output and in the last copy at a Z output. The procedure LOOK(Y) first calculates all the Y patterns with which the fault appears at a Y output. Then the CLOCK and LOOK(Z) procedures investigate to find the cases for which the fault appears in the next copy at a Z output. If such a pair of patterns Y_{j-1}, Y_j is found, the fault is testable. The test sequence is then determined in the same way as in STRAT(Z) by calculating back to the initialization result.

STRATZ nor STRATYZ can find a test may we conclude that a fault is fundamentally incapable of being tested.

In addition to the strategies described above the program TESTSC contains various other facilities, all of which are designed to limit computer time. The designer may for example supply a number of input sequences that he thinks likely to be able to detect faults. The ability of these sequences to detect faults is then determined by means of the procedure which does this in any case for the initialization sequence. It is also possible to generate sequences at random: in this way a large number of test sequences may sometimes be made available with relatively little computer time. These latter procedures are run through in the program TESTSC before the strategy STRAT.

Applications

The TESTCC and TESTSC programs can be used for a wide range of networks, including LSI circuits and networks formed on printed-wiring boards. More complex systems, such as computer subsystems, can also be handled as single entities. For practical reasons a limit is set to the number of lines in a network on which faults can be detected. Networks with a maximum of about 1000 gate circuits can be handled.

It is not easy to give figures for the computer time required, because it depends to a great extent on the type of network to be tested. Shift registers and counters in particular tend to increase the time considerably. However, if good use is made in such cases of the possibility of investigating in advance the usefulness of test sequences supplied, this effect can be eliminated. Taking as an example a sequential network with 1000 gate circuits, the TESTSC program on a computer such as the IBM 370/165 will generally need a few hours of computer time (and a storage capacity of 250 k bytes) to generate the whole structural test program, provided no untestable faults are found. The generation of one test pattern takes about a few hundredths of a second on the computer. If untestable faults are found, which tends to happen in the design phase, it may cost a few hours of computer time before they can be established with certainty. If it is possible however, to divide up the design of a sequential network in such a way that the investigation can be performed on a few combinational networks, a computer time of a few minutes to about half an hour will be sufficient.

What takes most time and is therefore most expensive is determining whether faults are fundamentally incapable of being tested. Nevertheless this is in fact one of the most important features of these programs, since they enable design faults to be detected that would otherwise not appear until a later stage, possibly not until manufacture or later. This can lead to considerable savings, expecially in the development of LSI circuits.

Summary. The functional test is made to determine whether the design of a logic circuit properly reproduces the desired logic function. In the complete test all possible combinations of input signals are fed to the network and the responses to these test patterns are compared with the responses of the fault-free circuit, which is known from the truth table. For a large circuit, such as a digital LSI network, this is not feasible; even if the test patterns were generated by a computer the test would take far too long (up to a few hundred years). It is possible, however, to make do with far fewer test patterns. Starting from the structure of a network a small group of test patterns can be derived that enable the test to be carried out in a short time with 100% certainty. This abbreviated test is referred to as 'structural'. This article describes two programs for deriving these patterns: the program TESTCC is used for combinational networks, in which the response depends only on the instantaneous input signal and in which a fault is detected by one test pattern. The TESTSC program is used for deriving test sequences for sequential networks, in which the response is determined by series of input patterns.

The article begins with a brief review of the faults that can occur during the various phases in the design of a digital network. In the structural test it is only necessary to take account of 'stuck at' faults on the various connecting lines in the network; the value on such a line remains fixed at '0' or '1' and does not respond to the other signals. A strategy is described that searches for a critical path for each s.a. fault in a combinational network. This is a path through the network that enables the fault to become visible at an output; the input pattern that brings this about is then the test pattern for this fault. To derive test sequences for a sequential network a model is used in which the network is represented as a series of identical combinational networks. These are derived from the sequential network by taking the internal feedback lines outside the circuit and breaking them. In this model successive input patterns are fed to a series of successive copies of the network. For each s.a. fault a search for a critical path is then made for a number of these copies. If the search is successful, the appropriate input patterns form the test sequence. If no critical path is found for a particular fault, the conclusion can be drawn that the fault is not capable of being tested. This usually indicates a design fault; the discovery of such faults, apart from finding the test patterns and sequences, is the principal purpose of the programs. Networks with a maximum of 1000 gates can be processed. The computer time required depends largely on the type of network; with a conventional large computer it may vary from a few minutes for a simple combinational network to a few hours for a complicated sequential network.

'MADGE', a microwave aircraft digital guidance equipment

II. The data link: data transmission and distance measurement

B. D. Corbett

This is the second of two articles about the MADGE portable aircraft-landing aid. The aid has been developed by the MEL Equipment Company in association with Mullard Research Laboratories for service with the British armed forces; it is also likely to find civil applications in the near future. The first article [*] described the principle of operation of the system and the interferometer receivers at the landing site which measure the angle of arrival of interrogating signals from the aircraft. This second article goes on to describe the digital data link between the landing site and the aircraft. The link passes angular guidance information from the interferometers to the aircraft and provides two-way communication for other purposes. It also measures the distance between the site and the aircraft. The link may be used simultaneously by up to two hundred aircraft, which share the available transmission time.

Introduction

The MADGE aircraft-landing system [1] is designed with both military and civil applications in mind. As was pointed out in part I, the basic concept of MADGE differs fundamentally from that of other systems which are at present in use or under development. The difference arises from the fact that the horizontal and vertical angle measurements needed for guidance are provided by passive arrays of interferometer receivers at the landing site. The angle-measuring interferometer receivers operate on the air-to-ground transmissions of a two-way data link between the aircraft and the site. The link is also used by the aircraft to measure the range from the site.

Two important features of the MADGE system are apparent. In the first place, no radio transmissions occur unless aircraft entering the approach zone require guidance or communication with the landing site. Secondly, the aircraft's three-dimensional position (in the form of angles and range) is made available simultaneously to the pilot and to an air-traffic controller at the site without the need for additional navigation systems or communication channels.

To a large extent the data link and the airborne instruments are independent of the arrangement and characteristics of the angle-measuring receivers on the ground. For instance the data link can accept a two-fold improvement in angular accuracy without modification. This gives the flexibility to provide for many different applications. Furthermore, the data link is not restricted to its guidance and navigation roles, but may be used for the simultaneous transmission of other digital messages at an information rate of over 1000 bits per second per aircraft (in both directions of transmission).

This second article is concerned with the airborne interrogator and ground transponder, which constitute the data link. Particular attention will be given to such topics as the choice of system parameters, traffic capacity, reliability of data transmission, data processing and the means whereby range and velocity are measured.

Data-link functions

The data associated with the elevation and azimuth guidance is communicated to the aircraft over the data link. The data-link antenna systems are designed to provide 360° of coverage in the horizontal plane, thus permitting range measurement and two-way data transmission on any bearing to a maximum radius of 30 nautical miles (about 55 km).

The aircraft has separate antennas in the nose and tail sections. The antenna switch searches in the ab-

B. D. Corbett, B.A., is with the MEL Equipment Co. Ltd., Crawley, Sussex, England.

[*] R. N. Alcock, D. A. Lucas and R. P. Vincent, 'MADGE', a microwave aircraft digital guidance equipment, I. General principles and angle-measuring units, Philips tech. Rev. 34, 225-241, 1974 (No. 9).

[1] See also S. J. Robinson, MADGE, a portable aircraft landing aid, Philips Telecomm. Rev. 32, 155-167, 1974 (No. 3).

sence of signals and locks when an adequate signal is acquired. The selection is also based on the rate of change of measured range. The nose antenna is selected as the range decreases and the tail antenna is selected when range increases.

The system operates in the band 5.00-5.25 GHz with different frequencies for air-to-ground interrogations and ground-to-air replies. Twelve alternative pairs of frequency channels may be selected. Amplitude modulation is used; messages consist of trains of pulses at a bit rate of about 1 megabit per second.

There are two types of interrogation and reply message, which are transmitted on the same pair of carrier frequencies but at different times, forming effectively two channels. These are known as the A- and the B-channel messages. The A channel is concerned with transfer of guidance data and the messages consist of an interrogation J_A and a reply Y_A. The B channel is concerned with transfer of other digital data and the messages consist of an interrogation J_B and a reply Y_B.

Interrogation

The pilot first selects frequency channel and ground-address code by means of his control unit, or 'controller' as it is called: he can then start to interrogate the ground station of his choice. The guidance interrogation word, J_A, is 25 bits (about 25 μs) long and contains a marker or 'location' pulse followed by address codes and error-checking bits (parity bits). The B-channel interrogation word, J_B, is 60 bits long. It is transmitted in the interval between A-channel interrogations when auxiliary data (or requests for data) is to be sent to the ground station. This auxiliary data includes range, which is derived in the air.

The interrogation word is transmitted at a mean rate of 50 times per second in the A or guidance channel. The interval between successive interrogations is randomly varied in order to avoid coherent addition of sampling errors caused by repetitive interference from other interrogating aircraft on the same frequency, or from other sources.

Processing at the ground station

Interrogations are received simultaneously on the ground by the interferometer receivers and by the transponder. Each interferometer generates an angle word within about 3 μs of the time of arrival of the first pulse in the interrogation word. The angle data is transmitted by cable to the transponder.

When the transponder detects a signal it stores the angle data received from the interferometers. It digitally filters the data before assembling the complete reply word for transmission. Digital filters are provided in the transponder for determining whether the measured angles fall within defined sector boundaries. The filters also generate a 'low-fly' warning when the measured elevation angle is below a safe limit which may be set by means of a switch on the transponder.

When the transponder has received the complete interrogation word it carries out a validity check. A-channel interrogations are subjected to a check of 16 bits including ground address and parity. The reply is generated 1 μs after the end of the interrogation word: it includes the air address (10 bits) which is sent back as received. Reply words Y_A and Y_B are each 60 bits long with a duration of about 60 μs.

Processing in the aircraft

The airborne logic unit carries out a validity check on each reply: this includes checks on ground and air address codes and parity. The range-measuring system is activated when more than 22% of interrogations result in valid replies.

Range is measured by means of a counter which is started from zero each time an interrogation is generated and stopped by a valid reply; the counter operates at a frequency of 16.1800 MHz, giving a range-measuring unit of 0.005 nautical miles or 9.26 metres. The range count is stored digitally and updated only for those replies that satisfy the validity check.

As soon as range is acquired the incoming replies are filtered by a *tracking gate*. This is a gate that will only accept replies whose leading edges fall within a window of 3-μs duration. The timing of the window is controlled by the stored range digits. The tracking gate ensures that, even when many aircraft are interrogating with the same address code, the probability of accepting false data is extremely low: it is typically less than 10^{-12}.

The pilot may select a glide slope angle β_0, an azimuth approach angle θ_0 and an azimuth overshoot angle on his controller. These reference angles are digitally subtracted from the elevation angle β and azimuth angle θ contained in each validated reply. This gives the angular deviation of the aircraft from the chosen flight path. The quantity $\beta - \beta_0$ is the vertical deviation and $\theta - \theta_0$ is the horizontal deviation.

As well as the digital outputs the logic unit generates a number of analog outputs for guidance purposes. Range r is derived from the range counter; range rate or velocity v is derived by analog differentiation of r. The angular deviations $\beta - \beta_0$ and $\theta - \theta_0$ are available in an analog form suitable for a standard flight instrument or autopilot. At close range the angular deviations become very large and the pilot prefers to be presented with linear rather than angular displacement. This is accomplished by a process known as 'beam-softening' in which the outputs become proportional

to $r(\beta - \beta_0)$ and $r(\theta - \theta_0)$ below a certain range (about 1 n mile). The logic unit performs this function automatically.

The absolute azimuth and elevation angles, θ and β, are also available in analog form: when displayed on suitable instruments, they can assist the pilot in locating the approach path when the aircraft first enters the coverage zone.

Direct-current outputs are provided for the operation of *fault flags* on the aircraft's instruments. These fault flags are small tabs or vanes that come into view as soon as they are de-energized; they give warning of malfunction and can be used to show that data is invalid or unreliable. Direct-current outputs are also provided for other indications, e.g. when the aircraft is flying low or is within the overshoot sector.

General description of the data link

Transmitters

A transmitting device with about 250 W of peak pulse power is needed at each end of the link to provide the required coverage whilst allowing an acceptable margin for possible system losses. The duty cycle of the transmitter in the ground transponder is determined by the number of aircraft which are interrogating it. In the military version, the duty cycle is limited automatically by blanking the transponder receiver periodically in order to prevent overload. A maximum duty-cycle limit of 6% is imposed, which allows 100 aircraft to receive guidance simultaneously from a single transponder at a mean reply rate of 16 replies per second per aircraft. A civil version of the system may use two transponders with provision for time-sharing; in this form the system is capable of handling up to 200 aircraft.

For the transmitters a travelling-wave amplifier tube was chosen with a peak power rating of 250 W at a duty cycle of 7%. This tube has a gain of 40 dB over the operating band and requires a drive power of only 25 mW. The air and ground transmitter circuits are essentially identical apart from minor differences arising from the higher mean power dissipation in the transponder; in the aircraft transmitter the mean radio-frequency power is never more than about 1 W.

The r.f. drive to the travelling-wave tubes is supplied by solid-state phase-locked microwave sources. The frequency reference is a crystal oscillator operating at about 90 MHz (the output frequency divided by 56). This type of source provides very high rejection of unwanted harmonics.

The pulse-modulator circuits for the travelling-wave tubes are shown schematically in *fig. 1*. The output of the travelling-wave tube is connected by coaxial cable

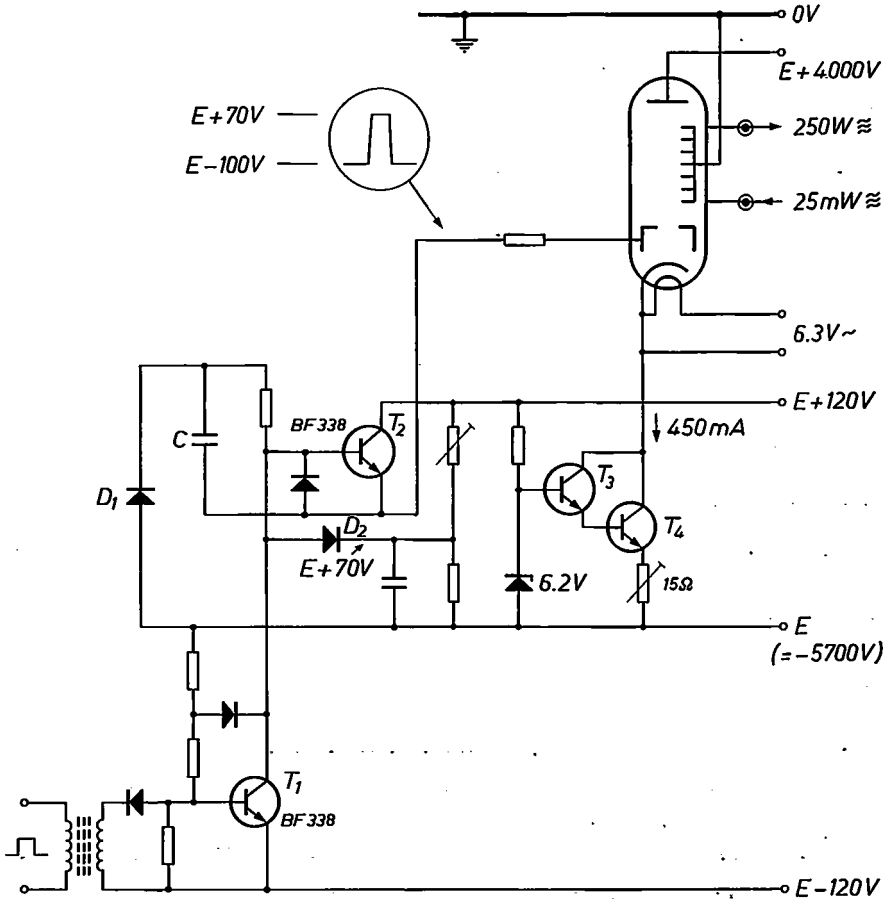


Fig. 1. Travelling-wave-tube pulse modulator. The cathode of the travelling-wave tube is at about −5.7 kV relative to ground. The tube is cut off by a grid voltage of −100 V relative to cathode. Modulating pulses at logic level are applied to the primary winding of a well insulated isolating transformer. In the absence of a pulse the transistor T_1 is conducting and the capacitor C is charged via diode D_1. An input pulse causes T_1 to switch off and T_2 to conduct. The charge on C holds T_2 on. The grid voltage rises to +70 V in 150 ns and is clamped by the diode D_2. The cathode current is controlled by means of the constant-current generator T_3, T_4. Its value is set to provide a peak r.f. output power of 250 W. At the end of the pulse the grid capacitance is rapidly discharged by the conduction of T_1.

to a microstrip transmit-receive module. This module is shown in *fig. 2*; it routes the signals to and from the antenna, contains the mixer stage of the receiver and has provision for power monitoring and test-signal injection in self-test routines.

Another important aspect of receiver performance is its behaviour in the presence of multipath reflections or echoes from obstacles such as buildings or other aircraft. These reflected signals can tend either to reinforce or cancel the direct signal; they affect the amplitude

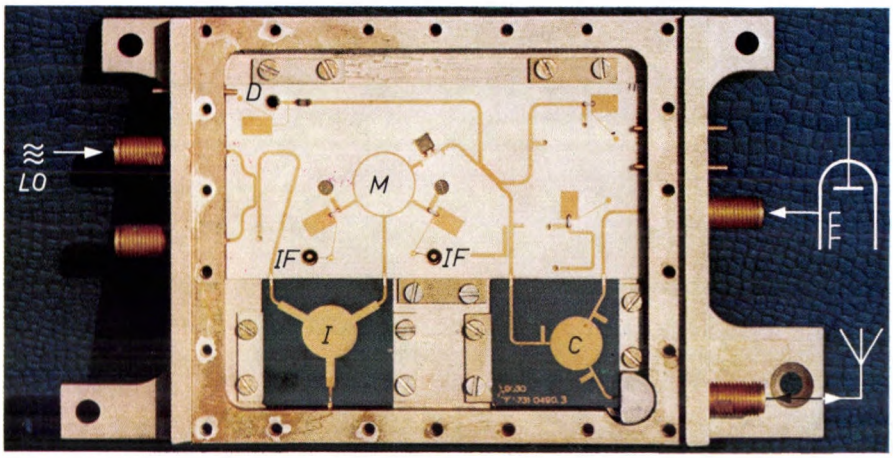


Fig. 2. Microstrip transmit-receive module on ceramic alumina substrate. The transmitter signal from the travelling-wave tube enters the module at the right and passes via a power-monitoring coupler-detector to the circulator *C* and to the antenna port (*lower right*). The received signal enters the antenna port and passes through the circulator *C*, image-suppression filter, excess-power monitoring coupler-detector and limiter to the balanced mixer *M*. Excess power at the detector (*at top right*) generates a signal for switching off the travelling-wave tube in order to protect the receiver. Power from the local oscillator is injected at the upper left-hand port *LO* and is divided. Some of this power passes through the isolator *I* to the mixer; the rest drives the travelling-wave tube. *IF* intermediate-frequency outputs. *D* noise diode for injection of a test signal.

Receivers

Both data-link receivers have a sensitivity of —89 dB relative to 1 mW and a bandwidth of about 3 MHz. The bandwidth is a compromise between the conflicting requirements of fast pulse response and adjacent-channel rejection.

Accurate range measurement, and the need for precise timing of the bit-sampling clocks, make stringent demands upon the design of the circuits controlling the gain and detection threshold. The transponder receiver needs to regain full sensitivity immediately after exposure to very strong signals, for example interrogations from nearby aircraft. It therefore uses a logarithmic amplifier in preference to an automatic-gain-control system with its inherent time delays.

Achievement of optimum sensitivity, consistent with the prevailing conditions of pulse and continuous-wave interference and circuit noise, is ensured by means of a *constant-false-alarm-rate* (CFAR) system; the general arrangement is shown in *fig. 3*. A control loop automatically adjusts the receiver threshold level so that it becomes less sensitive as the interfering pulse rate increases. This results in a constant noise-pulse frequency at the video output.

and distort the edges of pulses in the received messages. At typical aircraft speeds the change in pulse amplitude tends to be constant for the duration of a 60-μs word, but varies randomly from one word to the next. When the multipath delay approaches or exceeds half a microsecond the data may be 'corrupted'. This means that a 'one' might change to a 'zero' or vice versa. The peak detector incorporated in the threshold-control system is arranged to set the detection threshold at one-half of the amplitude of the first pulse in the word train (this pulse is always three or four microseconds long). The threshold level is then held for the duration of the incoming word even when the bit pattern includes long strings of 'zeros'. In this way multipath echoes which appear in the spaces of the pulse train are suppressed provided their amplitude is less than half that of the first pulse in the message train.

The transponder

In *fig. 4* a photograph of a prototype transponder is shown. It weighs 20 kg and is powered by an external 24-volt battery but is otherwise self-contained.

The preferred design of the antenna which provides 360° horizontal coverage is a rectangular slotted wave-

guide with a tapered matching section at its base. The vertical pattern of this antenna has to be closely controlled in order to obtain high gain at low angles of elevation; this has been achieved by computer-aided design techniques.

The interrogator

In *fig. 5* the main components of the aircraft equipment are shown. The transmitter-receiver unit and the logic unit are both powered from a 28-volt d.c. supply and weigh about 5 kg each. The pilot's controller is

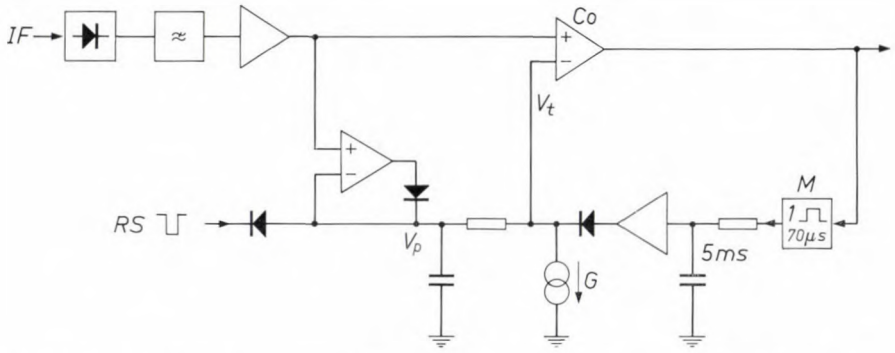


Fig. 3. Constant-false-alarm-rate (CFAR) threshold control in the ground receiver. The intermediate-frequency signal *IF* is derived from a logarithmic amplifier; it is detected, filtered, amplified and compared with a threshold voltage V_t. Any signal exceeding V_t results in a logic-one level at the output of the comparator *Co* and this triggers a fast-recovery monostable *M* with a period of 70 μs. *M* is triggered by a data-link message or any isolated noise spike or interfering pulse. Its output is smoothed by a lowpass filter with a time constant of 5 ms and added to a peak-detected component V_p of the detected signal. The constant-current generator *G* sets V_t lower than V_p by a constant amount; the threshold level is thus 6 dB below peak signal level. The threshold-control loop automatically adjusts V_t to maintain a constant pulse rate at the comparator output; in this way the receiver becomes less sensitive as the interfering pulse rate increases. The peak detector is reset by *RS* at the end of the received word.

The transponder can accept digital angle data from the interferometers over distances of up to 2 km. This data is sent serially on a balanced cable at the system bit rate of 1.01125 megabits per second. A self-clocking code is used: this means that timing information, needed for decoding the data, is included in the message and only one balanced cable is needed. The transponder can send B-channel and air-traffic-monitoring information by cable to a remote point; the same self-clocking code is used for this.

The transponder initiates a self-test sequence at 3/4-second intervals which checks the correct functioning of both the transponder and the interferometers. The results of the ground-unit checks are transmitted to the aircraft in each A-channel reply and may be sent by cable to a remote monitoring point. In this way any ground-unit malfunction is brought to the attention of both pilots and ground staff. In order to save battery power the transponder causes the interferometers to be switched off when no aircraft are interrogating the system.

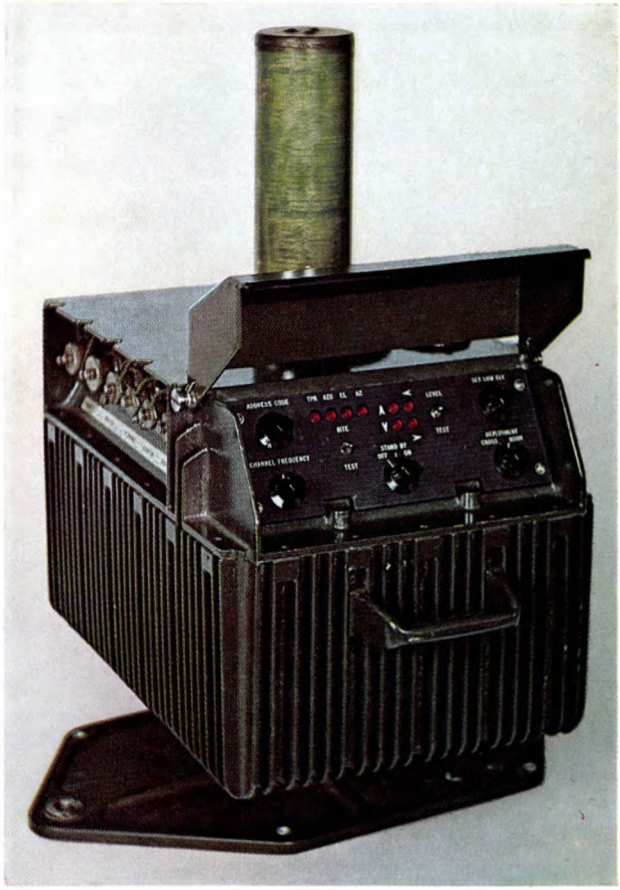


Fig. 4. The ground transponder (early military prototype). It weighs 20 kg and can withstand parachute dropping and manhandling over rough terrain. The control panel is seen at the front. The cylindrical omnidirectional antenna is mounted on top.

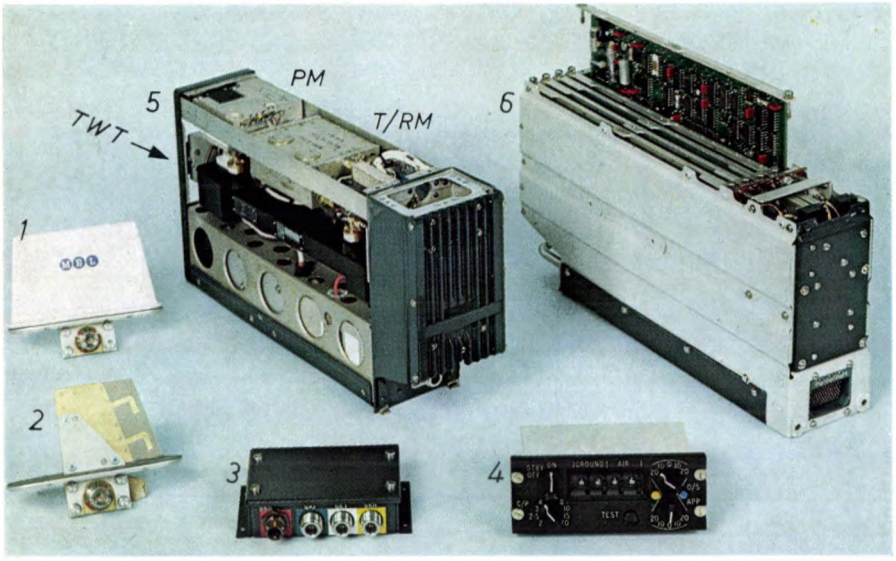


Fig. 5. The airborne interrogator set.
1 antenna
2 second antenna with radome removed
3 antenna-switching relay
4 pilot's controller
5 transmitter-receiver unit
6 logic unit
TWT travelling-wave tube
PM pulse modulator
T/RM transmit-receive module

used for the selection of addresses, frequency channels and flight-path angles; it contains only switches.

Two aircraft antennas are needed because the geometry of most aircraft makes it impossible to realize 360° of coverage in the horizontal plane from a single antenna. Each antenna consists of a printed-circuit stacked-dipole array housed in a thin plastic radome; the aerodynamic and structural characteristics make it suitable for supersonic aircraft. The vertical coverage is about ± 20° to the half-power points. The feeders are 50-ohm coaxial cables.

The pilot may check the operation of all circuits and flight instruments by means of a press-to-test button in the controller. System failures are indicated by fault flags on the aircraft's instruments. *Fig. 6* shows some of the instruments used with MADGE in a typical helicopter installation.

System parameters and traffic capacity

In MADGE, an *air-initiated* system has been adopted: the aircraft interrogates and the ground station transponds; range is derived in the air and transmitted to the ground. Aircraft interrogations cannot readily be synchronized and this results in a *random-access* system in which interrogations occur on a random basis and messages inevitably overlap.

The alternative choice would have been a *ground-initiated* system in which the ground station interrogates and the aircraft transpond; range would be derived on the ground and transmitted to the aircraft with the angle data. A ground-initiated system would be *time-ordered:* each aircraft would be interrogated in turn at an interval which allows sufficient time for the transponded message to be received. In this way there are no overlapping messages.

At first sight it might be thought that the air-initiated random-access system is at a disadvantage in a dense-traffic situation on account of overlapping interrogations arriving at the ground transponder. Erroneous data resulting from word overlap is eliminated by address matching and by the use of redundant codes. A number of replies are lost in this way, however, and this must be allowed for by making the interrogation rate higher than the required reply rate.

On the other hand, in a simple ground-initiated time-ordered system the number of aircraft which can use the system is limited by the fact that the ground station must allow enough time between interrogations in order to receive replies from aircraft at maximum range. Furthermore, the requirement for operation in the neighbourhood of other ground stations on the same frequencies makes effective time-ordering impractical. In the MADGE air-initiated system the ground station is capable in principle of handling up to 200 aircraft at one time. The same total number of aircraft can be handled by a number of ground stations within radio range of each other, all sharing the same radio-frequency channels.

Air initiation has the great added advantage that no special acceptance procedures are necessary when a new aircraft wishes to obtain guidance. This leads to flexibility of application, making the system equally suitable for temporary landing strips or large fixed installations.

Traffic capacity

The traffic capacity of a system may be defined as the maximum number of aircraft which can simultaneously obtain satisfactory guidance from one ground transponder. The reply rate from a transponder to a particular aircraft decreases as the total number of interrogating aircraft increases and once the reply rate falls

Fig. 6. Instrumentation in a MADGE-equipped Gazelle helicopter. *Top row, second from right:* two-axis horizon; the crossed pointers are actuated by the MADGE azimuth and elevation guidance outputs. *Top row, at left:* display developed for MADGE; a radial pointer indicates range up to 30 n miles on a nonlinear scale; an edgewise pointer gives azimuth deviations from the MADGE centre-line up to $\pm 45°$. *Bottom left:* pilot's controller (see *4* in fig. 5). *Above this:* 'doll's-eye' indicators for safe elevation angle and overshoot sector indication.

below a certain threshold value satisfactory guidance can no longer be obtained. The traffic capacity is equal to the number of interrogating aircraft corresponding to this threshold value. When several ground transponders share the same frequency channel and are within radio range of each other it is convenient to regard them as forming a single system and to define the traffic capacity in terms of the total number of aircraft receiving replies.

The reply rate received at the aircraft has been the subject of a detailed theoretical analysis. The validity of this analysis has been confirmed by a 'Monte-Carlo' computer simulation and by experimental evidence. A simplified account of the theoretical analysis is given in Appendix I. The results are shown by *fig. 7*, together with a number of plotted points obtained from the Monte-Carlo computer simulation.

In establishing the minimum acceptable validated-reply rate for adequate guidance, angle information is the critical variable; it is acquired when the validated-reply rate exceeds 15 per second, and is then continuously available until the rate drops below 8 per second.

Fig. 7 shows that with 200 aircraft the rate always exceeds the 15-per-second acquisition threshold by a suitable margin and is rather more than 20 per second in the case of a single ground station. This allows for some loss of data because of noise and interference originating outside the system. Thus the requirement for handling 200 aircraft is met in all cases considered. The CFAR system is organized in such a way that the probability of detection (unlike the probability of validation) is scarcely affected by traffic density. In fact the results are pessimistic for aircraft at relatively close range because the tendency for strong signals to suppress weaker ones has not been taken into account in the analysis.

numbers generated by this circuit have an additional purpose. They provide a random or universal air address which may be used by aircraft entering the system without an assigned address.

Data integrity and message format

The term 'data integrity' refers to the ideal case in which the data output at a receiving terminal corresponds exactly to the data addressed to an input terminal at the transmitting station. In practice we require that corruption of a message by crosstalk or interference should be eliminated or signalled in a reliable way when it occurs; furthermore, noise and distortion

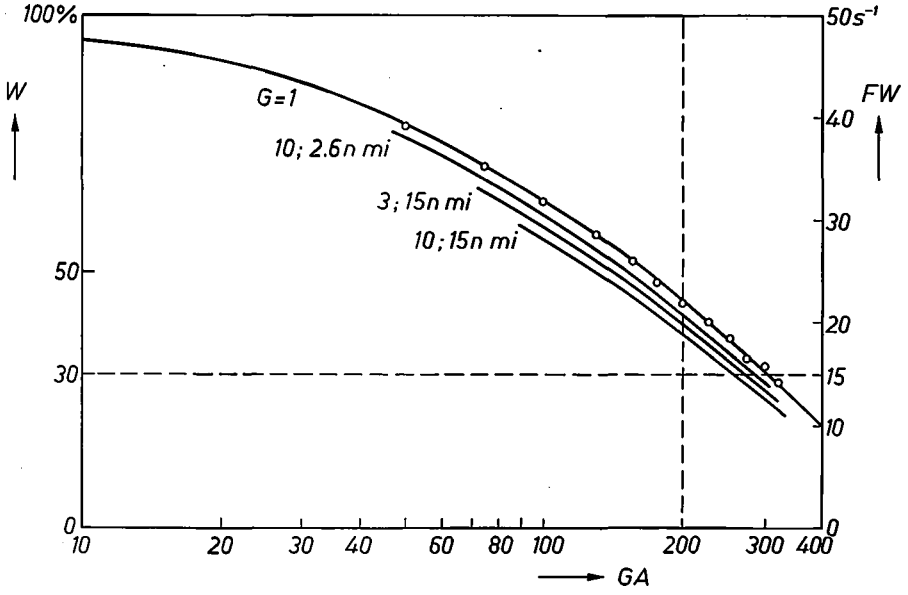


Fig. 7. Graph of reply efficiency W as a function of the number of interrogating aircraft GA. The graph shows the proportion of aircraft interrogations that results in uncorrupted replies. Since the interrogation rate is F per second, the right-hand scale FW shows the uncorrupted reply rate. The plotted points were obtained from a computer model, using the 'Monte-Carlo' method; the curves were obtained by calculation. Curves are shown for a number of ground stations G up to 10 on the same frequency channel and for maximum separations of 2.6 and 15 nautical miles. With 200 interrogating aircraft the valid-reply rate exceeds the 15-per-second rate required for acquisition of angle data. (In the case of multiple ground stations an equiprobable distribution of radii to the interfering ground stations has been assumed.)

Interrogation jitter and random addresses

The interrogations generated by the airborne unit do not follow one another at equal 20-ms intervals but are subjected to a defined amount of random jitter (± 2 ms peak). In fact the interval between one interrogation and the next is independent of the previous interval and is selected from 16 possible values. This prevents accidental synchronism with repetitive interference originating either within the system or from other sources.

The random-number generator which controls the interrogation rate is shown in *fig. 8*. The 4-bit random

of the signals must be reduced to acceptable levels and any possibility of malfunction of the selective addressing system must result in a rapid warning.

The attainment of these objectives in the MADGE data link has implications in the selection of system parameters, message format and data codes. In the MADGE system the data messages are conveyed by 60-bit words. The choice of the bit period is governed by two conflicting considerations. For a given information content a low bit-clock frequency would result in words of long duration and a consequent reduction of traffic capacity. On the other hand, the bandwidth

restriction of 3 MHz sets a limit to the minimum width of the pulses used for encoding the data. For these reasons a bit period of about 1 μs was selected for the system. The actual bit rate (1.01125 Mbits/s) is an outcome of the choice of range-measuring unit (0.005 n miles or 9.26 m): for convenience the bit-clock frequency is derived by dividing the range-measuring-clock frequency by 16.

General arrangement of format

Each 'one' is encoded by a pulse occupying the whole bit period (non-return-to-zero format). It can be seen from *Table I* that the message begins with a marker pulse which is three bits (about 3 μs) long. In the guidance channel the subsequent 'one' of the channel address extends the marker pulse to four bits so that the interrogation word provides a pulse of adequate duration for the interferometer to process. Another function of the marker pulse is to provide information to the receiver video-processing circuits to enable the comparator detection threshold to be set at an optimum level for detecting the remainder of the pulse train and maintaining synchronism with the bit clock.

Table I. Format of 60-bit message. Applies to A-channel reply (Y_A), B-channel interrogation (J_B), and B-channel reply (Y_B). The A-channel interrogation word (J_A) has only 25 bits of which the first 15 follow the sequence given in this table. Bit rate: 1.01125 megabits per second.

Sequential bit number	No. of bits	Function	Remarks
1-3	3	marker pulse	'111', extended by bit 4 in A channel
4-5	2	channel address	'10' = A channel, '01' = B channel
6-15	10	air address	includes 2 bits for special purposes
16-25	10	10-bit data word	Y_A: elevation data J_B, Y_B: auxiliary message
26-40	15	15-bit data word	Y_A: azimuth and flags J_B: range Y_B: auxiliary message
41-43	3	column-parity bits	cover 15 most significant bits, 20-34
44-46	3	row-parity bits	associated with 26-40
47-48	2	presence indicator	'10' = 26-40 valid
49-50	2	row-parity bits	associated with 16-25
51-52	2	presence indicator	'10' = 16-25 valid
53-54	2	procedure code	'10' = B-channel call-up in Y_A
55-60	6	ground address	59-60 are '00' in current systems

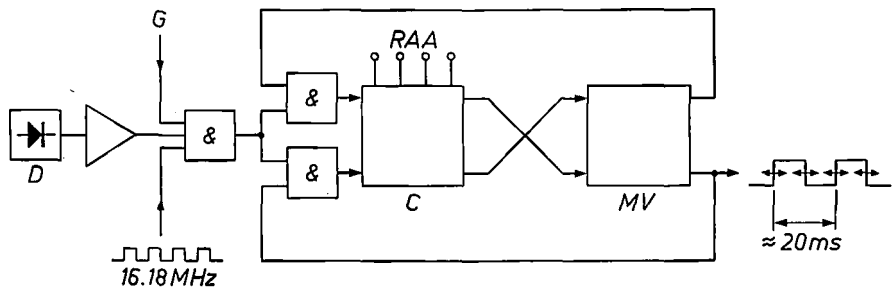


Fig. 8. Random-number generator. This circuit triggers interrogations at a mean interval of 20 ms with a random jitter of up to ± 2 ms. It also provides random aircraft addresses. The multivibrator circuit *MV* sets the interval between interrogations. The pulse widths are controlled by the states of two two-bit counters *C*. The counters are activated alternately by the 16-MHz clock. The clock is gated randomly by noise originating in a diode *D*. The counters generate the random air address at the terminals *RAA*. A gate level *G* is used to hold the random address *RAA* during the reply-checking period which follows an interrogation.

The message begins and ends with a sequence of address bits and check or parity bits, whilst the 25 bits of data occupy an intermediate position (16-40) with the most significant bits at the centre. This is done in order to reduce the probability of accepting corrupted data. When the message is overlapped at either end by an interfering word there will be a high probability that the signal will be rejected before the overlap extends into the data-carrying portion of the message.

The use of amplitude modulation on the radio-frequency carrier has a bearing on the characteristics and distribution of the check bits. Check bits at logic one are required in order to detect possible data-corruption by fading. But pulse interference is much more likely to falsify message bits at logic zero. It therefore follows that an adequate number of check bits in each state must be provided. The 25 bits of data are sent in *complementary* form (meaning that '0' is sent as '1' and '1' is sent as '0'); this is done in order to reduce the error-to-signal ratio caused by pulse interference when the value of the encoded variable is small.

Clock pulses have to be generated for the purpose of shifting and sampling the digital messages as they are received. Synchronization is controlled by the leading edge of the marker pulse; as the displacement of this edge (due to noise or interference) approaches a value of about 0.5 μs the resultant loss of synchronism will prevent validation and cause rejection of the incoming word.

Validation conditions and parity bits

Validation is the process of checking digital addresses against preset patterns and verifying that the redundant bits obey predetermined rules. The redundancy in the MADGE messages is based on parity codes. Channel-switching commands, and certain other single-bit message elements demanding high integrity, are sent as *complementary pairs*. In a complementary pair '0' is sent as '01' and '1' is sent as '10'. This may be interpreted as a parity grouping comprising one data bit and one parity bit. The other parity groups comprise five data bits and one parity bit. Both these parity groupings contain an odd-parity bit and an odd number of data bits. The advantage of this arrangement is that a total fade (giving a sequence of 'zeros') and continuous-wave interference (giving a sequence of 'ones') are detected as erroneous states and rejected.

The central data section of the 60-bit word consists of 25 bits of data, of which 15 bits are protected by *two-coordinate parity*. This arrangement is explained by reference to *fig. 9*. The data bits are assigned to various positions in a 5×5 matrix. The five rows form parity groups in association with the parity bits on the left (44, 45, etc., see Table I). Three of the columns, which contain all the more significant bits, form three additional parity groups associated with parity bits 41, 42 and 43.

It has been shown that codes using a two-coordinate parity check are well protected against noise occurring in random patterns [2]. In the MADGE data channels errors due to receiver noise are randomly distributed and affect single bits, but an interfering pulse may falsify an even number of adjacent spaces in the message. Under pulse interference a message format in which the parity groups follow each other sequentially would thus have poor error-detection characteristics. This defect is prevented in the MADGE format by breaking up and interlacing the parity groups in the transmission sequence corresponding to the pattern shown in fig. 9. The probability of detecting bursts of up to 14 consecutive errors is 1 for the azimuth word and 0.985 for the elevation word.

44	38	35	29	30	31
45	37	39	26	33	28
46	40	36	32	27	34
50	17	19	23	21	25
49	16	18	20	24	22
			41	42	43

Fig. 9. Arrangement of parity groups. The data bits numbered 16 to 40 in order of transmission (see Table I) are allocated positions in a 5×5 matrix. The horizontal rows form 6-bit parity groups containing parity bits 44, 45 etc. The 15 most significant bits (20-34) also appear in the vertical columns, forming additional 6-bit parity groups with parity bits 41, 42, 43. These bits are doubly checked. The most significant bits of the azimuth word (26-34) lie in the shaded region.

Acquisition of range data in the interrogator depends on the validation of the marker pulse, air and ground addresses, four complementary pairs and the three 6-bit column parity groups. For azimuth and elevation data the appropriate row-parity conditions are added to the range-validation criteria, thus providing a two-coordinate check on all except the least-significant data bits. The reply efficiencies for range, azimuth and elevation data are monitored separately.

Probability of validation

The probability of validation of a reply depends on the rate at which single-bit errors occur. It can be shown by an analysis of the parity-checking system adopted here (see Appendix II) that the validation conditions imposed for azimuth data imply a bit-error probability of 3.45% when the probability of validation (V) is 16%. This value of V represents a mean data-refreshment rate of eight per second: if the valid-reply rate becomes lower than this figure, the aircraft receives no angle guidance but range data may still be available.

The probability of bit errors varies with the signal-to-noise ratio. It is influenced by the detection-threshold control system (CFAR system) following the intermediate-frequency detector. These effects are analysed in Appendix III and the results are presented in *fig. 10*: this gives the probability of validation as a function of signal-to-noise ratio. There are different curves for messages having different validation conditions.

Undetected errors

When validated, a reply may nevertheless contain occasional errors if the system is working near the limit of sensitivity. These undetected errors result in noise at the data outputs of a receiving terminal. Such noise depends on the rate at which validated words are accepted. It is greater at a lower validation rate, because every error lasts as long as the interval between one validation and the next. For this reason the lower limit of eight per second is set for the angle-data validation rate below which the data is cancelled even if it is valid.

A computation of the probability of accepting one or more errors per validation is given in Appendix IV.

When applied to the nine most significant azimuth bits, it gives the probability of accepting one or more errors per validation as less than 4×10^{-3} when the bit-error probability is at its worst value of 3.45%. Under these conditions the validation rate is eight per second; the mean time between accepted errors is therefore about half a minute.

Analog processing and flag control

We shall now consider various analog circuits whose function is to make the output data compatible with the needs of instruments and pilots. The electrically operated warning tabs or flags with which aircraft instruments are fitted are de-energized and come into view when the data becomes invalid or unreliable. Separate flag outputs are provided for azimuth (approach and overshoot), elevation, range and velocity data. Additional two-state outputs of a similar kind provide information such as the availability of ground units for guidance service and a warning to the pilot that the elevation angle is too low.

The flag-control circuits consist of analog integrators which measure the duty cycle of a composite digital function. The integrator output is arranged to operate a trigger circuit with sufficient backlash to avoid flutter or ambiguity. The fault-flag integrators measure the reply efficiency for the data concerned. Normally the time delay between the failure of a signal and the appearance of a fault flag is not more than one second; in the case of the range flag, however, a short additional delay is generated in order to allow time to switch to the alternative antenna.

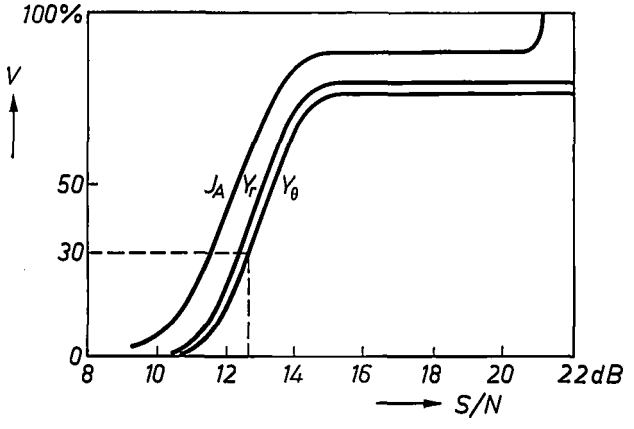


Fig. 10. Graph showing probability of validation V as a function of signal-to-noise ratio S/N. The curve J_A gives the probability of validation of the guidance interrogation at the transponder, Y_r and Y_θ give respectively the validation probabilities for range and azimuth-angle data at the interrogator. The plateau is caused by the CFAR control. In order to acquire azimuthal guidance a validated-reply rate of 15 per second or a validation probability of 30% is required.

Smoothing and rate-limiting

Since a higher noise builds up in the analog circuits as the data-refreshment rate becomes lower, a limit is set by cancelling the data when the rate is too low. In order to accept as low a rate as possible, noise is reduced in the analog circuits for range and angle data by means of an integrating capacitor. Its time constant (150 ms) may be augmented by additional lags in the read-out instruments. The analog noise resulting from undetected errors contains components of very low frequency, which are stored in the integrating capacitor. The additional lags in the read-out instruments do not help in suppressing these components.

In the case of the angle-processing circuits the undetected errors may produce full-scale excursions at the digital-to-analog interface. These errors are held at the interface until another reply is validated. Even when the validation rate is at its maximum value of 50 per second, the resulting excursions at the analog output would be unacceptable if reliance were placed on the time constant τ of 150 ms alone. For this reason the angle-processing circuits include rate-of-change limiters that restrict the rate at which the integrating capacitor can charge. The limit adopted is 50% of full-scale voltage per second: this may be written as V_m/τ_m where V_m is the full-scale voltage and τ_m is 2 seconds. The effect of the time constant τ by itself is to multiply the digital noise by the reducing factor $1/\sqrt{2VF\tau}$, where the product of the validation probability V and the interrogation rate F gives the data-refreshment rate. A rate limit of V_m/τ_m volts per second multiplies noise of amplitude V_m by the additional reducing factor τ/τ_m. Noise voltages at the integrator input are thus in general limited to $V_m\tau/\tau_m$.

The sensitivity of the guidance outputs can be adjusted to suit the aircraft type. In the case of azimuth guidance, angle errors are greatest at a low sensitivity, corresponding to 10° full-scale. When the validation rate has dropped to its minimum value (eight per second) the noise due to transmission errors has an r.m.s. value of 0.026°. The largest component of this noise is due to highly weighted single-bit errors accepted by the cross-linked portion of the parity array.

The mean amplitude of these rare transients at the analog output is 0.5° when the validation rate is eight per second: their decay time constant is 150 ms. With a typical external time constant of 300 ms, the peak amplitude is reduced to less than 0.2° at the guidance instrument. A momentary decrease of the validation rate below eight per second will trip the flag trigger before these transients can increase substantially.

[2] B. J. Beukelman, The undetected-error probability of codes using two-coordinate parity check and of cyclical codes, Philips Telecomm. Rev. 29, 188-204, 1971.

A simple rate-of-change limiter is shown in *fig. 11*. This circuit is used for the azimuth data, because the peak digitization noise generates voltage ramps here which are not steeper than the rate-of-change limit. In the case of the elevation guidance output this condition is not met. The elevation rate limiter is actuated by any excursion of the signal amplitude which exceeds the expected maximum.

'Beam-softening'

At short distances the pilot prefers the deviation from the flight path to be presented to him in metres instead of degrees. This is achieved by the procedure known as 'beam-softening'. Before display the angular differences $\theta - \theta_0$ and $\beta - \beta_0$ are multiplied by a factor which is a constant k at ranges above a certain value r_0 and linearly proportional to range below r_0. For example, the azimuth-guidance output A_0 is given by:

$$A_0 = k(\theta - \theta_0) \qquad \text{for } r \geqslant r_0,$$

$$A_0 = \frac{kr}{r_0}(\theta - \theta_0) \qquad \text{for } r < r_0,$$

where r_0 is set at 0.99 n miles (1.833 km). When k is chosen so that the full-scale deflection at long range is equivalent to a 5° angular displacement, at range r_0 it is equivalent to a displacement of 5° or 160 m; below r_0 is it interpreted as 160 m. This multiplying process is achieved by means of a digital-to-analog converter which uses an analog voltage derived from the range digits as a reference.

When the reply message indicates that the aircraft is within the overshoot sector the sensitivity factor k is automatically halved in order to coarsen the sensitivity of the guidance outputs: this is done to make it easier for the pilot to pick up the missed-approach flight path. At the same time the sign for both azimuth outputs is reversed in order to preserve the correct convention for left and right indication.

Range and velocity measurement

Range is derived by the aircraft from measurements of the time lapse between the end of the interrogation and the arrival of the reply. As stated in the description of the data-link functions, the time lapse is measured by means of a range clock operating at a frequency of 16.1800 MHz, one count being equivalent to a distance of 0.005 n miles (9.26 m).

As long as the proportion of validated replies remains below 22% (11 per second) the aircraft system operates in the *search mode*. The timing sequence of the range-measuring system in this mode is shown in *fig. 12*. At a precisely timed moment t_0 near the end of each interrogation word J the interrogator starts a fixed count D_1, which lasts approximately as long as the duration of the reply word Y. The range-measuring counter is started at the termination of D_1. It is stopped as soon as a reply received from the transponder has been validated; this takes place at an instant t_v near the end of the reply word. The range count R_1 then gives the range measured.

During the search mode it is important that time should not be wasted in processing interference. The first 15 bits of the message are therefore checked serially on receipt, and as soon as an error is detected the logic becomes free to process another train of pulses. In practice this means that even in dense-traffic conditions range is normally acquired within a fraction of a second.

As long as the reply efficiency stays below 22% the range count is ignored. When, however, it reaches this value the range count is transferred to the range store. The range fault flag then withdraws and the system adopts the *tracking-gate mode* which is illustrated in fig. 12*b*. In this mode range signals are available so long as the data-refreshment rate exceeds a mean value of 4.5 per second.

The stored range value indicates the time after the interrogation at which the reply may be expected. The tracking gate is arranged to open at 16 range-clock intervals (about 1 μs) in advance of the expected time of arrival of the reply and it closes about 2 μs after it. In this way a 3-μs window is provided for the leading edge of the reply word: this edge is used as a time reference for bit-clock synchronization and range measurement. At an approach speed of 1000 km/hour the 1-μs margin in advance allows 22 successive failures to validate, or 'misses', before reversion to the search mode becomes necessary.

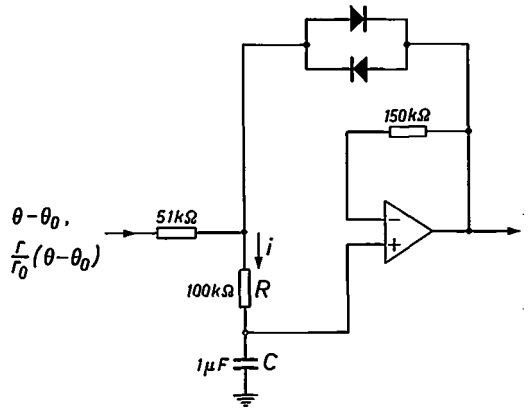


Fig. 11. Rate-of-change limiter circuit. For slowly varying signals the charging current i is small and the voltage iR developed across the resistor R is smaller than the forward voltage drop across the silicon diodes (450 mV). Limiting occurs when $iR \geqslant 450$ mV. The rate of change of voltage is then limited to 450 mV/RC = 4.5 V/s.

It is shown in fig. 12*b* that in the tracking-gate mode the fixed count D and the range count R occur in reversed order. The range count is started at t_0 and is stopped after the first bit of the marker pulse has been checked. Its value is then R_2, which is 32 range-clock units (2 μs) too great, first because the transponder introduces a 1-μs delay before replying to an interrogation, and, secondly, because the first bit of the reply word has been included in the range count. These 32 range-clock units are subsequently subtracted from R_2 before storing the range count.

When the range-clock count has reached a value equal to the stored value R_s obtained from the preceding validation, the tracking gate opens. Because of the 1-μs delay at the transponder this will be 1 μs ahead of the next arrival, discounting any small change in the range value due to the motion of the aircraft.

Both modes of operation use the same reference time, t_0; the relationship between t_0 and the trailing edge of the interrogation word may be finely adjusted so as to cancel timing errors due to circuit propagation delays.

Errors in the range measurement occur for a number of different reasons. One is the finite rise-time of the pulse whose leading edge serves as a time reference. Others are multipath propagation and thermal noise, which both distort the leading edge. Additional sources of error arise in the processing circuits.

Errors may be classified as *bias* or *noise*. Bias errors are those which may be treated as invariant over a period of about 15 minutes. Noise may be subdivided into *drift* and *jitter* components. Drift noise consists of variations in the band 0.001-2 Hz, which are too slow to be affected by analog filtering, or by digital averaging of 10 successive samples. Jitter components involve random changes occurring at the validation rate.

Pulse rise-time effects

The leading edge of the marker pulse is used as a timing reference for range measurement because all subsequent pulse edges are subject to possible distortion by multipath reflections originating from earlier parts of the message. The shape of this reference edge at the input to the video comparator is governed by the pulse response of the tuned intermediate-frequency amplifier and post-detector filter.

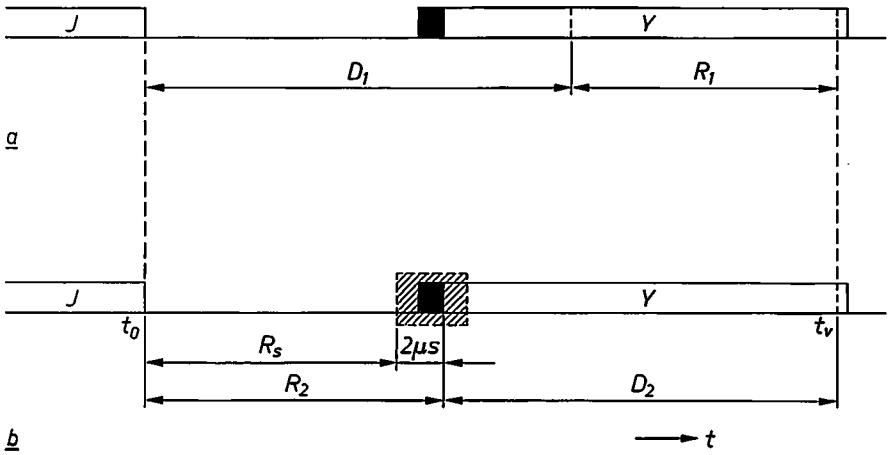


Fig. 12. Range timing diagrams. *a*) Timing sequence in the search mode; no data is displayed. At a particular moment t_0 near the end of the interrogation J a fixed count D_1 is started. At the end of D_1 the range count R_1 is started and stops at the exact moment t_v when the reply is validated. D_1 is selected so as to remove fixed timing errors. *b*) Timing sequence in the tracking mode. The range count R_2 is started at t_0 and stopped after the first bit of the marker pulse (shown black). The range-tracking gate defines the interval shown by the dashed box; it opens at a time equivalent to the previous stored range R_s and closes 3 μs later. A fixed count D_2 starts at the end of the range count R_2 and defines the moment t_v when the range store may be updated following validation. During D_2 the count R_2 is reduced by 32 clock bits (about 2 μs) in order to generate an updated value of R_s.

At the comparator the leading edge of the marker pulse is compared with a threshold level. The moment at which the edge crosses the threshold is accepted as the arrival time of the pulse. The time delay from the start of the pulse to the threshold-crossing point is a function of the signal strength (see *fig. 13a*), and, consequently, of the range. This may also be visualized in terms of a fixed pulse shape which is intersected by the threshold at a level proportional to the range of the signal (fig. 13*b*).

Obviously the range errors introduced by the different time delays depend on the shape of the pulse edge, and on its rise time as determined by the filter characteristics in the receiver. The rise time t_r of a filter is defined as the time delay between the 10% and 90% levels in response to a unit step (fig. 13*b*).

Weak signals cross the threshold in the part of the rise-time characteristic which is approximately linear and has a slope defined by t_r. For strong signals the threshold is located below the 10% level so that it becomes necessary to consider the form of the curvilinear foot of the characteristic. For both the logarithmic and linear amplifiers it has been found that the logarithm of the delay is proportional to the signal strength in decibels. This relationship holds over a range of signal strengths extending from the weakest signal capable of being validated to a level about 55 dB higher.

interrogation; the recovery time constant is equivalent to a range of 3 km.)

Propagation errors and thermal noise

Signals are normally subject to fading caused by various factors such as meteorological conditions, ground reflections and aircraft movements like pitch and bank. Certain conditions may also produce reinforcement of the signal. Since fig. 14 is plotted with a logarithmic range scale the resultant changes of the range reading δr can be read off the curve by noting

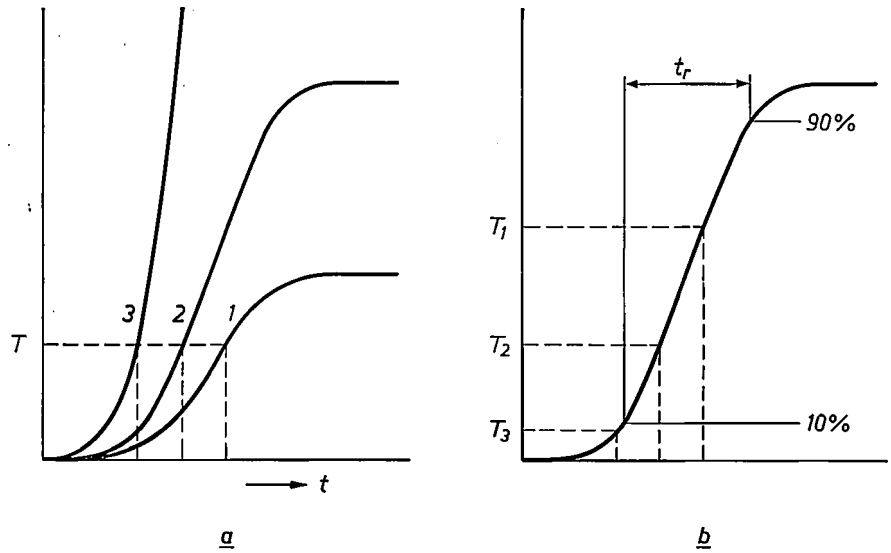


Fig. 13. Edge-detector timing errors. *a*) The leading edges of pulses *1*, *2*, *3* of different signal strength cross the threshold level T at different points in time. This causes the timing of the pulses, and hence the measured range values, to show errors which vary with signal strength, i.e. with range. *b*) Alternative representation. The leading edge of a pulse of fixed shape intersects different threshold levels T_1, T_2, T_3 corresponding to different ranges. t_r rise time of the pulse.

A relationship between the range r and the range error δr can be obtained from empirical data:

$$\delta r = D(r/R)^{\alpha},$$

where

$$D = (ct_r/4)/(1 - 5^{-\alpha}).$$

Here D is defined as the range error for a signal that crosses the threshold at half the pulse height and R is the range for this condition; it is derived from typical power-budget figures. c is the velocity of light and α is a constant which depends on the filter characteristics; it has a value of 0.39 in the interrogator and 0.425 in the transponder.

The total range-measuring error is the sum of the contributions of the down-link and the up-link. *Fig. 14* shows the composite error curve. The error figures have been normalized for calibration adjustment at a range of 100 metres. (The up-link contribution includes the effect of a swept threshold control in the airborne receiver. This reduces the sensitivity at the timing edge in the period immediately following the end of each

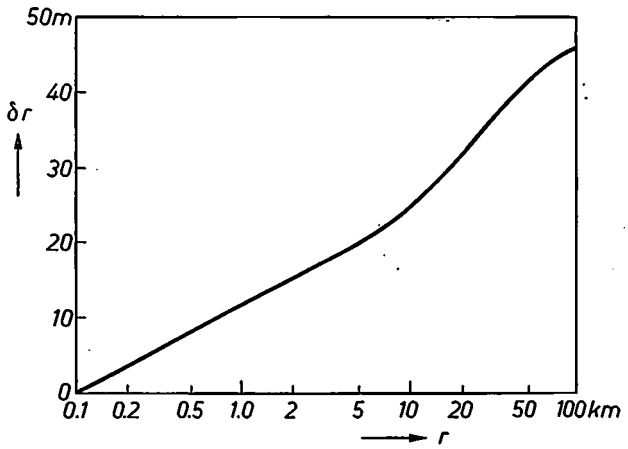


Fig. 14. Graph of predicted range error δr as a function of range r. As the range increases the signal grows weaker and crosses the detection threshold at a later point in time, thus increasing the range error. The range error is arbitrarily set to zero at a range of 100 m. In calculating the curve a rise time of 310 ns has been assumed for the ground receiver, and 460 ns for the air receiver.

that a 6-dB fade is equivalent to doubling the range. For the purpose of specifying accuracy it is normal to assume that the system must handle fades of up to 10 dB in normal flight.

Pulse distortion due to multipath reflections is highly variable in nature and may cause periodic misses, or, in relatively rare circumstances, extended dropouts due to message corruption. As a source of range-measuring errors, however, such distortion is less serious since the detection threshold for acceptable signals is always less than 60% of the pulse amplitude. This means that only the lower portion of the leading edge is used for timing valid signals and this is practically immune from severe distortion.

However, the rate at which the edge rises is affected by multipath components arriving with a delay which is very small compared with t_r [3]. These components, which we shall call 'zero-delay echoes' are in fact the principal source of range errors arising from multipath transmission. They may either reinforce or attenuate the leading-edge build-up by a factor which depends on the relative strength of echo and signal. In typical circumstances relative motion and random variations at reflecting surfaces cause the echoes to vary in phase between one interrogation and the next. In this case range noise is generated at the validation frequency. For zero-delay echoes that are 6 dB lower than the signal the resultant noise (before averaging) is equivalent to signal fluctuations having an r.m.s. value of about 3 dB. The mean error or bias in these circumstances is a reinforcement effect of 1 dB, equivalent to a small decrease of range. The resulting range error may be read from fig. 14.

When a reflecting object is much closer to the flight path (or its extension) than 1/20 of the range, the phase of the echo may be relatively static. In this case a zero-delay echo at a level of —6 dB could generate a bias error corresponding to signal-strength limits of +3.5 and —6 dB.

Range noise is also generated by random or thermal noise in the receiver circuits. The magnitude of the effect is governed by the noise-voltage to threshold-voltage ratio at the video comparator, and this in turn depends on the controlled false-alarm rate. In our transposed model of the edge-detection process (fig. 13*b*) the thermal noise produces fluctuations of the height of the threshold-crossing point above the base-line of a standard pulse. The corresponding changes in the nominal value of the range can be introduced on the abscissa of the error curve in fig. 14 to find the fluctuations of the range error. The noise-voltage to threshold-voltage ratio is 0.43 when the probability of a false alarm has its nominal value of 1% as determined by the CFAR system, so that thermal noise generates range errors corresponding to signal-strength fluctuations of about 4 dB at the validation frequency.

Table II. Summary of range errors. Totals represent 95% probability limits over operating-temperature range.

Range (km)	Digital (metres)			Analog (metres)		
	bias	noise	total	bias	noise	total
0.1	4	4	6	15	6	16
1	13	7	15	20	8	22
10	25	12	28	44	26	52
30	37	15	40	100	63	120
50	42	16	50	160	102	200

Additional range errors

Apart from the errors due to propagation and thermal noise, bias errors and digitization noise are generated by the timing circuits at both ends of the link. Minor additions are the calibration error of the range-clock oscillator and a bias error arising from cable delays depending on the arrangement of the equipment. Digitization noise is a jitter effect due to the timing uncertainty (± 31 ns) in the synchronization process at the receivers; in the two-way path it has a total value of 3.8 metres r.m.s. before averaging.

The bias errors have an unpredictable, random value. They may be characterized by their 'two-sigma value'; if a normal distribution is assumed, 95 per cent of the values lie between the limits -2σ and $+2\sigma$, where σ is the standard deviation of the distribution. The two-sigma value of the error in the timing circuits is estimated as 5 metres, including a drift component of 1.5 metres.

The analog range output is subject to additional zero and calibration errors plus a velocity-dependent noise component due to nonlinearity of the digital-to-analog converter. The two-sigma value of these errors does not exceed 20 metres plus 0.5 m/km at all operating temperatures, with an increment of 30 metres at ranges above 2 n miles (3.7 km).

Two-sigma values of the total range error resulting from the various components are summarized in *Table II*. The values are based on long-term sampling with ± 3 dB of multipath noise and a 20% probability of 10-dB fades.

Velocity measurement

Differentiation of the analog range signal with respect to time is used to obtain the velocity of the aircraft. For a satisfactory accuracy it is necessary to have close tolerances even in the least-significant bit values of the range digital-to-analog converter. This is because inaccuracies in the bit values generate velocity noise

[3] See also C. J. Hirsch, A TACAN-compatible accurate DME for short ranges, IEEE Trans. ANE-12, 68-76, 1965.

which is proportional to the fractional errors and independent of the weights of the bits. Bits of a lesser weight change at a proportionally higher rate so that the time derivative is the same for bits of different weights.

A simplified circuit of the range differentiator is shown in *fig. 15*. This is basically a conventional analog differentiator with preamplification. There is a digitally controlled back-off system, which subtracts from the analog range voltage a large fraction derived from the five most-significant range bits (of weight from 1 n mile to 16 n miles). By subtracting this slowly varying fraction which is of no use for the determination of the velocity, a more efficient use is made of the linear voltage range of the preamplifier and more preamplification can be applied.

if C is 1 μF the charging current corresponding to 1 knot is 306 pA when no preamplification is used. Operational amplifiers are available for measuring currents as low as this but the limiting factor is the leakage (over the working temperature range) of available capacitors of suitable size, even with the best dielectric available. The leakage current introduces a range-dependent velocity error. The automatic back-off system divides such errors by the voltage gain of the preamplifier; it multiplies the charging current I by the same factor, so that the leakage and drift requirements become much less stringent.

The noise in the velocity output is about 2.5 knots r.m.s. at 5 n miles and 7.5 knots at 15 n miles. In order to supplement the 150-ms time constant in the range analog amplifier, the velocity circuit incorporates a

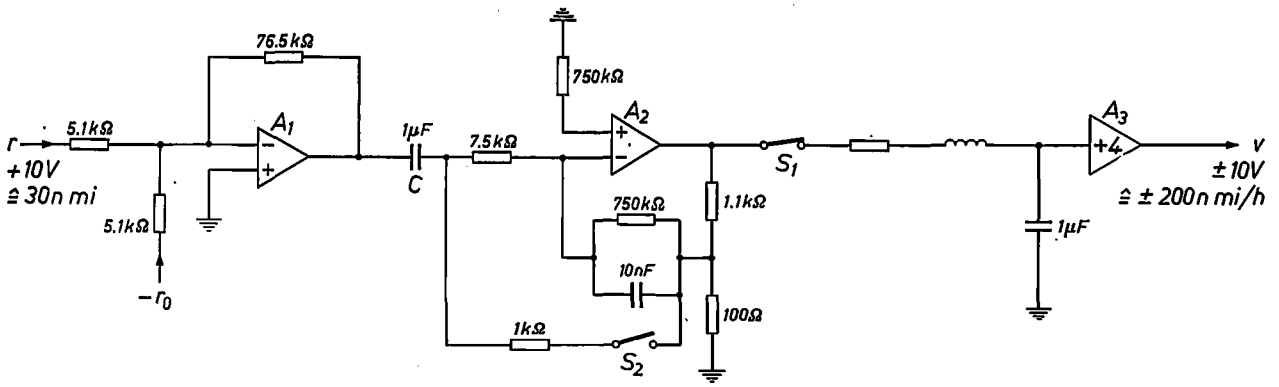


Fig. 15. Simplified schematic diagram of velocity derivation. Velocity is derived by analog differentiation of the range output r in the differentiator A_2 with a time constant of 750 ms. The value of the five most significant range bits is converted into analog form (r_0) and subtracted from the range r in the amplifier A_1. The voltage r_0 is derived by way of a digital backlash circuit so as to avoid unwanted changes due to random fluctuations of the least-significant range digit. r_0 changes step-wise by amounts equivalent to 1 n mile; at each step switch S_1 opens and S_2 closes for 50 ms to allow time for capacitor C to charge to the new level. The back-off system reduces the signal dynamic range so that sufficient preamplification may be used to minimize errors from leakage in the differentiating capacitor C. The amplifier A_3 is part of a lowpass filter, shown in full detail in fig. 16.

The back-off digits are coupled to the back-off voltage by way of a digital backlash system. This is necessary because random fluctuations of the least-significant range bit would otherwise generate instability of the voltage, due to carries, when the total weight of the bits that are suppressed in generating the back-off voltage approaches 1 n mile.

The necessity for preamplification may be explained in the following way. Analog differentiation of a voltage V requires the measurement of a charging current I such that $I = C\mathrm{d}V/\mathrm{d}t$, where C is the differentiating capacitance. The full-scale value of the velocity output is ± 200 knots (1 knot = 1 n mile/hour). The voltage available from the range-measuring system has a full-scale value of 10 V (for 30 n miles) and at 1 knot it is changing at a rate of about 93 μV/s. Space limitations make it necessary to restrict the size of the capacitor C; second-order lowpass filter. The digital noise translates into voltage steps; it is therefore desirable that the filter should have a fast settling time without overshoot. The Bessel transfer function provides these features with the added advantage of a constant time delay in the passband. An active Bessel-filter circuit is employed, which is shown in *fig. 16*.

Future applications and modifications

The MADGE data link has built-in flexibility. In its present form two different functions (guidance and two-way digital communication) are provided by time-division multiplexing. The principle could be extended. For instance, a long-range navigation facility can be provided simultaneously with the short-range guidance

role by associating the transponder with an additional rotating antenna; in this case azimuth data for navigation purposes is provided by a shaft encoder.

In the basic system as described the flight path of each aircraft is under the control of the pilot. In fact the logic hardware of the data link already provides for an alternative guidance mode in which flight-path deviation data is determined by a computer at the ground station. In this mode of operation it is possible to control complex traffic patterns, with individual nonlinear tracks allocated to each aircraft that enters the system; the flight paths could be continuously modified to suit the total traffic situation at any time. This ground-controlled mode retains the procedural advantages of random access from air to ground and it is again possible to provide simultaneous control for as many as 200 aircraft.

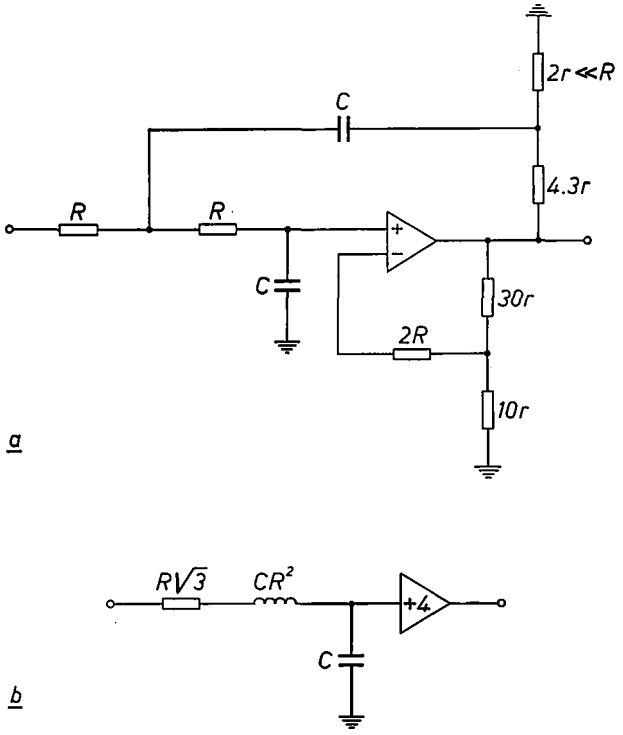


Fig. 16. Second-order Bessel lowpass filter [4]. This filter reduces digital noise in the velocity output and has a fast settling time without overshoot. Only two component values (R and C) are used to fix the frequency-scale factor. With $R = 220$ kΩ and $C = 1$ μF the half-power cut-off frequency is 0.57 Hz. *a*) Complete circuit diagram, *b*) equivalent circuit.

[4] R. P. Sallen and E. L. Key, IRE Trans. CT-2, 74, 1955.

Appendix I. Traffic analysis

Symbols:

- A number of aircraft per ground transponder
- F interrogation rate per aircraft (50 per second).
- f rate of interfering interrogations
- G number of transponders using the same frequency pair
- h ground dispersion factor, $0 \leqslant h \leqslant 1$
- J width of A-channel interrogation word (24.7 μs)
- k fraction
- U transponded replies per interrogation
- W uncorrupted replies received per interrogation
- Y width of A-channel reply word (59.3 μs)

Ground response efficiency

Two radio frequencies are assumed: one for all down-links and another for all up-links.

An interrogation arriving at the transponder will generate no ground reply either when the transponder is replying to another aircraft or when the transponder is processing an interrogation from another aircraft. The fraction of interrogations whose leading edges are overlapped by the reply dead time is

$$k_1 = (A-1)UFY. \quad (1)$$

The interfering interrogation rate is

$$f = (AG-1)F. \quad (2)$$

The transponder is organized so that any interrogation can generate a processing dead time J unless it is completely masked by a reply dead time Y; the proportion thus masked is $(A-1)UF(Y-J)$. Hence the rate at which interrogations are processed is

$$f_1 = f\{1-(A-1)UF(Y-J)\}.$$

The fraction of interrogations overlapped by earlier interrogations is therefore

$$k_2 = f_1 J = fJ\{1-(A-1)UF(Y-J)\}. \quad (3)$$

The probability p_1 that the leading edge of an interrogation is free for processing is then

$$p_1 = 1-k_1-k_2. \quad (4)$$

The probability p_2 that an interrogation is not overlapped by a later interrogation is approximated by Poisson statistics, because the mean number of signals during J is small compared with a possible maximum of $(AG-1)$. Whence

$$p_2 = \exp(-fJ). \quad (5)$$

Since $U = p_1 p_2$ it follows that

$$p_1 = U\exp(fJ). \quad (6)$$

Combining eqs. (4) and (6) gives:

$$U = (1-fJ)/\{\exp(fJ) + (A-1)F(Y-fJY+fJ^2)\}; \quad (7)$$

where $f = (AG-1)F$.

Air reply efficiency

Other ground stations may produce reply crosstalk, in which case the reply efficiency is $W < U$; W is the system-traffic component of the reply efficiency V in the aircraft.

The interfering reply rate is

$$f_2 = A(G-1)UF. \quad (8)$$

When there is only one ground station ($G = 1$), then $f_2 = 0$ and $W = U$. In the general case when $G > 1$ reply crosstalk at the aircraft receiver is partially inhibited by overlapping of the corresponding interrogations at an interfering transponder. The vulnerable periods, which precede and follow the leading edge of a reply in the aircraft, each have the value

$$Y_2 = Y-J+hJ, \quad (9)$$

where the ground dispersion factor h is zero when the inter-

fering transponder is less than 4.8 km from the home transponder, and reaches a value of unity ($Y_2 = Y$) when the spacing is 11.8 km or greater. Applying Poisson statistics, as in eq. (5), gives

$$W = U \exp(-f_2 \,.\, 2Y_2).$$

Substituting from eqs. (8) and (9) we get:

$$W = U \exp\{-A(G-1)UF \,.\, 2(Y - J + hJ)\}. \quad (10)$$

A graph of W versus GA is presented in fig. 7.

Appendix II. Validation and bit-error probability

Symbols:

V probability of accepting an incoming message
q probability of error for each bit checked

The probability of validating m invariant bits (such as address bits) is $(1-q)^m$ where m is defined as the *confidence index* for the set of bits.

Single-coordinate parity groups

Let n be the number of bits in a parity group (including one parity bit); $p = 1-q$; $s = p-q = 1-2q$. The probability that r bits will be in error in the group of n bits is

$$\frac{n!}{r!(n-r)!} q^r p^{n-r}.$$

Only odd numbers of errors break the parity code and are therefore detected; these correspond to the negative terms in the expansion of $(p-q)^n$. Since $(p+q)^n = 1$ it may be deduced that for the group of n bits the probability of acceptance is

$$A(n) = \tfrac{1}{2} + \tfrac{1}{2}s^n, \quad (11)$$

and the probability of rejection is

$$R(n) = \tfrac{1}{2} - \tfrac{1}{2}s^n. \quad (12)$$

The probability that one particular bit is in error in any word accepted is given by:

$$Q(n,1) = q \Big/ \left(p \frac{1+s^{n-1}}{1-s^{n-1}} + q\right). \quad (13)$$

From eq. (13) it follows that the bit-error improvement factor due to the parity check is about $1/(n-1)q$ when $nq \ll 1$.

The confidence index for the n-bit group is a function ϕ such that $(1-q)^\phi = A(n)$, see eq. (11). When $nq < 1/3$ this gives the approximate relation

$$\phi = n - \tfrac{1}{2}n(n-1)q. \quad (14)$$

Validation of complete message

When parity checking is applied along x columns and y rows (as shown in fig. 9) the error rate in the matrix of cross-linked rows and columns is very small, so that the doubly checked bits may be taken as a set of xy invariant bits. The total confidence index for the message is defined as

$$M = a - bq = m + xy + \Sigma\phi, \quad (15)$$

where the ϕ are obtained from eq. (14) using, in each case, a value of n which excludes any doubly checked bits in the group. a and b are constants which characterize the validation conditions. Hence the probability of validating a complete message is

$$V = (1-q)^{a-bq}. \quad (16)$$

Bit-error probability as a function of V

The inverse function of eq. (16) is obtained by an approximation which is valid when $q < 10\%$. If $M = a$, $q \approx 1/\{a/\ln(1/V) + \tfrac{1}{2}\} = j$, say; then for $M = a - bq$ we obtain

$$q = 2j/\{1 + \sqrt{1 - 4bj^2/\ln(1/V)}\}. \quad (17)$$

For azimuth data $M = 53 - 22q$; if $V = 16\%$, eq. (17) gives $q = 3.45\%$.

Appendix III. Validation and signal-to-noise ratio

Symbols:

N r.m.s. value of noise voltage
S r.m.s. value of signal voltage
T threshold voltage

Under weak-signal conditions the probability q_0 of an error on a true logic-zero is determined by the controlled-false-alarm rate f and is

$$q_0 = f/B, \quad (18)$$

where B is the bandwidth defined by the intermediate-frequency amplifier and post-detector filter. The probability q_1 of an error on a true logic-one depends on the signal amplitude. The mean bit-error probability is taken as

$$q = \tfrac{1}{2}(q_0 + q_1) \quad (19)$$

on the assumption that zeros and ones are equally distributed in the message. We define $x_0 = T/N$ and $x_1 = (S-T)/N$. Assuming Gaussian statistics we obtain the following relations for determining q in conjunction with eqs. (18) and (19):

$$x_0 = G^{-1}(q_0),$$
$$q_1 = G(x_1) = G(S/N - x_0),$$

where $G(x_1) = \tfrac{1}{2} - \tfrac{1}{2}\,\mathrm{Erf}\,(x_1/\sqrt{2})$ and $G^{-1}(q_0)$ is the inverse Gaussian function, so that $G(x_0) = q_0$.

Threshold-raising

When the signal voltage S reaches a critical value S_1 the control system raises the threshold, thus reducing q_0. In the general case the bit-error probabilities are given by the following relations:

$$x_0 = G^{-1}(f/B), \quad (20)$$
$$q_0 = G\{x_0 + \tfrac{1}{2}(S - S_1)/N\}, \quad (21)$$
$$q_1 = G\{S/N - x_0 - \tfrac{1}{2}(S - S_1)/N\}, \quad (22)$$

where the terms in $S - S_1$ are taken as zero when $S < S_1$.

If the bit-error probabilities are derived in this way as a function of the signal-to-noise ratio, the probability V of validation can be calculated from (16). V is shown as a function of S/N in fig. 10.

Appendix IV. Undetected errors and data noise

One-coordinate parity

Eq. (13) in Appendix II gives the probability of accepting an error in any particular bit of a parity group checked along one coordinate. In a group containing $(k-1)$ data bits and one parity bit the noise variance is

$$Q(k,1)\sum_{i=1}^{k-1} w_i^2, \quad (23)$$

where the w_i represent the weights of the data bits. Owing to saturation effects in analog processing circuits a parity group may contain a sub-group of data bits which are not additive. A generalized form of eq. (13) gives the probability of accepting one or more errors per validation within a sub-group r where $1 < r < k-1$:

$$Q(k,r) = 1 - p^r(1 + s^{k-r})/(1 + s^k). \quad (24)$$

For r bits each with a non-additive weight of w the noise-variance contribution is

$$Q(k,r)w^2. \quad (25)$$

Two-coordinate parity

We now consider the undetected errors in an array of n-bit parity groups in which z^2 data bits are checked along two coordinates and $z < n-1$. The square matrix z^2 is bounded by $2z$ partial parity groups of $n-z$ bits each. The probability that

there will be no errors in the matrix z^2 is

$$p^{z^2}A^{2z}(n-z),$$

where $p = 1-q$ and $A(n-z)$ is the probability of accepting any one of the partial parity groups; see eq. (11) in Appendix II.

The probability of accepting exactly one error in the matrix is derived from the fact that there are z^2 equiprobable positions for an error and in each case two of the partial parity groups must contain an odd number of errors, which would normally result in rejection. The probability is

$$z^2p^{z^2-1}qR^2(n-z)A^{2z-2}(n-z),$$

where $R(n-z)$ is the rejection probability given by eq. (12) in Appendix II.

We now define a function $E_1(n,z^2)$ as the probability of accepting exactly one error in the block z^2 expressed as a fraction of the probability of blocks devoid of errors; n is the number of bits in each of the cross-linked parity groups. By division we get

$$E_1(n,z^2) = z^2t^2u, \tag{26}$$

where $t = R(n-z)/A(n-z)$ and $u = q/p$. The equivalent function for exactly two errors involves two classes of error pattern. Proceeding as before we get

$$E_2(n,z^2) = z^2(z-1)t^2\{1+\tfrac{1}{2}(z-1)t^2\}u^2. \tag{27}$$

The probability of accepting three errors or more is normally small enough to be neglected.

The probability of accepting one or more errors per validation in the matrix is

$$\sum_{r=1}^{z^2} E_r \Big/ (1+\sum_{r=1}^{z^2} E_r).$$

In practice $E_1 \ll 1$ and we may take the probability as $E_1 + E_2$. With $n = 6$ and $z = 3$ we have $E_1 + E_2 < 4\times10^{-3}$ when $q = 3.45\%$. Eqs. (23) and (25) give expressions for the noise contributions of the partial parity groups when k is taken as $(n-z)$. For the square matrix the noise-variance contribution is

$$E_1\sigma_1^2 + E_2\sigma_2^2, \tag{28}$$

where σ_1 is the r.m.s. value of the bit weights and σ_2 is the r.m.s. value of all two-bit combinations, with due allowance for saturation effects.

Summary. This second article about the MADGE aircraft-landing aid describes the digital data link between the landing site and the aircraft. The link is an air-initiated random-access system. It passes horizontal and vertical angular guidance information from ground-based interferometers to the aircraft. The airborne terminal converts this information into a form suitable for aircraft instrumentation; it also derives the slant distance to the landing site from the message go-and-return time and transmits this distance to the ground terminal. The ground terminal can output the aircraft's identity and three-dimensional position to a remote display.

Up to 200 aircraft may be served by one ground terminal. Messages are protected by address codes and two-coordinate parity. The airborne processing circuits include guidance rate limiters, 'beam-softening' and distance and velocity measurement.

Recent scientific publications

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

Philips Research Laboratories, Eindhoven, The Netherlands *E*
Mullard Research Laboratories, Redhill, Surrey, England *M*
Laboratoires d'Electronique et de Physique Appliquée, 3 avenue Descartes, 94450 Limeil-Brévannes, France *L*
Philips GmbH Forschungslaboratorium Aachen, Weißhausstraße, 51 Aachen, Germany *A*
Philips GmbH Forschungslaboratorium Hamburg, Vogt-Kölln-Straße 30, 2000 Hamburg 54, Germany *H*
MBLE Laboratoire de Recherches, 2 avenue Van Becelaere, 1170 Brussels (Boitsfort), Belgium *B*
Philips Laboratories, 345 Scarborough Road, Briarcliff Manor, N.Y. 10510, U.S.A. (by contract with the North American Philips Corp.) *N*

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, The Netherlands.

P. B. van Aken, H. J. van Daal & K. H. J. Buschow: Kondo sideband effects in the Seebeck coefficient of $Ce_{1-x}La_xAl_3$ compounds.
Physics Letters **49A**, 201-203, 1974 (No. 3). *E*

E. Arnold: Disorder-induced carrier localization in silicon surface inversion layers.
Appl. Phys. Letters **25**, 705-707, 1974 (No. 12). *N*

C. Belouet, E. Dunia (Université de Paris VI) **& J. F. Pétroff** (Univ. de Paris VI): X-ray topographic study of defects in KH_2PO_4 single crystals and their relation with impurity segregation.
J. Crystal Growth **23**, 243-252, 1974 (No. 4). *L*

F. Berz: Diffusion near an absorbing boundary.
Solid-State Electronics **17**, 1245-1255, 1974 (No. 12). *M*

P. Blood, G. Dearnaley (AERE, Harwell) **& M. A. Wilkins** (AERE, Harwell): The origin of non-Gaussian profiles in phosphorus-implanted silicon.
J. appl. Phys. **45**, 5123-5128, 1974 (No. 12). *M*

A. J. van Bommel & J. E. Crombeen: LEED, Auger electron spectroscopy (AES) and photoemission studies of the cesium covered GaAs(110) surface.
Surface Sci. **45**, 308-313, 1974 (No. 1). *E*

H. van den Boom & J. H. Haanstra: Raman spectra of vibrational modes in spinel $CdIn_2S_4$.
J. Raman Spectr. **2**, 265-274, 1974 (No. 3). *E*

J. van den Boomgaard, D. R. Terrell, R. A. J. Born & H. F. J. I. Giller: An *in situ* grown eutectic magneto-electric composite material, Part 1. Composition and unidirectional solidification.
J. Mat. Sci. **9**, 1705-1709, 1974 (No. 10). *E*

D. J. Breed: Non-ionic room temperature instabilities in MOS devices.
Solid-State Electronics **17**, 1229-1243, 1974 (No. 12). *E*

C. H. J. van den Brekel: Growth rate anisotropy and morphology of autoepitaxial silicon films from $SiCl_4$.
J. Crystal Growth **23**, 259-266, 1974 (No. 4). *E*

J. C. Brice & T. M. Bruton: The stability of facets on growing crystals.
J. Crystal Growth **26**, 59-60, 1974 (No. 1). *M*

J. C. Brice, T. M. Bruton, O. F. Hill & P. A. C. Whiffin: The Czochralski growth of $Bi_{12}SiO_{20}$ crystals.
J. Crystal Growth **24/25**, 429-431, 1974. *M*

J. Brokken-Zijp: Isomerisation and decomposition of alkylaryldiazosulphides by light.
Tetrahedron Letters 1974, 2673-2676 (No. 31). *E*

M. Brouha, K. H. J. Buschow & A. R. Miedema: Magneto-volume effects in rare-earth transition metal intermetallics.
IEEE Trans. **MAG-10**, 182-185, 1974 (No. 2). *E*

K. H. J. Buschow, B. van Laar (Reactor Centrum Nederland, Petten) **& J. B. A. A. Elemans** (Kamerlingh Onnes Laboratorium, Leiden): Magnetic properties of NdZn.
J. Physics F **4**, 1517-1521, 1974 (No. 9). *E*

C. Canali, J. W. Mayer, G. Ottaviani (all with Institute of Physics, Modena, Italy), **D. Sigurd** (California Institute of Technology, Pasadena) **& W. van der Weg** (Philips Research Labs., Amsterdam Division): Solid-phase transport and epitaxial growth of Ge and Si.
Appl. Phys. Letters **25**, 3-5, 1974 (No. 1).

F. M. A. Carpay & W. A. Cense: In situ growth of composites from the vitreous state.
J. Crystal Growth **24/25**, 551-554, 1974. *E*

H. J. van Daal, P. B. van Aken & K. H. J. Buschow: The Seebeck coefficient of $YbAl_2$ and $YbAl_3$.
Physics Letters **49A**, 246-248, 1974 (No. 3). *E*

R. Davies: Basic design principles of frequency stable microwave oscillators.
Comm. Int. **2**, No. 1, pp. 32, 34 & 36, Jan. 1975. *M*

A. M. van Diepen & F. K. Lotgering: Mössbauer effect in $LaFe_{12}O_{19}$.
J. Phys. Chem. Solids 35, 1641-1643, 1974 (No. 12). *E*

H. Durand: Het gebruik van fotovoltaïsche energie-omzetting. Een technologische en economische uitdaging.
Ingenieur **87**, 9-12, 1975 (No. 1). *L*

D. den Engelsen & B. de Koning: Ellipsometry of spread monolayers, Part 2. Coloured systems: chlorophyll a, carotenoic acid, rhodamine 6G and a cynanine dye.
J. Chem. Soc., Faraday Trans. I **70**, 2100-2112, 1974 (No. 11). *E*

G. Eschard, R. Polaert & E. Roaux: Comparison between two methods of SNR measurement of image tubes.
Appl. Optics **13**, 2187-2189, 1974 (No. 10). *L*

E. D. Fletcher: An engineer's guide to holography.
Electronic Engng. **46**, Nov. 1974, 53-55 & 57 (No. 561). *M*

S. R. Fletcher (Imperial College, London), **E. T. Keve & A. C. Skapski** (Imperial College, London): The crystal structure of triglycine sulphate at low X-ray dosage and after irradiation/field treatment.
Ferroelectrics **8**, 479-482, 1974 (No. 1/2). *M*

C. T. Foxon, B. A. Joyce, R. F. C. Farrow (Royal Radar Establishment, Great Malvern) **& R. M. Griffiths** (SERL, Baldock): The identification of species evolved in the evaporation of III-V compounds.
J. Physics D **7**, 2422-2435, 1974 (No. 17). *M*

G. Frank & S. Garbe: GaP-$Al_xGa_{1-x}As$-GaAs heterostructures with improved transmission photoemission.
Phys. Stat. sol. (a) **26**, K 91-93, 1974 (No. 2). *A*

G. Frens: Aerodynamic drag on bursting bubbles.
J. phys. Chem. **78**, 1949-1953, 1974 (No. 19). *E*

Y. Genin: An algebraic approach to *A*-stable linear multistep-multiderivative integration formulas.
BIT **14**, 382-406, 1974 (No. 4). *B*

Y. Genin: Hurwitz sequences of polynomials.
Philips Res. Repts. **30**, 89-102, 1975 (No. 2/3). *B*

J. J. Goedbloed & H. Tjassens: Parasitic oscillations in IMPATT-diode oscillators.
Proc. 4th European Microwave Conf., Montreux 1974, pp. 328-332. *E*

J. J. Goedbloed & M. T. Vlaardingerbroek: Noise in IMPATT-diode oscillators at large-signal levels.
IEEE Trans. **ED-21**, 342-351, 1974 (No. 6). *E*

J. M. Goethals: Two dual families of nonlinear binary codes.
Electronics Letters **10**, 471-472, 1974 (No. 23). *B*

H. C. de Graaff & R. J. van der Wal: Measurement of the onset of quasi-saturation in bipolar transistors.
Solid-State Electronics **17**, 1187-1192, 1974 (No. 11). *E*

G. C. Groenendaal & R. R. Wilson (NKF Kabel B.V., Delft): Mathematical and numerical analysis of coaxial cable.
Philips Res. Repts. **30**, 162-176, 1975 (No. 2/3). *E*

J. C. M. Henning, J. H. den Boef, R. P. van Stapele & D. Polder: Theory of flexural-strain-modulated electron spin resonance (SMESR).
Solid State Comm. **15**, 1535-1538, 1974 (No. 9). *E*

D. Hennings & H. Pomplun: Evaluation of lattice site and valence of Mn and Fe in polycrystalline $PbTiO_3$ by electron spin resonance and thermogravimetry.
J. Amer. Ceramic Soc. **57**, 527-530, 1974 (No. 12), & **58**, 152, 1975 (No. 3/4) *(reply to discussion)*. *A*

J. H. C. van Heuven: A new integrated waveguide-microstrip transition.
Proc. 4th European Microwave Conf., Montreux 1974, pp. 541-545. *E*

J. H. C. van Heuven: Conduction and radiation losses in microstrip.
IEEE Trans. **MTT-22**, 841-844, 1974 (No. 9). *E*

W. H. de Jeu, W. J. A. Goossens & P. Bordewijk (University of Leiden): Influence of smectic order on the static dielectric permittivity of liquid crystals.
J. chem. Phys. **61**, 1985-1989, 1974 (No. 5). *E*

W. H. de Jeu & Th. W. Lathouwers: Dielectric properties of some nematic liquid crystals for dynamic scattering displays.
Chem. Phys. Letters **28**, 239-241, 1974 (No. 2). *E*

E. Klotz, U. Tiemens & H. Weiss: Display of the three-dimensional modulation transfer function of tomography and tomosynthesis X-ray systems.
Optics Comm. **12**, 282-286, 1974 (No. 3). *H*

M. Laguës & M. Chaumont (Laboratoire R. Bernas, Orsay): Photoémission de semiconducteurs à grand gap activés par ségrégation à la surface d'alcalins implantés.
C.R. 4ème Congrès Int. 'AVISEM 74', Toulouse 1974, pp. 285-298. *L*

J. Lemmrich: Frequenzanaloge Universalregler.
Regelungstechn. Praxis **16**, 249-254, 1974 (No. 10). *H*

F. K. Lotgering: Magnetic anisotropy and saturation of $LaFe_{12}O_{19}$ and some related compounds.
J. Phys. Chem. Solids **35**, 1633-1639, 1974 (No. 12). *E*

J. Magarshack, A. Rabier & R. Spitalnik: Optimum design of transferred-electron amplifier devices in GaAs.
IEEE Trans. **ED-21**, 652-654, 1974 (No. 10). *L*

R. Memming: The role of energy levels in charge transfer processes at semiconductor-electrolyte interfaces.
Proc. Symp. on Electrocatalysis, San Francisco 1974, pp. 178-194. *H*

R. F. Mitchell & D. W. Parker: Synthesis of acoustic-surface-wave filters using double electrodes.
Electronics Letters **10**, 512, 1974 (No. 24). *M*

B. J. Mulder: Topochemical preparation of silver sulphide crystals from cadmium sulphide.
J. Crystal Growth **23**, 214-216, 1974 (No. 3). *E*

A. van Oostrom: Surface analysis by FIM and AP.
CRC Crit. Rev. Solid State Sci. **4**, 353-369, 1974 (No. 3). *E*

J. A. Pals: Properties of Au, Pt, Pd and Rh levels in silicon measured with a constant capacitance technique.
Solid-State Electronics **17**, 1139-1145, 1974 (No. 11). *M/E*

D. E. Penna, D. W. Parker, F. W. Smith & R. Stevens: Interdigital acoustic-surface-wave transducer with low interactions.
Electronics Letters **10**, 489-490, 1974 (No. 23). *M*

D. Pohl (Philips Medical Systems Division, Eindhoven): Ein Verfahren zur Ermittlung der durch Rauschen begrenzten Detailwahrnehmbarkeit bei Bildverstärker-Fernsehanlagen.
Thesis, Aachen 1974. (Philips Res. Repts. Suppl. 1975, No. 5.)

Th. J. A. Popma & A. M. van Diepen: Magnetization and Mössbauer spectra of non-crystalline $Y_3Fe_5O_{12}$.
Mat. Res. Bull. **9**, 1119-1128, 1974 (No. 9). *E*

H. Rau: Vapour composition and critical constants of selenium.
J. chem. Thermodyn. **6**, 525-535, 1974 (No. 6). *A*

F. C. de Ronde & S. Shammas (Pahlavi University, Shiraz, Iran): MIC bandfilters using open-ring resonators.
Proc. 4th European Microwave Conf., Montreux 1974, pp. 531-535. *E*

F. Rondelez & A. Mircea-Roussel: Dielectric relaxation in the radio frequency range for the nematic phase of M.B.B.A.
Mol. Cryst. liq. Cryst. **28**, 173-178, 1974 (No. 1/2). *L*

T. E. Rozzi & F. C. de Ronde: The analytical approach to the full-band matching of waveguide discontinuities applied to capacitive irises and steps.
Proc. 4th European Microwave Conf., Montreux 1974, pp. 551-555. *E*

A. M. J. G. van Run, D. R. Terrell & J. H. Scholing: An *in situ* grown eutectic magnetoelectric composite material, Part 2. Physical properties.
J. Mat. Sci. **9**, 1710-1714, 1974 (No. 10). *E*

B. Schneider: Effect of crystalline anisotropy on the magnetostatic spin wave modes in ferromagnetic plates, II. Results for finite plates.
Phys. Stat. sol. (b) **66**, 99-106, 1974 (No. 1). *H*

J. M. Shannon & J. Lohstroh: JFET optical detectors in the charge storage mode.
IEEE Trans. **ED-21**, 720-728, 1974 (No. 11). *M, E*

F. W. Smith: Acoustic surface wave bandpass filters.
SERT J. **9**, 13-15, 1975 (No. 1). *M*

W. T. Stacy: Dislocations, facet regions and growth striations in garnet substrates and layers.
J. Crystal Growth **24/25**, 137-143, 1974. *E*

G. H. A. M. van der Steen & E. Papanikolau (Philips Glass Development Centre, Eindhoven): Introduction and removal of hydroxyl groups in vitreous silica, Part I. Influence of the melting conditions on the hydroxyl content in vitreous silica.
Philips Res. Repts. **30**, 103-119, 1975 (No. 2/3).

A. Thayse: New method for obtaining the optimal Taylor expansions of a Boolean function.
Electronics Letters **10**, 543-544, 1974 (No. 25/26). *B*

D. L. A. Tjaden & E. J. Tercic: Theoretical and experimental investigations of digital magnetic recording on thin media.
Philips Res. Repts. **30**, 120-161, 1975 (No. 2/3). *E*

W. Tolksdorf, G. Bartels, P. Holst & W. T. Stacy: Dependence of lattice parameter on composition in substituted yttrium iron garnet epitaxial layers.
J. Crystal Growth **26**, 122-126, 1974 (No. 1). *H, E*

H. van Tongeren & J. Heuvelmans: Positive column of Na-Ar and Na-Ne-Ar dc low-pressure discharges.
J. appl. Phys. **45**, 3844-3850, 1974 (No. 9). *E*

H. J. L. Trap: Glas.
Natuur en Techniek **42**, 94-107, 1974 (No. 2). *E*

A. Turos (California Institute of Technology, Pasadena), **W. F. van der Weg** (Philips Research Labs., Amsterdam Division), **D. Sigurd & J. W. Mayer** (both with Calif. Inst. Technol., Pasadena): Change of surface composition of SiO_2 layers during sputtering.
J. appl. Phys. **45**, 2777-2779, 1974 (No. 6).

J. D. B. Veldkamp: Short-fibre-reinforced materials.
Thesis, Twente 1975. (Philips Res. Repts. Suppl. 1975, No. 4.) *E*

A. T. Vink: The dependence of the radiative transition probability of donor-acceptor pairs on pair separation.
J. Luminescence **9**, 159-179, 1974 (No. 3). *E*

A. T. Vink, R. L. A. van der Heijden & A. C. van Amstel: The kinetics of donor-acceptor pair transitions with strong phonon coupling in GaP.
J. Luminescence **9**, 180-211, 1974 (No. 3). *E*

J. C. A. Vreeswijk (Philips Lighting Division, Plant Terneuzen), **R. G. Gossink & J. M. Stevels:** Nucleation kinetics and critical cooling rate of glass-forming liquids.
J. non-cryst. Solids **16**, 15-26, 1974 (No. 1). *E*

L. Vriens & M. Adriaansz: Light scattering from Ne, Ar, and Xe arcs.
J. appl. Phys. **45**, 4422-4431, 1974 (No. 10). *E*

F. W. Willmott & R. J. Dolphin: A novel combination of liquid chromatography and electron capture detection in the analysis of pesticides.
J. chromatogr. Sci. **12**, 695-700, 1974 (No. 11). *M*

M. Zimmer & W. Tolksdorf: Analysis of substituted yttrium iron garnet epilayers by an electron microprobe.
J. Crystal Growth **23**, 331-334, 1974 (No. 4). *H*

PHILIPS TECHNICAL REVIEW

VOLUME 35, 1975, No. 11/12

Light

The double issue that brings this year's volume to a close is given over to the theme of 'Light', or more accurately, to research on incandescent and discharge lamps. Subjects such as light emitting diodes (LEDs) are therefore not included — even though Philips are active in that field — nor are applications. Neither — but with one important exception — do we include anything on the subject of phosphors, since these have been extensively treated in recent years.

In the incandescent lamp the main emphasis is on the problem of efficacy, which in fact is physically identical with the problem of lamp life. Our knowledge of the processes that determine the life of present-day incandescent lamps is now fairly complete. Although, in relation to the prevention of these processes, a halogen lamp with a net mass loss of zero for each element of filament is still a distant goal, we have gained further insight into the ways in which that goal might be attained. Recently, considerations quite unconnected with efficiency have prompted investigations into whether the carbon-filament lamp might perhaps have more to offer than it did 70 years ago.

Although the efficacy is also important in discharge lamps — more than ever in recent years — the most urgent problems with these lamps are to be found elsewhere. While gas discharges have a much higher efficacy than incandescent lamps, the colour rendering is not so good, and the low-pressure discharge lamps have the disadvantage of a lower luminance. Solutions to the problem of colour rendering are being sought both by developing lamps that radiate a 'better' spectrum with reasonable efficacy, and also from investigations to determine the minimum number of spectral lines necessary for good colour rendering. This work has led to some very surprising results, and in the meantime a new class of fluorescent lamps has been based on them.

Almost everything we have to say about incandescent lamps in this special issue is contained in a single three-part article. Discharge lamps and related studies, however, form the subject of a number of separate articles, some more concerned with the physical principles, others with the technology. The first of these separate articles is intended as a survey of the field, and includes a table listing the types, characteristics and applications of modern discharge lamps.

The tower of Berne Cathedral, illuminated by conventional incandescent lamps (pressed glass, 150 W) and halogen lamps (quartz glass, 2000 W).

Research on incandescent lamps

I. Survey

Since Edison made the first incandescent lamp in 1879, with a carbon filament, the principle of the incandescent lamp has not changed very much. The principle is still to pass an electric current through a solid body so as to heat it to such a high temperature that at least some of its radiation is emitted in the visible part of the spectrum. The major improvements since then have been the replacement of carbon by tungsten, the use of an inert-gas filling — previously the bulb was evacuated — the coiling of the filament and finally the addition of halogens.

The great advantage of tungsten compared with other filament materials is that its vapour pressure is low, so that loss of the filament material by evaporation is slow. Other advantages are its selective emissivity and its great strength at elevated temperatures. Most subsequent improvements have also aimed at reducing the rate of evaporation. The inert gas does this by inhibiting the transport of tungsten. The principal feature of modern iodine and bromine lamps is that the bulb is made clear, which allows smaller bulbs to be used. This in turn makes it possible to use a higher inert-gas pressure, which further reduces the rate of tungsten evaporation.

The evaporation rate of the filament material has received so much attention because it is vitally important to the life and luminous efficacy of most lamps. The higher the temperature of the filament the more efficient is the energy conversion, but the greater the rate at which the filament evaporates, resulting in a shorter life. In a type of lamp whose life is completely determined by vapour transport of tungsten, the life varies in a way determined by the temperature dependence of the vapour pressure of tungsten. This implies, for example, that a 70% increase in efficacy, achieved by raising the filament temperature from 2800 to 3200 K, will shorten the life of such a lamp by a factor of about 100. An improvement of efficacy by raising the temperature is therefore only possible if ways of suppressing the adverse consequences of the temperature increase can be found. This calls for a thorough knowledge of the filament defects and transport mechanisms that govern the life of the lamp. In part II of this article a closer look will be taken at the transport problem in incandescent lamps and at the burn-out mechanism. If life is determined by the transport of tungsten in the vapour phase and by hot-spot development [1], a regenerative cycle based on chemical transport could prolong the life. The aim is then to prevent loss of mass from any part of the filament. This aim is not completely achieved in halogen lamps. Although the iodine or bromine cycle returns tungsten from the bulb to the filament, the hottest parts of the filament — where evaporation losses are greatest — do not in principle have any more tungsten returned to them than the cooler parts. In part III the possibilities of a truly regenerative cycle will be discussed. An example of prolonging the life of a lamp by processing the coiled filament to improve the temperature distribution along the coil is dealt with in a separate article in this issue [2].

So far we have only considered increasing the efficacy by raising the filament temperature while maintaining an acceptable life. Another way to increase efficacy is to improve the selective emissivity of the lamp. Efforts have been made in the past to find materials that have a better selective emissivity than tungsten, and a special study has been made of high-melting oxides. These selective radiators were not however a success, mainly because of two unwanted effects: thermal instability at high temperatures and the gradual decrease in selectivity with rising temperature [3]. The best way of improving efficacy by increasing the selectivity appears to be to use a filter that is transparent in the visible part of the spectrum but reflects the infrared radiation back to the filament. A filter of this type was successfully introduced some years ago to improve the efficacy of low-pressure sodium lamps [4]. In an incandescent lamp, however, the situation is rather more difficult because the desired part of the radiation spectrum is directly adjacent to the part to be reflected. This means that the optical properties of the filter must undergo a sharp change within a very narrow frequency range, which calls for a filter characteristic with steep sides [3].

Investigations into ways of improving the efficacy of incandescent lamps along the lines indicated above

---

[1] H. Hörster, E. Kauer and W. Lechner, Philips tech. Rev. 32, 155, 1971.
[2] J. Schröder, this issue, p. 332.
[3] R. Groth and E. Kauer, Z. angew. Physik 16, 130, 1963. See also E. Kauer, Philips tech. Rev. 26, 33, 1965.
[4] See, for example, H. J. J. van Boort and R. Groth, Philips tech. Rev. 29, 17, 1968.

form a prominent part of incandescent-lamp research at Philips. This does not mean, however, that the results of the work are only to be evaluated in terms of the improvement in efficacy obtained. A better understanding of the burn-out mechanism may also lead to a smaller spread in lamp life, which is also an improvement of quality. The process by which a better temperature distribution along the filament is obtained not only results in a longer life but also in a higher light output with a better distribution. Improvement of the efficacy of incandescent lamps is therefore not just an end but also an incentive.

II. Transport and burn-out in incandescent lamps

E. Fischer, J. Fitzgerald, W. Lechner and W. Lems

The only way of increasing the efficacy of incandescent lamps — except by improving their selective emissivity — is to increase the filament temperature. Such an increase in temperature is however always accompanied by a decrease in life. This means that increasing the efficacy is mainly a problem of extending the life. An understanding of the defects and transport mechanisms that determine life is therefore necessary to ensure that the correct measures are taken.

The oldest model used in explaining the finite life of filaments is the hot-spot model, which was introduced in the early days of lamp research. A more comprehensive version of this model has been developed in our laboratories, and this has been compared with experiments on straight wires in vacuum and in lamps filled with an inert gas [1]. The essential features of the hot-spot model will now be briefly recapitulated.

A hot spot is a small region of the filament where the temperature is higher because of a local inhomogeneity. The temperature difference ΔT between the spot and the rest of the filament increases with time, because tungsten evaporates from the hot spot at a faster rate. The final result of this process is that the lamp fails through localized melting or fracture of the filament. The essential features of the model are the existence of an initial inhomogeneity and tungsten transport by evaporation. (In the case of an inert-gas filling the transport rate is controlled by the rate of the tungsten diffusion through the inert gas.) The hot-spot model yields a relationship between the life of the lamp and the initial relative temperature difference $\Delta T/T$ of the hot spot.

Recent work on burn-out has supplemented the simplified model by taking quantities such as the spot length [2] and the heat conduction along the wire into account [3]. The causes of hot spots, and tungsten transport inside the coiled filament (such as turn-to-turn diffusion) are now being studied, and investigations are in progress of cases where there is reasonable ground for assuming that lamp life is determined by transport mechanisms other than evaporation. We shall make no attempt to present a balanced survey, let alone a complete account, of all that has been done in this field; that would be far beyond the scope of this article. We shall limit ourselves to a discussion of the causes of initial spots and of the tungsten transport through the inert gas. After a discussion of some recent observations on coiled filaments, with particular attention to turn-to-turn diffusion and to effects caused by the observed development of crystal planes ('facets'), we shall close with an attempt to draw a few conclusions about the nature of the mechanisms that determine lamp life and the operating conditions under which they become important.

Dr E. Fischer, J. Fitzgerald, W. Lechner and Dr W. Lems are with Philips GmbH Forschungslaboratorium Aachen, Aachen, West Germany.

Causes of hot spots

The inhomogeneities usually considered as causes of initial spots are constrictions in the wire and variation in electrical resistivity, in emissivity or in filament geometry. Since the initial temperature difference ΔT of a spot is typically of the order of a few degrees Celsius, an accurate determination of the initial variations is hardly possible.

From the observation that filaments that were given an extra smooth finish had a greatly increased 'deadly weight loss' — i.e. the weight loss of the filament at the end of life — H. Hörster *et al.*[1] have concluded that the life of the filaments they studied did not depend on a bulk property of the material. The correlation that they observed between the temperature distribution along the wire after switching on and the original variations (measured along the length) in wire diameter led them to the conclusion that constrictions were the main cause of spots. Photography of a coiled fila-

ment heated by a current pulse (flash testing) also indicated that constrictions are an important cause of initial spots [4]. Since constrictions not only increase the electrical resistance but also reduce heat capacity, flash testing is a sensitive method for the detection of variations in wire diameter.

However, the fact that wire constrictions have been shown to be the main cause of hot spots does not mean that other causes can be ruled out. Local variations in the pitch of the filament coil could be relevant, and inhomogeneous distribution of potassium-filled bubbles has recently been suggested as a possible source of local variation in the electrical resistivity. (Potassium is added to produce a wire with interlocking crystals, which does not sag.) It is not clear, however, whether such a distribution exists right from the beginning or whether the inhomogeneity develops through bubble migration and growth, which might occur in the presence of a temperature gradient [5]. The initial question here of course, is about the cause of the temperature gradient. A purely statistical fluctuation in bubble density cannot account for an effective local increase in resistance over a characteristic spot length, because the number of bubbles initially present in such a volume element is too large — of the order of 10^{10}, depending on the wire diameter.

Tungsten transport in an inert gas

The hot-spot model gives some impression of the processes that set limits to the life of incandescent lamps, but no information about the transport rate of the tungsten. To calculate this rate, a detailed investigation of transport processes in the filling gas is necessary. For straight incandescent wires the transport of tungsten from the filament to the wall is by diffusion and convection. In the case of coiled filaments, there is also transport inside the coil, and in certain circumstances this transport may determine the life. In this section we shall confine ourselves to the transport from the filament to the wall; the transport inside the coil will be discussed in the next section.

The treatment of the problem of transport from the filament to the wall is usually based on Langmuir's stationary-film hypothesis [6]. G. R. Fonda [7] was the first to tackle the problem, and W. Elenbaas [8] extended the theory. This model assumes that the filament is surrounded by a gas layer, with the tungsten transport by diffusion inside this layer and by free convection outside it. This assumption of free convection is however only a valid approximation for large lamps. If the effect of the lamp geometry on the transport is to be taken into account, then the temperature and velocity field must be accurately calculated. E. Fischer and J. Fitzgerald [9] have made such calculations for a cylindrical lamp operating horizontally, and have verified them experimentally. Some details of their results will be presented below. The more general case can be described qualitatively as follows.

The main parameter that determines the life of an incandescent filament is the temperature, which controls the vapour pressure p_v of tungsten at the surface of the wire. The rate of evaporation, and hence the mass loss, per unit time, is proportional to p_v. For two identical filaments — identical in geometry and the magnitude of the defects — the life will be inversely proportional to the vapour pressure of the tungsten:

$$\tau \propto p_v^{-1}.$$

If the effect of the inert-gas pressure p and the lamp geometry on the operation is to be considered, it is necessary to know whether the transport is mainly by diffusion or convection. If diffusion predominates, the flow rate of the tungsten is determined by the diffusion coefficient and the concentration gradient. The diffusion coefficient is proportional to p^{-1}, the concentration gradient to R^{-1}, the reciprocal of the distance between the filament and the envelope. Under these conditions the life τ is proportional to pR. In lamps with high values of R or a high pressure p, the transport is mainly by convection. In this case there is an additional transport, proportional to p^2R^3, so that for large R and p the life decreases with increasing R and p. It follows that for a given inert-gas pressure there is an optimum lamp diameter, which will give the maximum life.

This optimum lamp diameter has been determined by Fischer and Fitzgerald in their investigations of a horizontally operating incandescent lamp [9]. The first step was the determination of the temperature and flow fields of the inert-gas filling, both experimentally and theoretically, as mentioned above. The spatial temperature distribution was measured by holographic interferometry; the flow field was made visible by introducing very fine tracer particles of quartz glass (25 μm). *Fig. 1* shows a comparison between the calculated temperature fields, obtained from a numerical solution of

[1] H. Hörster, E. Kauer and W. Lechner, Philips tech. Rev. 32, 155, 1971, and J. Illum. Engng. Soc. 1, 309, 1972.
[2] C. Baker, J. Illum. Engng. Soc. 4, 163, 1975.
[3] B. A. Moys, GEC J. Sci. Technol. 40, 17, 1973.
F. J. Harvey, J. Illum. Engng. Soc. 3, 295, 1974.
[4] D. A. Davies, in: F. Benesovsky (ed.), Hochschmelzende Metalle (3. Plansee Seminar, Reutte/Tirol 1958), p. 69.
G. Comenetz, H. A. Johansen and J. W. Salatka, J. Illum. Engng. Soc. 2, 366, 1973.
C. W. Dawson, J. L. Nelson, W. E. Radzelovage and H. G. Sell, J. Illum. Engng. Soc. 2, 381, 1973.
[5] C. W. Dawson, Metall. Trans. 3, 3103, 1972.
[6] I. Langmuir, Phys. Rev. 34, 401, 1912.
[7] G. R. Fonda, Phys. Rev. (2) 21, 343, 1923, and (2) 31, 260, 1928.
[8] W. Elenbaas, Philips Res. Repts. 18, 147, 1963, and 22, 5, 1967.
[9] E. Fischer and J. Fitzgerald, J. appl. Phys. 45, 2895, 1974.

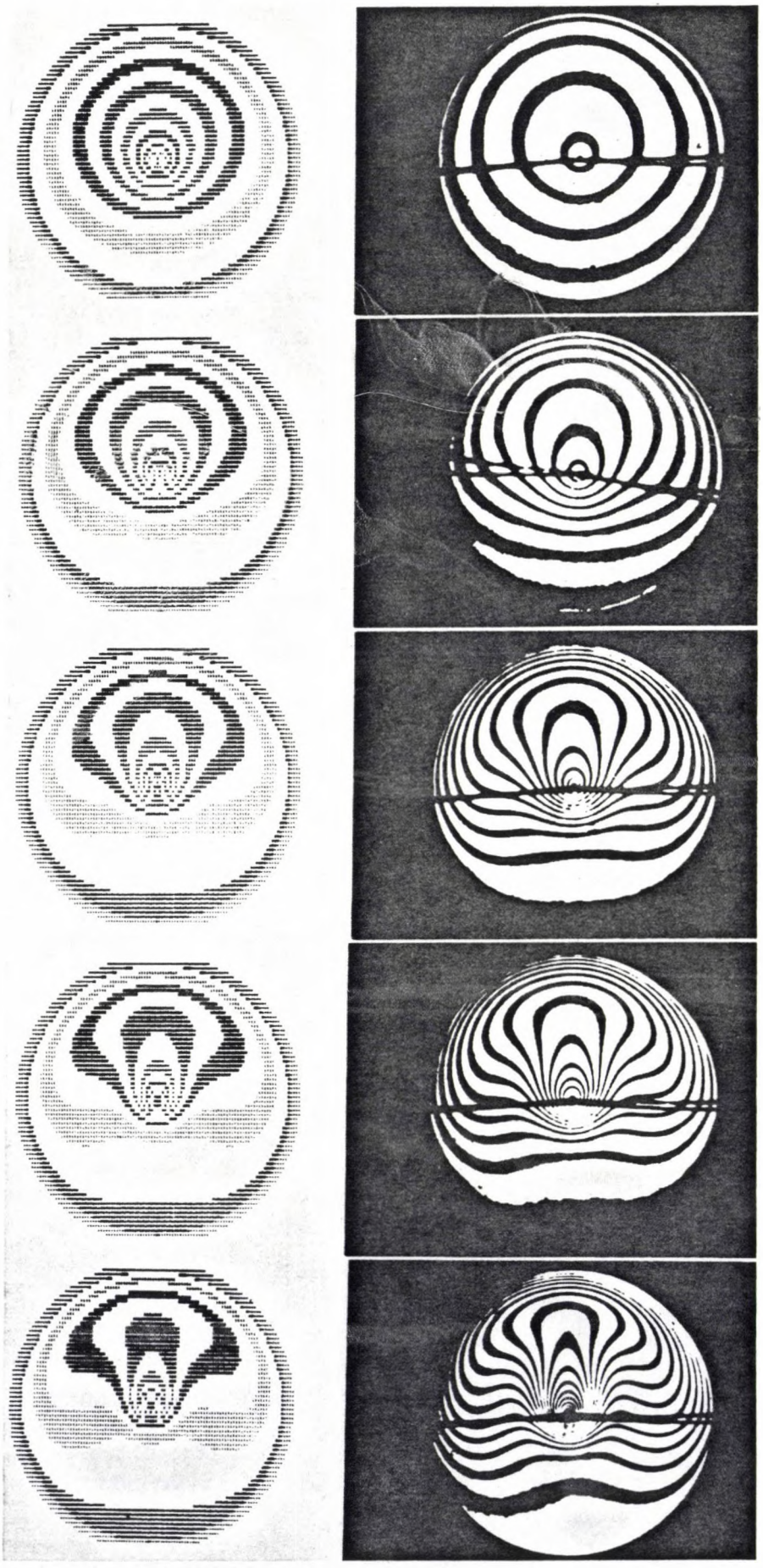
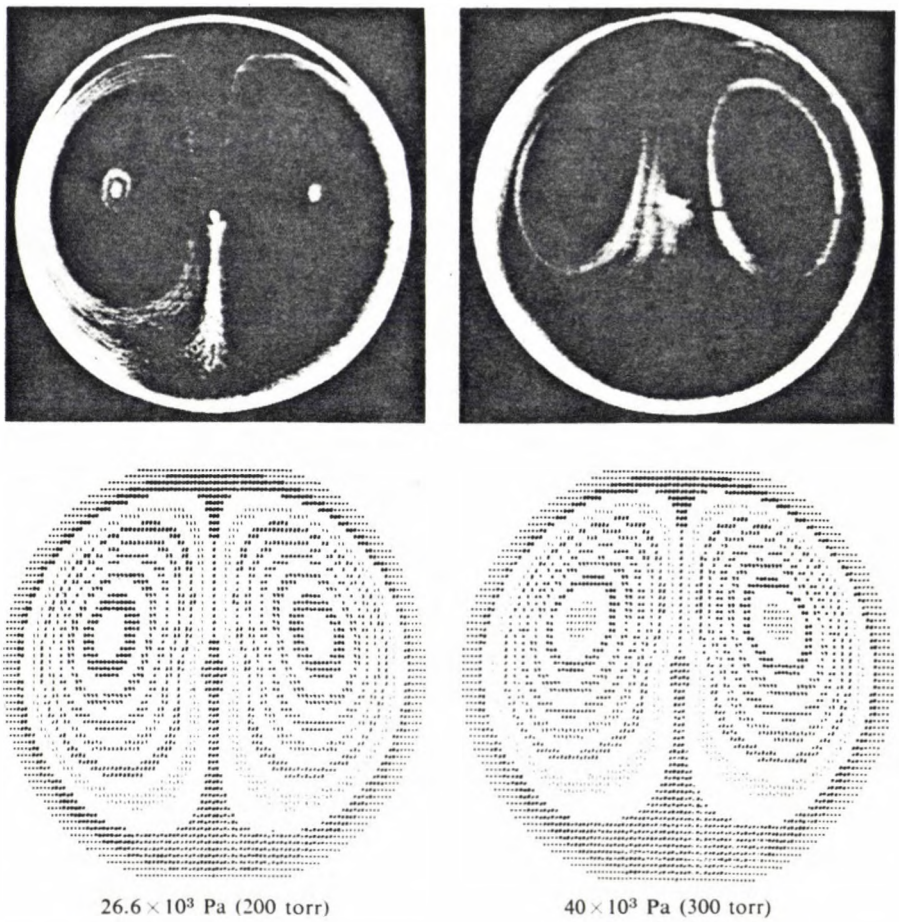


Fig. 1. Calculated temperatur distributions (*left*) and inter ference patterns (*right*) in a axially symmetric argon-fille incandescent lamp (argon pres sure p) in a plane perpendicula to the filament. In this lamp eac fringe represents an isotherm The diameter of the tube is 4 cm diameter of wire 200 μm.

the basic equations of flui dynamics, and the fring pattern obtained by inte ferometry. For the cylin drical lamp, and with th observer looking in th axial direction, the fringe represent the isotherms. Th calculated and experimen ally determined isotherm differed by less than 10 % Good agreement betwee theory and experiment wa also obtained for the flo field, as shown in *fig. 2.* the velocity field of the ga flow is known, the tungste transport from the inca descent wire can be ca culated, taking convecti transport into account a well as diffusion by conce tration gradients. Becau of the low tungsten conce tration in the vapour phas an experimental verificatic of the tungsten transpo was not possible in this cas The results of calculatio of the tungsten transpo rate as a function of lam size at a constant inert-g pressure are shown in *fig.* For the example chos — argon at a pressure 10^6 Pa (10 atm), wire diar eter 1 mm, wire temper ture 3200 K — the tran port rate is a minimum at tube radius of about 3 m

Lamps of this filling a size will always be halog lamps. The formalism us can also be applied in th



26.6×10^3 Pa (200 torr) 40×10^3 Pa (300 torr)

Fig. 2. Flow field, calculated and found experimentally, in the lamp of fig. 1 for two values of the argon pressure p.

case by simply changing the boundary conditions for the tungsten concentration at the wall. The optimum dimension for a minimum in the radial transport rate is hardly affected by the addition of a halogen.

As we said earlier, with coiled filaments it is also necessary to take into account the axial transport inside the coil. This effect, which we shall discuss in the next section, can be the factor that determines the life if there is a high temperature gradient along the filament.

Observations on coiled filaments

When a straight wire is coiled the free surface area, i.e. the area responsible for mass and energy losses, is reduced to about half of the corresponding total wire surface. This means that the evaporation rate is also reduced by a half [10]. A coiled filament should therefore have about twice the life of a straight wire of the same wire diameter, but if the hot-spot model is valid the deadly weight loss should be identical [1]. In the cases where the coiled filament has a smaller deadly weight loss than the corresponding straight wire, defects introduced by coiling, other transport mechanisms, and possibly changes in shape might also be of significance. Of course, other failure mechanisms than those

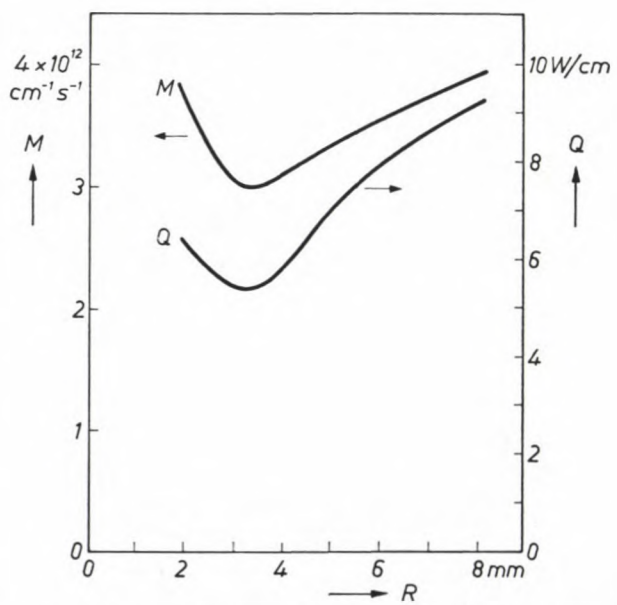


Fig. 3. A calculation of the number of tungsten atoms M leaving the filament per second, and the heat flow Q (both per unit of length) in axially symmetric argon-filled lamps as a function of the radius R for an argon pressure of 10^6 Pa (10 atm), a wire diameter of 1 mm and a wire temperature of 3200 K. The curves show a minimum at $R \approx 3$ mm.

[10] G. R. Fonda and A. A. Vernon, J. Opt. Soc. Amer. **22**, 223, 1932.

described by the spot model could also be present.

We shall now summarize the work on transport inside the coil with particular reference to the case of a high temperature gradient (turn-to-turn diffusion). We shall then present some recent results on the formation of crystal facets on the wire surface; these facets can influence the life by changing the energy balance as well as the mechanical stability of the coil.

earlier article [1]. Observations on low-voltage automobile lamps (type H3, 12 V/55 W) and projection lamps (12 V/100 W) have shown symmetrical hot-spot development on both sides of the centre. The filaments of these lamps do not burn out at the centre, where the temperature is highest at the beginning of life, but at locations where there is a maximum of the net axial component of the tungsten flow. The presence of

Fig. 4. Coiled-coil filament of a gas-filled incandescent lamp (220 V/200 W) shortly before the end of life. The wire has 'facets' from which the development of tungsten single crystals can be identified. The crystals extend over several turns. The photograph was taken with a scanning electron microscope.

Axial tungsten transport

The assumption that evaporation is the main transport mechanism does not necessarily mean that the filament will eventually burn out at the place where the initial filament temperature is highest. In the case of a low-voltage lamp (with a relatively large ratio of wire diameter to length) the influence of the heat conduction along the filament produces a marked temperature gradient in the axial direction. As a consequence there is a high concentration gradient of tungsten in the vapour phase along the coil. This sets up an axial mass transport from turn to turn. The conditions under which this axial transport determines the life, with experimental evidence of burn-out at positions of maximum axial transport, have been described in an

bromine or iodine in the envelope does not affect this transport as long as the amount of tungsten which is locally transported back to the filament in the form of tungsten-halide molecules is small compared with the evaporation rate. Because of the low thermal stability of the compounds formed by tungsten with iodine and bromine, this condition will be met in the region of the filament where the temperature is highest.

A quite different situation would arise if the dissociation of the tungsten-halide compound took place at such a high temperature that it only occured at the hottest parts of the filament. In principle a zero net tungsten transport via the vapour phase could then be achieved. This possibility, which might perhaps be attained with a gas mixture containing fluorine, will be discussed in part III of this article.

When axial transport determines the life of low-voltage lamps, a longer life should be obtainable if the temperature distribution along the coil is altered in such a way that the net mass flow per turn is constant along the coil. The probability of burn-out will then be equal for all turns. Such a temperature distribution can be obtained by treating the coil with fluorine [11].

Facet formation

Depending on the temperature of the incandescent filament and the pressure of the inert gas (e.g. if the pressure is about 10^5 Pa at temperatures below 3000 K) a tungsten coil develops flat surfaces, known as facets. A scanning-electron-microscope photograph of a coiled-coil filament is shown in *fig. 4* as an example. These facets can be identified as crystal planes. From the orientation of the facets the orientation of the single crystal in question, which extends over several turns, can easily be identified. The observation of facets has raised the question as to whether they affect the life of the filament. Development of facets might perhaps change the energy balance, causing a change in average or local temperature, or the mechanical stability might be affected if constrictions are formed.

The results of an investigation [12] on straight coils in argon (wire diameter 36 μm, inert-gas pressure 8×10^4 Pa) can be summarized as follows. At temperatures above 3000 K the filament wire remains circular in cross-section throughout its life. The life-determining transport mechanism appears to be evaporation, as identified by the activation energy. Below 3000 K, facets predominantly of {100} type develop. Between 3000 K and 2700 K the life-determining transport mechanism again appears to be evaporation. However, the observed life does not correspond to an extrapolation (based on the activation energy) of the results above 3000 K. We have found that as a result of the development of facets the temperature decreases and hence the life increases as compared to the situation (constant-voltage operation and equal initial temperature) in which the cross-section of the wire remains circular. The main cause of this is probably that the formation of facets increases the resulting effective emissivity of the filament.

Facets also occur below 2700 K, but the situation with regard to life is different. In this temperature region the results depend on the kind of current. When operated with direct current, the lamp has a considerably shorter life than when operated with alternating current. Scanning-electron-microscope pictures of coils operated with a.c. and d.c. current have shown that the coils look much the same outside, but are very different at the interior. Whereas the interior of an a.c.-operated coil remains more or less circular the inside of a d.c.-operated one was found to be lozenge-shaped.

Fig. 5. Whereas the inside of an a.c.-operated facetted coil remains more or less circular (see fig. 4) the inside of a d.c.-operated coil becomes lozenge-shaped. The coil shown was taken out of the lamp shortly before the end of life, and then broke.

Fig. 5 shows a d.c.-operated coil, which was switched off near the end of life, and broke after it was taken out of the lamp. As can be seen, the facets on the outside and inside of the coil are not parallel, which leads to constrictions. These constrictions apparently develop further into deep notches until the coil becomes mechanically unstable and breaks. *Fig. 6* shows that little is left of the initial cross-section at the position where the coil broke.

The temperature dependence of life below 2700 K indicates that evaporation is no longer the life-determining transport mechanism for d.c.-operated coils. Just as in d.c.-operated vacuum lamps below 2300 K [13], surface migration appears to be the mechanism responsible for failure. In vacuum lamps this limiting temperature is lower than in the gas-filled lamp because the inert gas reduces the evaporation rate by about two orders of magnitude, whereas surface migration is probably not affected. The site of failure was found in most cases to correspond to a hot spot that was present before the facets had formed. This indicates that the facetting observed for single-coil filaments does not

[11] J. Schröder, this issue, p. 332.
[12] W. Lems, H. Kinkartz and W. Lechner, Philips Res. Repts. **30**, 218, 1975.
[13] N. P. Demas, Illum. Engng. **66**, 267, 1971.

give rise to new hot spots during life. Facetting could, however, lead to the faster development of an initial hot spot.

In view of the foregoing, it appears generally useful to warn against extrapolating the results of high-temperature 'forced' life tests to lower operating temperatures. This could for example lead to the extrapolation of results obtained on non-facetted coils at high temperatures to the region where, because of the development of facets, the temperature decreases during life, and finally to a temperature region where it is not in fact the evaporation that determines life but more usually the surface migration.

Although the mechanisms that determine life are reasonably well understood, only modest improvements in the lamp's efficacy can be attained by improving the quality of the filaments. This, however, does not imply that such improvements are unimportant.

Only by establishing a truly regenerative cycle that will suppress the development of spots caused by evaporation — with no local transport via the vapour phase — can any substantial improvement in lamp life or efficacy be expected. In that case other burn-out mechanisms will probably determine the life of incandescent lamps.

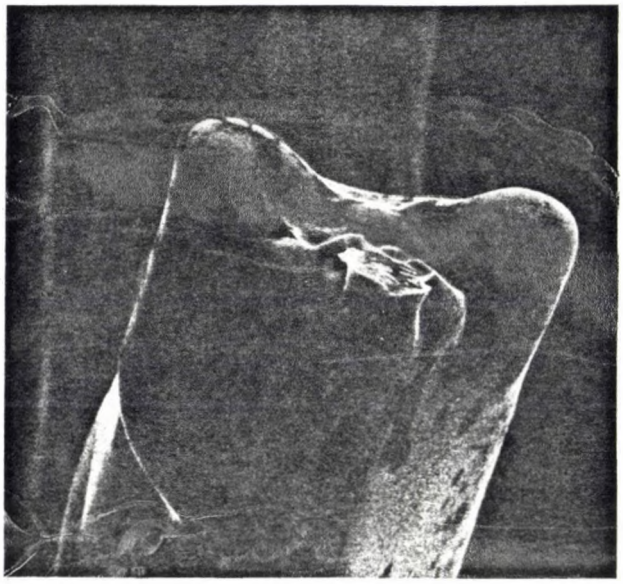


Fig. 6. At the place where the coil in fig. 5 broke its cross-section was much smaller than the initial value.

Summary. The description of the burn-out mechanism of incandescent filaments in terms of the development of a hot spot from an initial inhomogeneity through enhanced evaporation appears to be qualitatively correct in the great majority of cases. An important cause of initial hot spots is the presence of local constrictions in wire diameter. In cylindrical gas-filled lamps there is a minimum in the evaporation as a function of diameter. In low-voltage lamps the transport of tungsten occurs mainly in the axial direction of the coil (turn-to-turn diffusion). Such filaments do not necessarily burn out at the hottest place (the centre) but at the places were the axial tungsten transport is greatest. Facets, which appear on the wire when the operating temperature is lower than 3000 K, do not seem to be a cause of hot-spot development during life. A substantial improvement in lamp life (or efficacy) can only be expected if the development of hot spots is suppressed by a truly regenerative cycle.

III. Regenerative chemical cycles in tungsten-halogen lamps

J. H. Dettingmeijer, G. Dittmer, A. Klopfer and J. Schröder

Halogen lamps

Incandescent lamps in which iodine [1] or bromine [2] is added to the inert-gas filling have been widely used for about fifteen years. They are mainly used in projectors and in reflector-type lamps. Their efficacy is about twice that of the ordinary incandescent lamp because the temperature of the filament in a halogen lamp is higher.

In the ordinary incandescent lamp the size of the bulb must be matched to the temperature of the filament to keep blackening of the bulb — and hence light losses — within reasonable bounds. In halogen lamps, on the other hand, the evaporated tungsten is returned to the filament by a chemical transport process, which prevents blackening of the bulb. This makes it possible to use small bulbs, and because of their greater mechanical strength, these bulbs can be filled with inert gas at a higher pressure. The small volume of such bulbs also permits the use of a heavier

J. H. Dettingmeijer is with the Philips Lighting Division, Eindhoven; Dr G. Dittmer, Dr A. Klopfer and Dr J. Schröder are with Philips GmbH Forschungslaboratorium Aachen, Aachen, West Germany.

inert gas than argon, even though it is more expensive. Both measures reduce the transport of tungsten and thus permit a higher filament temperature [3].

As will be shown below, in iodine or bromine lamps the addition of halogen does not in itself have any moderating effect on the processes that lead to burn-out in lamps filled with inert gas. Fluorine seems to offer the best prospects of a true regenerative cycle, but the development of a fluorine lamp will be long and difficult, and there is no certainty as yet that it will succeed.

Halogens as chemical transport media

The transport media most often considered for chemical cycles are the halogens iodine, bromine, chlorine and fluorine, possibly combined with oxygen. This is because the tungsten halides and oxyhalides are sufficiently volatile at the relatively low temperature of the bulb. A simplified picture of the chemical transport can be obtained by considering that the evaporated tungsten from the hot incandescent filament diffuses through the inert gas towards the bulb wall, reacting in the cooler zones with the transport medium to form a volatile compound. In their turn the reaction products diffuse in the direction of the filament. In certain temperature zones the volatile compound dissociates completely or partly, depending on its thermal stability, and in the process tungsten is released. The transport medium is then available again for renewed reactions in the cooler zones. The result of the process is that there are no tungsten deposits on the wall, and therefore there is no overall weight loss for the filament during life.

Although the tungsten transport towards the bulb wall may be suppressed by the halogen cycle, there may nevertheless be a transport along the filament coil, as in inert-gas-filled lamps without halogen. This happens when the halogen compounds dissociate in zones where the temperature is substantially lower than that of the incandescent coil. Since in this case the tungsten transport is largely determined by the vapour-pressure curve, the development of a hot spot is not prevented [4]. For this reason attempts have long been made to use those halogens that form the most stable compounds with tungsten [5]. The thermal stability of the tungsten halides increases from the iodide to the fluoride. However, the chemical reactivity increases in the same order, and this is one reason why the first halogen lamps put on the market contained iodine and bromine.

Transport and equilibrium conditions

We shall now take a closer look at the processes that take place in the halogen lamps. Besides the physical processes that also occur in ordinary gas-filled lamps, such as evaporation and diffusion, there are also chemical processes in which tungsten compounds are continuously being formed and broken down again. At the incandescent coil itself and at the filament supports the reactions between halogen and tungsten are heterogeneous, whereas in the space between bulb and coil the reactions are homogeneous because they take place completely in the gas phase. Since no tungsten settles on the bulb wall in these lamps, we only need to consider what happens at the filament coil, where we examine the heterogeneous equilibrium of solid tungsten and halogen. It is thus assumed that local thermodynamic equilibrium exists in the immediate vicinity of the coil, which means that the heterogeneous reactions are fast compared with the diffusion processes. The composition of the gas phase can then be calculated as a function of temperature and pressure, provided of course that we know which chemical compounds are present and their thermodynamic characteristics. Often, however, very little accurate thermodynamic information or none at all is available on the compounds at high temperatures. In such cases there is no option but to calculate with estimated values and then to verify and improve the estimates by matching them with equilibrium compositions found experimentally.

It is more convenient, particularly in complex systems, to treat the tungsten-containing components of the gas (free tungsten and compounds) not separately but collectively. This is done by taking the sum of the partial pressures, each multiplied by a weighting factor equal to the number of tungsten atoms per molecule of the compound in question. We shall refer to this summed pressure as the total effective tungsten pressure and express it as Σp_W.

The transport of tungsten now proceeds from places near the coil where the total effective tungsten pressure is high to places where it is low, and tungsten is therefore deposited. Because of the high pressure of the inert gas, the mass transport takes place by diffusion. Since a quantitative description of the tungsten transport is extremely complicated, owing to the different diffusion coefficients of the various components, we shall confine ourselves here to a qualitative description of the reactions in the lamp.

[1] E. G. Zubler and F. A. Mosby, Illum. Engng. 54, 734, 1959. J. W. van Tijen, Philips tech. Rev. 23, 237, 1961/62.
[2] G.R. T'jampens and M. H. A. van de Weijer, Philips tech. Rev. 27, 173, 1966.
[3] W. Elenbaas, Philips Res. Repts. 18, 147, 1963.
[4] A treatment of the hot-spot model has been given in H. Hörster, E. Kauer and W. Lechner, Philips tech. Rev. 32, 155, 1971.
[5] See for example J. Schröder, Philips tech. Rev. 26, 111, 1965.

Characteristic features of lamps containing iodine, bromine, chlorine or fluorine as transport medium

Iodine

Originally iodine was chosen as an additive because it is the least reactive of the halogens. It was assumed at the time that the return transport took place by means of the compound WI_2. An indication of what was actually happening came from the observation that small traces of oxygen were indispensable to the cyclic process in the iodine lamp. The explanation was found when it was shown that at low temperatures WO_2I_2 is the stable compound with sufficient volatility [6], and that the cycle stopped as soon as the system was deprived of oxygen by the addition of carbon [7].

Fig. 1 gives the total tungsten pressure Σp_W as a function of temperature at an initial pressure of 10^3 Pa (10^{-2} atm) of I_2 with an addition of 0.01 % of oxygen. The filling pressure of iodine is relatively insignificant, since Σp_W is determined mainly by the added oxygen. The total quantity of oxygen that may be present in the gas phase is in turn limited by the vapour pressure of the solid tungsten oxide that settles on the colder parts of the lead-in wires. The tungsten will preferentially be transported to the point at which the curve has a minimum. At an operating temperature of 3200 K the tungsten transport between neighbouring turns of the coil that have slightly differing temperatures takes place almost completely as tungsten vapour, and hence goes from the hotter to the cooler turn. At operating temperatures corresponding to the minimum in the Σp_W curve, and where tungsten is transported from the cooler to the hotter turns, the cool ends of the coil are attacked. This effect increases with the amount of oxygen (or water vapour) present [8].

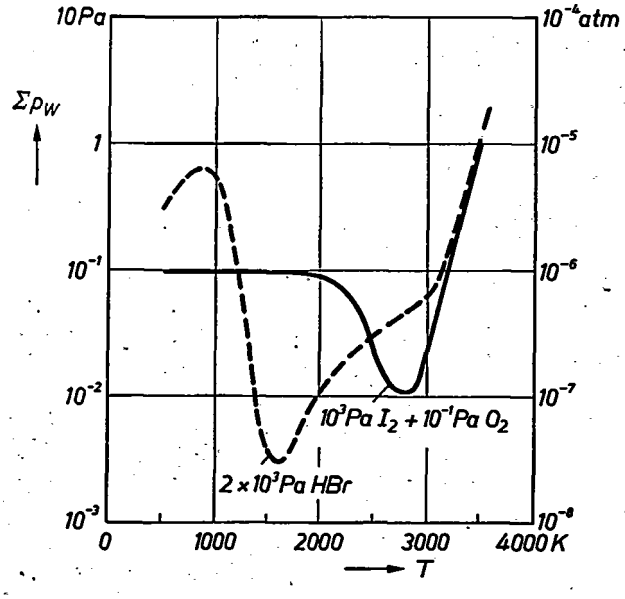


Fig. 1. Temperature dependence of the total effective tungsten pressure Σp_W in a heterogeneous equilibrium system with I_2 (10^3 Pa) and O_2 (0.1 Pa) or with HBr (2×10^3 Pa). Σp_W is the sum of the partial pressures of the gaseous tungsten compounds, when the number of tungsten atoms per molecule for each compound is taken into account.

Bromine

Whereas the halogen is used in its elemental form in iodine lamps, this has long been discontinued for bromine lamps. This was because it turned out that the dosage of pure bromine was extremely critical, since any excess of bromine attacks the relatively cool filament ends and supports. It was not until hydrogen bromide was added to the filling that these undesired transport effects could be almost completely eliminated [2].

The total effective tungsten pressure in hydrogen-bromide lamps with a filling pressure of 2×10^3 Pa of HBr is also shown in fig. 1 as a function of temperature. At low temperatures HBr hardly reacts at all with tungsten. Only WBr_4 and WBr_5 then have a small partial pressure. The total effective tungsten pressure is approximately three thousand times smaller than the partial pressure of HBr, which means that the amount added is not very critical. At high temperatures the HBr is largely dissociated. The compounds WBr_2 and WBr, which have low partial pressures, are in equilibrium with atomic bromine. As can be seen from the curve, Σp_W at the bulb wall, which has a temperature of 800 K, is greater than at the much hotter filament; consequently there is no blackening of the bulb. Because of the low thermal stability of the tungsten bromides, Σp_W at the filament will mainly be determined by the tungsten vapour pressure. Under these circumstances the same burn-out mechanisms as in the iodine lamp and the ordinary gas-filled lamp must therefore prevail.

It can thus be seen that the application of the total effective tungsten pressure for the heterogeneous equilibrium system explains very satisfactorily why the halogen cycle with oxyiodide or bromide is regenerative for the bulb wall but not for the filament.

The shape of curves like those shown in fig. 1 depends to a great extent on the enthalpy of formation of the appropriate tungsten-halogen compounds. The curve for iodine in fig. 1 is based on values found in the literature. In calculating the other curve a value of 234 kJ/mol (56 kcal/mol) was taken for the enthalpy of formation of gaseous WBr_2. The stability taken in fig. 2 for gaseous WCl_2 and WF_2 [9] is also lower than has usually been assumed by other investigators.

Chlorine and fluorine

As we noted earlier, the thermal stability of the tungsten compounds increases from bromine to fluorine. A comparison of the total effective tungsten pres-

sure as a function of temperature on the addition of the pure halogens bromine, chlorine and fluorine is shown in *fig. 2*. Stable compounds are still found at very high temperatures in the tungsten-fluorine system; if an appropriate filling pressure is used the total effective tungsten pressure only falls off at high temperatures.

For this reason an experimental investigation was made of the tungsten-fluorine system. The tungsten transport in a lamp was measured under stationary conditions and also under dynamic conditions, as in a gas-flow system [9]. *Fig. 3* shows the composition of the gas phase. At the low temperature the compounds WF_6 and WF_5 are stable; their chemical and thermodynamic properties are well known [10].

At high temperatures, WF_4 predominates whereas WF_3, WF_2 and WF occur in equilibrium only when there is an excess of atomic fluorine. Since it has not yet proved possible to isolate these lower tungsten fluorides, their thermodynamic properties were determined indirectly by varying the reaction enthalpies until the calculated Σp_W values corresponded to the experimental values. The Σp_W curve corresponding to this gas-phase composition has already been shown in fig. 2. Just as in the case of iodine and bromine, the marked decrease towards the higher temperatures causes a continuous tungsten transport from low-temperature to high-temperature sites, which again gives corrosion of the cooler filament ends. In the next section we shall see how this problem may perhaps be overcome.

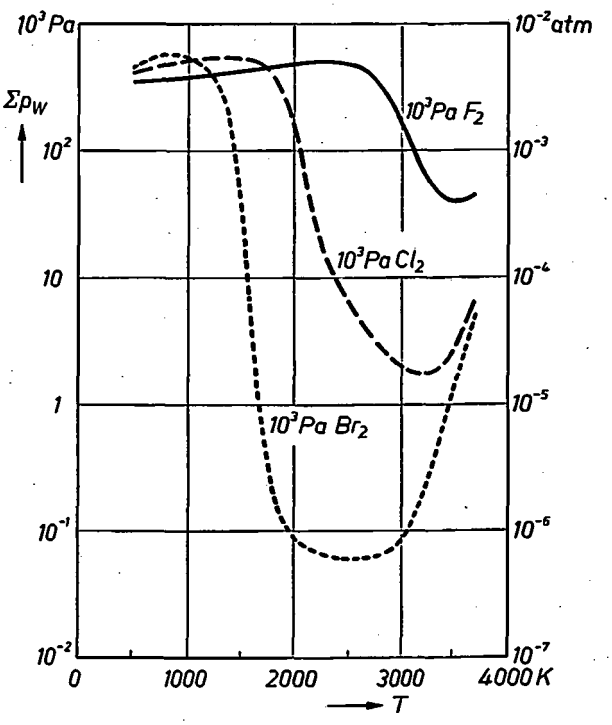


Fig. 2. As fig. 1, but now for heterogeneous equilibrium systems with Br_2 (10^3 Pa), Cl_2 (10^3 Pa) or F_2 (10^3 Pa).

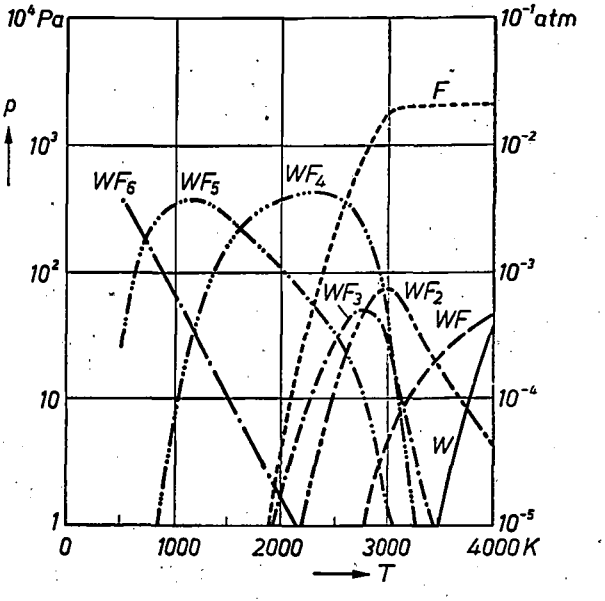


Fig. 3. The partial pressure in the heterogeneous equilibrium system with 10^3 Pa of F_2 above solid tungsten, as a function of temperature.

A regenerative cycle using F_2, SiF_4 and Cl_2

Present investigations are aimed at discovering chemical transport systems that operate in such a way that the filament contains no zones from which tungsten is transported preferentially. Translated into terms of total effective tungsten pressure this means discovering systems that keep Σp_W constant within the range of temperatures found in the lamp, for which this quantity first starts to decrease at temperatures above the operating temperature.

In the tungsten-fluorine system Σp_W must be increased both at low temperatures and at very high temperatures to the level required by the WF_4. To overcome the deficiency at the high temperatures, the partial pressure of the lower-valency tungsten fluorides must be increased. This can be done by adding a volatile fluorine compound that first starts to release atomic fluorine at temperatures in the region of 3000 K. Silicon tetrafluoride is a suitable compound, since it dissociates above 2500 K into SiF_2 and 2F. The deficiency at the lower temperatures can be cor-

[6] J. H. Dettingmeijer and B. Meinders, Z. anorg. allgem. Chemie 357, 1, 1968.
J. Tillack, *ibid.* 357, 11, 1968.
H. Schäfer, D. Giegling and K. Rinke, *ibid.* 357, 25, 1968.
[7] J. H. Dettingmeijer, J. Tillack and H. Schäfer, Z. anorg. allgem. Chemie 369, 161, 1969.
[8] J. H. Dettingmeijer, B. Meinders and L. M. Nijland, J. less-common Met. 35, 159, 1974.
[9] G. Dittmer, A. Klopfer, D. S. Ross and J. Schröder, J. Chem. Soc. chem. Comm. 1973, 846.
[10] J. Schröder and F. J. Sieben, Chem. Ber. 103, 76, 1970.
J. Schröder and F. J. Grewe, Chem. Ber. 103, 1536, 1970.

rected by adding chlorine. *Fig. 4* shows the curve for the total effective tungsten pressure when the additions to the tungsten-fluoride system are optimally dosed.

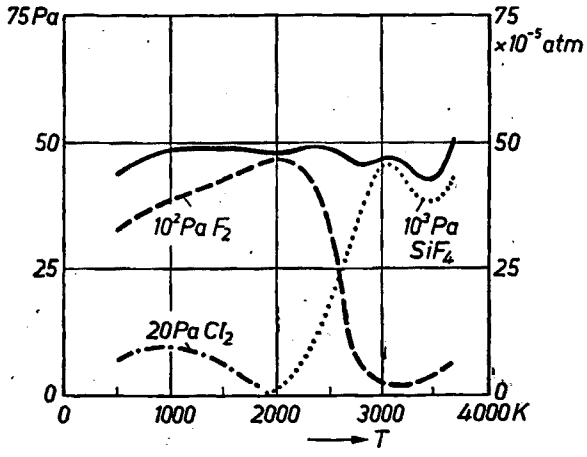
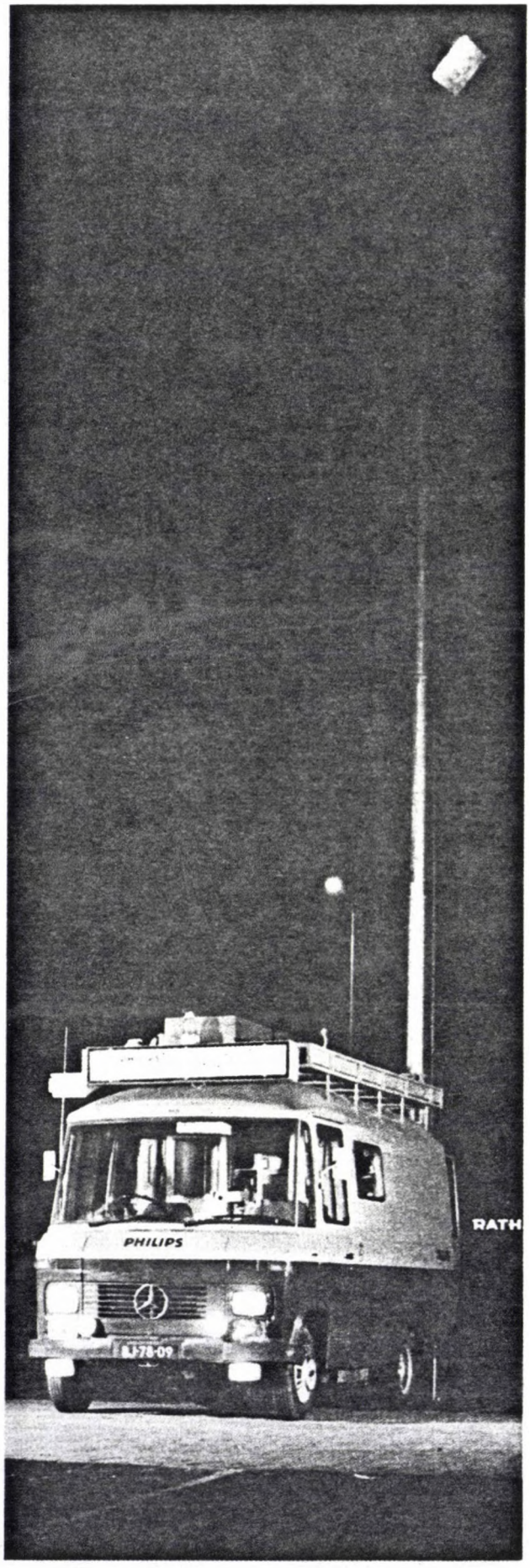


Fig. 4. The solid curve gives the total effective tungsten pressure in a heterogeneous equilibrium system with SiF_4 (10^3 Pa), F_2 (10^2 Pa) and Cl_2 (20 Pa). The other curves refer to the corresponding systems with SiF_4 or F_2 or Cl_2 separately.

The curve is almost horizontal, apart from a smal ripple.

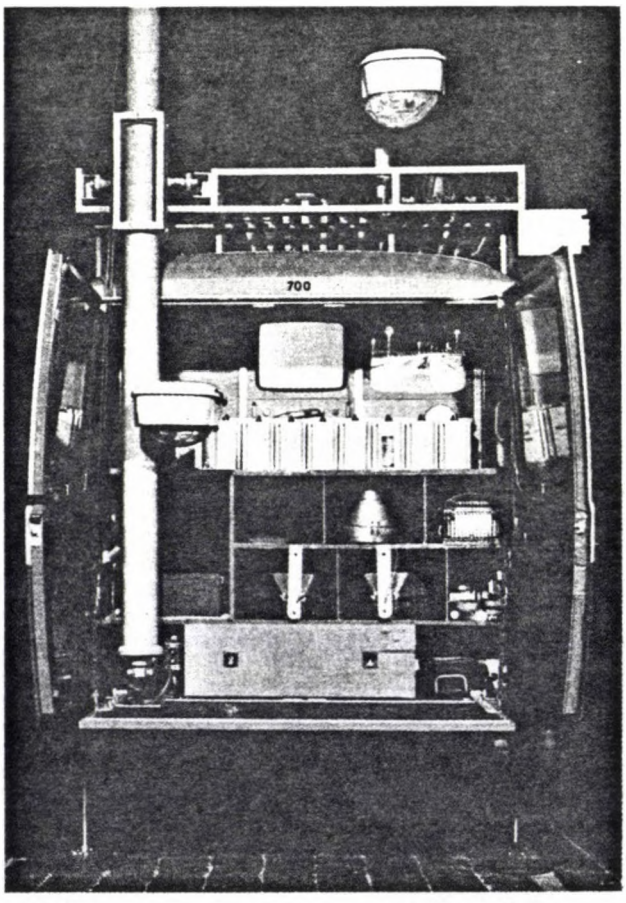
The curve in fig. 4 relates to the heterogeneous chem ical equilibrium system. To give a good description o the transport phenomena in the lamp it is necessary tc bear in mind that the diffusion coefficients of the react ing components are different. Only when this is done is it possible to determine the effect of the ripple in the curve on the tungsten transport and the extent to which a regenerative cycle can be achieved or approximated.

Summary. The tungsten cycle in halogen lamps is described and discussed with the aid of thermodynamic-equilibrium calculations. The chemical cycles in commercially available lamps keep the bulb clear but are unable to prevent the formation of hot spots. The use of a mixture of fluorine, silicon tetrafluoride and chlorine may well bring us a step nearer to a chemical transport system in which no spot is formed on the filament from which tungsten is transported at a faster rate than elsewhere. Final conclusions about the practicability of such a system have not yet been reached.



Demonstration and test vehicle for exterior lighting

The vehicle shown, with its extensible mast, can readily be used anywhere for demonstrating the effect of lamps and lanterns for external lighting. In addition, the place where the driver normally sits — on the right of the picture in this case — is occupied by equipment that can measure and record the luminance of the road surface. It can also make a video recording of the 'driver's' view of the road and traffic. Provided a telephone line is available, a terminal in the vehicle can be connected to a computer in Eindhoven. This enables the crew to make a rapid design for any lighting project.



Discharge lamps

R. Bleekrode, M. Koedam and L. Rehder

Introduction

The best-known gas-discharge sources of light, the sodium lamp and the mercury-vapour lamp, originally developed in completely opposite directions. To increase the useful output from the discharge, the gas pressure and current density in the sodium lamp always had to be made smaller, whereas in the mercury lamp they always had to be made larger. This can be attributed to the fact that 'resonance radiation' — emitted when an atom returns from an excited state to the ground state — is enhanced by low pressure and low current density, whereas the radiation from transitions between higher energy levels is enhanced by high pressure and high current density. In sodium vapour the visible radiation is resonance radiation, and the radiation from higher transitions is primarily in the infrared. In mercury vapour the visible radiation comes from higher transitions; the resonance radiation is in the ultraviolet [1].

Among the lamps listed in *Table I* the *low-pressure sodium lamp* and the *super-high-pressure mercury lamp* are the ultimate outcome of this development. The first type of lamp has the highest luminous efficacy of all radiation sources, and is widely used for road lighting. The other type, together with the super-high-pressure xenon lamp, is outstanding for its very high luminance. This makes it possible to obtain a high luminous flux from a compact source (CS), giving lamps that are particularly useful for optical systems (searchlights, cinema projectors).

The light from the low-pressure sodium lamps is virtually monochromatic. It creates distinct contrasts but colours cannot be distinguished in it. In road lighting, however, this disadvantage is acceptable. The high pressure mercury lamp is somewhat better in this respect; its visible output consists of several lines. An increase of pressure generally leads to line-broadenin

Table I. A selection from the discharge lamps marketed by Philips, and the tin-halide lamp, now in an advanced stage of development. The efficacy figures given under η in the first column exclude ballast losses; the figures given under η' in the second column include them. L luminance. R_a general colour-rendering index [2].

	Code	η (lm/W)	η' (lm/W)	L (cd/cm²)	R_a	Application
Low-pressure sodium lamp 180 W	SOX	183	150	10		Road lighting
High-pressure sodium lamp 250 W	SON	100	88	25 [+]		Road lighting
High-pressure mercury lamp [*] 250 W	HPLN	51	47	10		Road lighting
Super-high-pressure mercury lamp 100 W	CS	20		170 000		Projection, searchlights, etc.
Super-high-pressure xenon lamp 2500 W	CSX	40		72 000	90	Projection, searchlights, etc.
Fluorescent discharge lamps standard 40 W	TL33	80	67	0.75	65	Offices and other work spaces
de luxe 40 W	TL84	80	67	0.75	85	Offices and other work spaces
Metal-iodide lamp 375 W	HPI	83	78	600 (14[+])	66	Floodlighting
Tin-halide lamp 250 W		65	60	400 (9[+])	90	Indoor lighting

[*] Outer envelope coated with fluorescent layer.
[+] For a lamp with outer envelope coated with diffusing layer.

Dr R. Bleekrode is with Philips Research Laboratories, Eindhoven; Dr M. Koedam is with the Philips Lighting Division, Eindhoven; Dr L. Rehder is with Philips GmbH Forschungslaboratorium Aachen, Aachen, West Germany.

[1] See the first article in the very first volume of this journal G. Heller, Comparison between discharge phenomena i sodium and mercury vapour lamps.

[2] This quantity is defined in J. J. Opstelten, D. Radielović an J. M. P. J. Verstegen, this issue, p. 361.

and enhanced continuum radiation. The spectrum of a *super-high-pressure xenon lamp*, for example, consists mainly of a recombination continuum in the visible region (*fig. 1a*).

The development has diversified along various paths. The desire for at least some colour discrimination at certain traffic routes and road intersections can now be met by the *high-pressure sodium lamp*. Because of the marked broadening of the sodium lines (fig. 1*b*) this lamp allows some colour discrimination, yet maintains a very high luminous efficacy. *Low-pressure mercury-vapour lamps* in the form of fluorescent lamps have been in use for some decades. In these lamps the dominant ultraviolet radiation is no longer accepted as a loss but converted into visible radiation. The nature of the fluorescent powders determines the spectrum. These lamps combine good luminous efficacy with good colour rendering. Their main use is for lighting in offices and other workplaces. A fluorescent powder, emitting in the red, is also used nowadays to improve the colour rendering of the high-pressure mercury lamp. Like the high-pressure sodium lamp, it is used for outdoor lighting. Although its colour rendering is better, it has a much lower luminous efficacy than the high-pressure sodium lamp.

Many attempts have been made in the past to improve the colour rendering of discharge lamps by using metals that give a better spectrum than mercury.

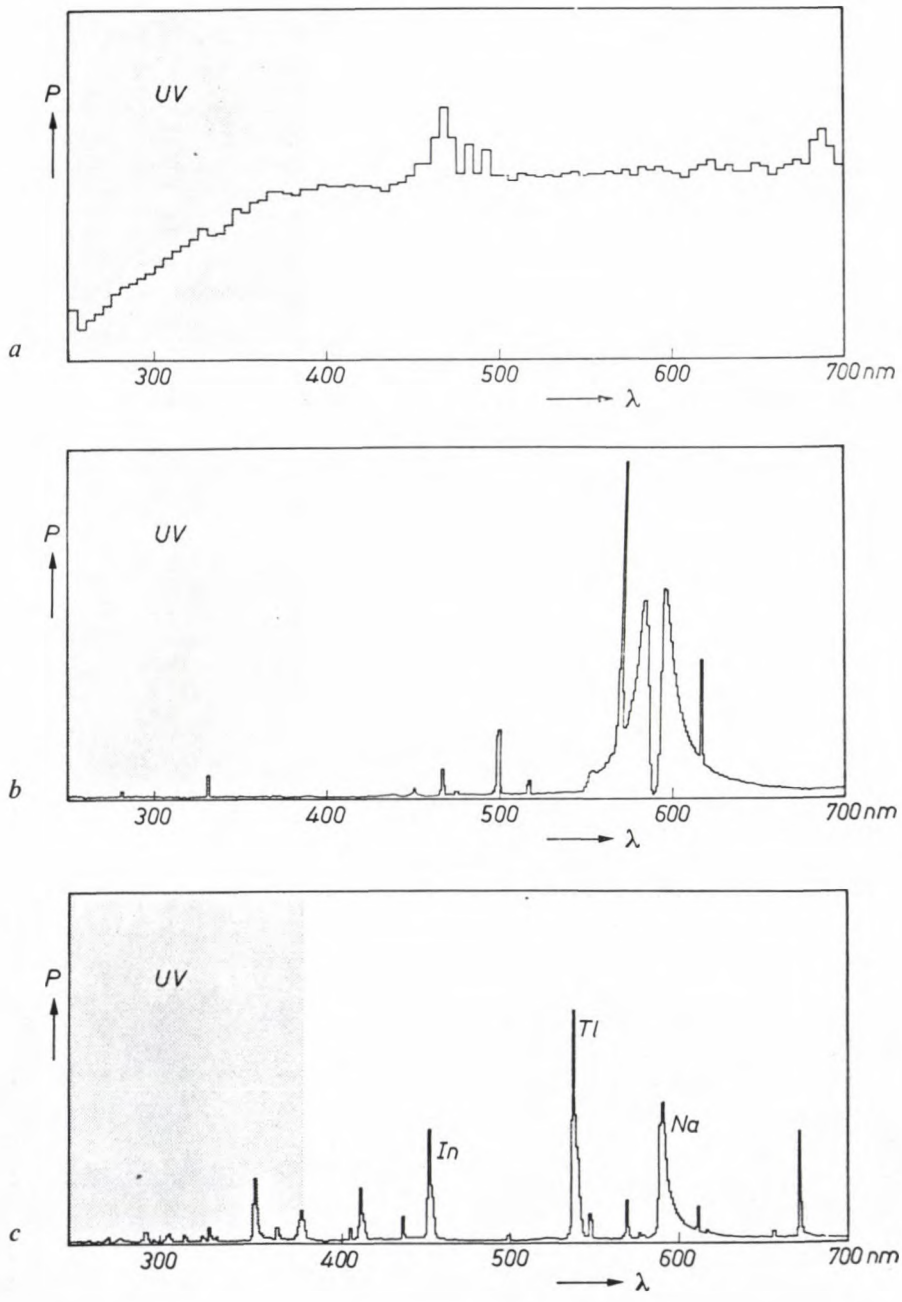


Fig. 1. Spectral power distributions $P(\lambda)$ of the radiation from three of the lamps in Table I. *a*) Super-high-pressure xenon lamp, *b*) high-pressure sodium lamp, *c*) HPI lamp. Good colour rendering is obtained with a continuous spectrum (*a*), but also with a limited number of lines in the visible spectrum (*c*). The pressure-broadened sodium line of (*b*) gives some colour discrimination, though only to a limited degree. This line clearly shows the self-absorption so characteristic of optically dense plasmas. The vertical scales for the power are arbitrary and differ for (*a*), (*b*) and (*c*). There are steps in the spectra because the measuring equipment records the total power in intervals of 5 nm (*a*) or 2 nm (*b* and *c*).

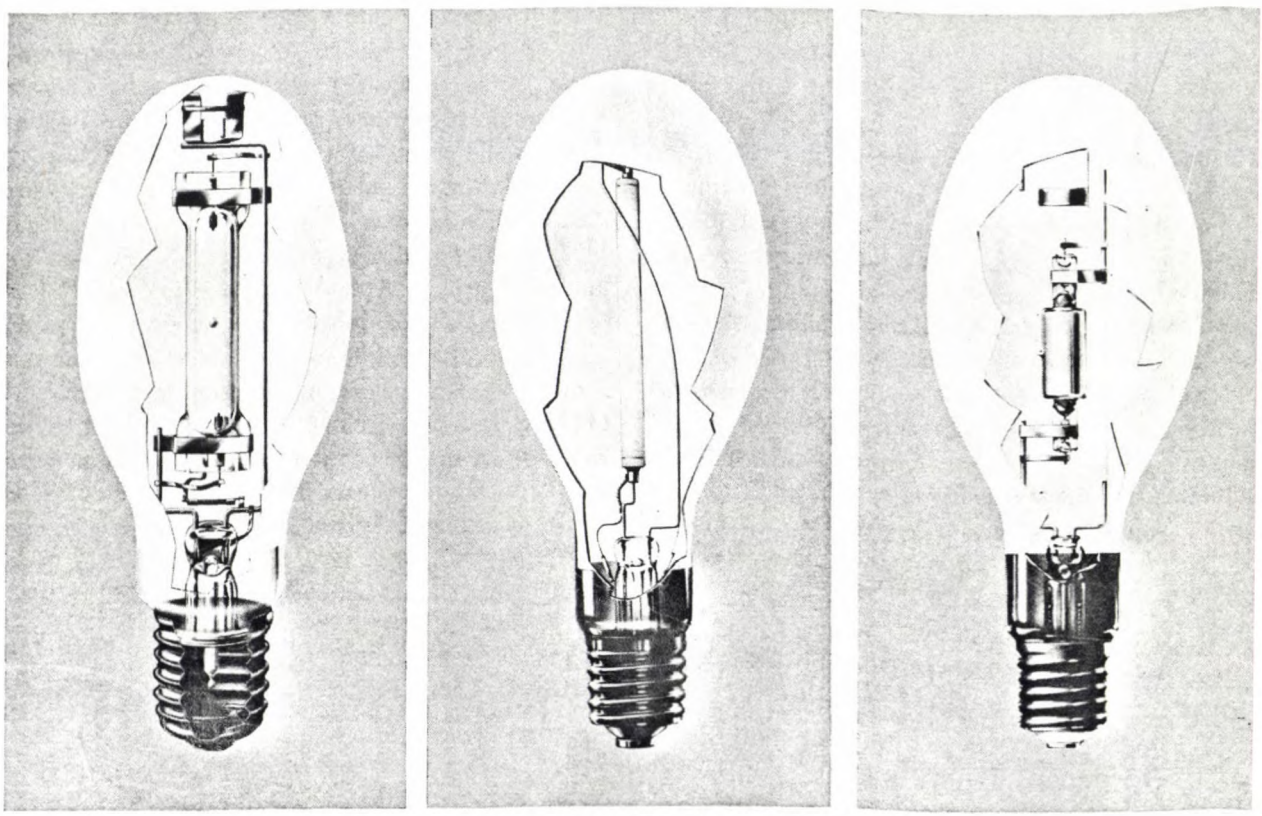


Fig. 2. Three high-pressure discharge lamps. From left to right: a high-pressure mercury lamp (HPLN), a high-pressure sodium lamp (SON) and a metal-iodide lamp (HPI). The outer envelopes of the lamps are made with a layer that provides diffusion and in the HPLN lamp also provides fluorescence. (SON and HPI lamps with clear outer envelopes are shown on pages 336 and 349 of this issue.)

Owing to chemical difficulties, however, no success was achieved until these metals were introduced into the lamp as an iodide or some other halide. In the discharge itself the halide dissociates and the radiation emitted is from the metal. The resulting *metal-halide lamps* may be roughly divided into three types: those that emit a spectrum with a relatively small number of lines, those emitting a large number of more or less equally distributed lines, and those whose emission is mainly a continuous spectrum of molecular radiation. Examples are the HPI lamp with iodides of sodium, thallium and indium (see fig. 1*c*), lamps with iodides of dysprosium and other rare-earth metals, and the 'tin lamp', respectively. This tin lamp contains mercury, tin iodide and tin chloride, and in some cases lithium chloride is added to make the light still more white. Metal-halide lamps are mainly used for lighting sports stadiums, roads and shopping streets. The tin lamp is more suitable for indoor lighting. Its colour temperature can be selected within a wide range by an appropriate choice of the tin halides.

These lamps, three of which are shown in *fig. 2*, are the subjects of continuous research [3], some idea of which will be given in this article. Other articles in this issue will deal with more specific aspects. The greatest activity is of course to be found in the younger branche of the lamp family — the high-pressure sodium lam and the metal-halide lamp, which are barely ten yea old.

The research may broadly be divided into the phy ical investigation of the gas discharge itself and wor on the technological aspects of lamp manufactur Research on the discharge is ultimately aimed at in proving the luminous efficacy and colour quality of th lamps, and also at improving the current-voltage cha acteristic with respect to the power supply. The goals may not always be apparent, since improvemer can only be achieved by means of a better understan ing of the complex processes taking place inside th lamp, and therefore the investigations are initially co cerned with acquiring that understanding. The techn logical work is primarily aimed at extending the usef life of the lamps.

Research on gaseous discharges

The radiation from these lamps is produced in th gas-discharge column, which takes up almost the enti space between the electrodes. The physical state in th column, and in particular the field-strength, is virtual

independent of the distance to the electrodes. The ionized gas (the 'plasma') is practically neutral. Space charges are found only in layers near the electrodes [4]. In this section we shall confine ourselves to the phenomena in the column. It will also be assumed that the column has cylindrical symmetry.

Electrical conduction and light production in the column arise from the following basic processes. The free electrons are accelerated by the field. Through inelastic collisions they excite and ionize atoms or molecules. In the steady state the ionization is in equilibrium with the recombination of electrons with ions. In low-pressure discharges the recombination takes place mainly at the wall, in high-pressure discharges mainly in the interior. The current is carried by the highly mobile electrons, compared with which the ions and the atoms are almost at rest. The emitted radiation originates from excited atoms or molecules returning to lower energy levels. Radiation may also be emitted as a result of the recombination of electrons with ions.

We shall briefly recall here the functions of the 'buffer gas' which is nearly always used, e.g. the inert gas in a low-pressure sodium lamp. The buffer gas limits the mobility of the electrons, and this has two advantages: firstly, the electrons do not reach the wall so quickly and thus remain longer available to the discharge, and secondly the current is kept at practical values. Since the lowest excitation energy of the inert gas is much higher than that of the sodium, the collisions of the electrons with the inert-gas atoms are virtually elastic and only a small fraction of the energy is transferred per collision. Nevertheless, because of the large number of such collisions, energy transfer does take place from the electron gas to the inert gas, which means that the inert gas has another important function: it makes an essential contribution to the heat balance of the lamp. Finally, it is the discharge in the inert gas that makes it possible to start the *cold* lamp, which of course initially contains virtually no sodium *vapour*. In high-pressure lamps the role of buffer gas is usually taken by mercury. Ignition, however, is again by means of an inert gas.

This simple qualitative description of the processes inside the lamp is not sufficient if we wish to optimize the design of a particular lamp, in other words to choose the gas filling, the gas pressure, the tube diameter and so on in such a way that the light production meets the specifications as closely as possible. To do this it is necessary to go into more detail, and many complications are then encountered. These arise from the fact that, although the properties of the plasma do not depend upon the position in the axial direction, they are certainly not independent of the distance from the axis: the degree of ionization of the plasma decreases from the centre towards the wall. Radial gradients are therefore present in the concentrations of electrons, atoms and ions, and also in temperature, and these gradients cause diffusion currents of particles to and from the centre and also set up an outwardly directed heat flow. Furthermore, there is interaction between all the various kinds of particles, not only through elastic and inelastic collisions but also through the absorption and emission of radiation.

A complete physical description of such a complicated situation is not at all easy to produce. The usual approach is therefore to describe systems of interest by models that are simple enough for mathematical treatment yet agree with reality in the essentials. It is evident that the models for high-pressure and low-pressure discharges will be very different. *Table II* shows how widely some of the practical quantities differ for various types of discharge. Note in particular how

Table II. Some characteristic quantities for low-pressure, high-pressure and super-high-pressure discharges.

		Low-pressure	High-pressure	Super-high-pressure
Pressure	(Pa)	100-1000	10^5-10^6	10^6-10^7
Current density	(A/cm^2)	0.1-2	5-50	100-2000
Supplied power / cm	(W/cm)	0.3-3	50-200	1000-8000
Radiated power / cm	(W/cm)	0.2-2	20-100	400-4000
Electron density n_e	(cm^{-3})	10^{12}-5×10^{13}	5×10^{14}-5×10^{16}	5×10^{16}-10^{18}
Gas temperature T_g	(K)	400-1000	4000-7000	5000-9000
Electron temperature T_e	(K)	7000-20 000	4000-7000	5000-9000
Degree of ionization		10^{-4}-10^{-1}		

[3] There are many books on the physics, technology and applications of discharge lamps; we should like to mention in particular:
J. F. Waymouth, Electric discharge lamps, M.I.T. Press, Cambridge Mass. 1971, and:
W. Elenbaas, Light sources, Philips Technical Library, 1972.

[4] A. Bouwknegt, H. Nienhuis, D. J. Schipper and P. A. W. Tielemans, this issue, p. 356.

steeply the supplied and radiated power per centimetre length of discharge increases with increasing pressure. We shall return presently to the considerable difference between electron temperature and gas temperature in the low-pressure discharge.

Models are tried out in conjunction with experiments, of course, which may indicate how to find the model or improve it, verify the chosen model or give the values for the parameters. In view of the complexity of the subject matter, the investigator requires an extensive range of measuring methods; the most important ones are spectroscopy, in both the optical and the microwave bands, and the use of electrostatic probes. The measurements may be made with lamps, but are also made on model systems. The quantities measured are the electric field-strength, the concentration, temperature and velocity distribution of the electrons, the concentrations of the various kinds of particles (particles in the ground state, excited or ionized particles), the gas temperature, the intensity and spectral distribution of the radiation, and all quantities should preferably be measured as a function of place or of time. This subject of plasma diagnostics [5] will not be dealt with here. A method that seems to open up new perspectives — measurements of light scattering due to the plasma — is the subject of a short report in this issue [6]. In the remainder of this section we shall be concerned with a rather brief outline of models used in practice.

High-pressure discharges

The high-pressure discharge [7] is the one whose theory is the least complicated. Owing to the high particles density the energy transport and the exchange of energy (kinetic, excitation, dissociation and ionization energy) between the particles take place almost entirely by way of collisions. This means that *local thermal equilibrium* (LTE) will prevail everywhere in the discharge. We can thus assign to each small volume element a temperature T, which, in accordance with the rules of statistical thermodynamics, completely determines the velocity distributions of all types of particles, the numbers of electrons, atoms (in different excited states) and ions, and also — in the complex case of metal-halide lamps — the direction of the chemical reactions [8]. The only exception in this respect is the radiation field. In *complete* thermal equilibrium the spectral distribution of the radiation field would be a Planckian distribution at the given temperature T, but this is never the case because the radiation field interacts freely with the cold environment of the discharge tube.

The problem thus reduces to that of determining T as a function of r, the distance from the axis. In principle the function $T(r)$ follows from the *energy-balance equation*, which in the simplest case is:

$$\sigma E^2 = \operatorname{div}(-\kappa \operatorname{grad} T) + U(T). \qquad (1)$$

Here σ is the electrical conductivity, E the electric field-strength, κ the thermal conductivity and U the radiated power per unit volume. Equation (1) is merely a formal way of saying that the Joule heat generated is partly dissipated by heat conduction while the remainder is emitted in the form of radiation. If the radiation originates from atoms that return from energy state *2* to state *1*, then U is determined by T through the population n_2 of energy level *2*. After eq. (1) has been solved for the given boundary conditions, the total radiated power can also be calculated. One of the difficulties in this calculation is that σ and κ are themselves functions of T, which must be known from theory or experiment.

In some high-pressure discharges, particles diffuse in the form of atoms from the hot centre to the cooler outer regions, where they recombine to form molecules, and then diffuse back to the interior, where the molecules again dissociate. The dissociation-energy transport thus produced can be an important part of the total energy transport [9]. Excitation and ionization energy may be transported in a similar way. In such cases we must add another term to eq. (1):

$$\sigma E^2 = \operatorname{div}(-\kappa \operatorname{grad} T) + U(T) + \operatorname{div} \sum_i j_{Di} u_i. \qquad (2)$$

Here j_{Di} is the diffusion-current density and u_i the excitation, ionization or dissociation energy of particles of the kind i. To solve eq. (2) we must know the relevant diffusion coefficients as well as σ and κ as a function of temperature.

The greatest theoretical difficulty, however, is found in the radiation term U. In an optically thin plasma — a plasma from which every generated photon $h\nu$ is simply emitted — the term U is equal to $A_{21}n_2h\nu$, where A_{21} is Einstein's coefficient of spontaneous emission. In practice, however — in low-pressure discharge lamps as well — the radiative transport is much more complex, because most of the photons are re-absorbed by atoms, which then jump from state *1* to state *2*; U is therefore only a small fraction of $A_{21}n_2h\nu$. This effect is most pronounced at the centre of a spectral line, and in inhomogeneous optically dense plasmas it produces 'self-reversal' (see fig. 1*b*). The theoretical treatment [10] is often difficult. In the first place, U is the small difference between a large emission term and a large absorption term. Secondly, all parts of the plasma interact with one another through the radiative transport. Thirdly, the spectral distribution of the net emitted power is sensitive to the exact form of the

atomic spectral line and to the probability of photons being extinguished on the way.

All the parameters in the outlined model calculations depend only on r; they have a cylindrically symmetrical distribution and are constant in the axial direction. In fact, deviations from such constancy are clearly observed in some vertically operating metal-halide discharges. In such discharges axial segregation of the gas components may occur as a result of the *convection* in the plasma (not considered here) in combination with diffusion [11].

Low-pressure discharges

In low-pressure discharges the existence of LTE cannot be assumed, which makes it much more difficult to set up a model. In addition to the energy-balance equation we now have a 'balance equation' for every kind of particle. We shall explain this by taking the simplest conceivable case, a 'two-level model'. The model only contains atoms in one state (the ground state), ions and electrons. Owing to electroneutrality, the ion concentration is everywhere equal to the electron concentration (n_e). Assuming a low degree of ionization, we can take the atom concentration n_1 to be constant. We then only need to consider one particle-balance equation, which is the one for the electrons:

$$-D_a \operatorname{div} \operatorname{grad} n_e = k n_1 n_e. \qquad (3)$$

The left-hand side of the equation gives the net number of electrons flowing out of the volume element as a result of ambipolar diffusion (where D_a is the ambipolar diffusion coefficient). This number must be equal to the net number of electrons created by collision ionization, which is proportional to n_e and to n_1 (the right-hand side).

In the usual discussion of eq. (3) it is assumed that k is a constant. In the case of cylindrical symmetry the solution $n_e(r)$ is then a Bessel function of order zero, and from the boundary condition $n_e(R) = 0$ (R is the radius of the tube) a relation is found between k, D_a, n_1 and R ($2.405/R = \sqrt{k n_1/D_a}$). The collision cross-section q — giving the probability that an electron of energy E will ionize an atom on colliding with it — depends, however, on E, and therefore k depends on $f(E)$, the electron-energy distribution:

$$k = \int_0^\infty q(E) f(E) \sqrt{E}\, \mathrm{d}E.$$

If for example the electron gas is hotter in the interior than at the wall, then $f(E)$ is not the same in the interior as at the wall, and therefore k is in general by no means a constant. The two-level model is not very realistic, since it does not even include the production of radiation — the ultimate purpose of the lamp; it does however indicate something that is also true for other models, i.e. that the model is not complete without the knowledge of the collision cross-sections $q(E)$ and the energy-distribution function of the electrons $f(E)$. Multilevel models [12] include several particle-balance equations containing various collision terms. Usually it is only the collisions of the *electrons* with other particles (atoms in different states and ions) that are important, since these are the ones that mainly cause ionization, excitation and de-excitation. The collision cross-sections in these processes are often reasonably well known; it is $f(E)$ that is generally the stumbling block.

In most non-LTE models it is assumed that thermal equilibrium exists between the heavy particles (atoms, ions) and also between the electrons but not between the heavy particles and the electrons. This is a reasonable assumption, since there is a large exchange of energy when two particles of approximately the same mass collide, whereas the energy exchange is very small for colliding particles of very different mass. The function $f(E)$ is then a Maxwell-Boltzmann distribution, determined by one parameter, the electron temperature T_e. We have already seen that T_e and the gas temperature T_g may differ widely (Table II).

If we cannot assume a Maxwell-Boltzmann distribution for the electrons the situation becomes even more complicated. Considerable deviations from this distribution have been observed in neon and argon discharges, for example. These are mainly due to the fact that electrons whose energy is greater than the lowest excitation energy E_1 of the atom interact much more strongly with the gas, through inelastic collisions, than the other electrons, which of course only undergo

[5] The many handbooks on this subject include:
R. H. Huddlestone and S. L. Leonard (ed.), Plasma diagnostic techniques, Academic Press, New York 1965, and:
W. Lochte-Holtgreven (ed.), Plasma diagnostics, North-Holland, Amsterdam 1968.

[6] L. Vriens and M. Adriaansz, this issue, p. 344.

[7] High-pressure discharges are described in W. Finkelnburg and H. Maecker, Elektrische Bögen und thermisches Plasma (Electric arcs and thermal plasma), in: S. Flügge (ed.), Handbuch der Physik, Vol. XXII, Springer, Berlin 1956, pp. 254-444.

[8] Deviations from LTE in the high-pressure sodium discharge are dealt with by J. J. de Groot, J. A. J. M. van Vliet and J. H. Waszink, this issue, p. 334.

[9] E. Fischer, J. appl. Phys. 45, 3365, 1974. See also P. C. Drop, E. Fischer, F. Oostvogels and G. A. Wesselink, this issue, p. 347.

[10] C. van Trigt, Phys. Rev. A 1, 1298, 1970. See also V.V. Ivanov, Transfer of radiation in spectral lines, National Bureau of Standards Special Publication No. 385, 1973.

[11] P. C. Drop, E. Fischer, F. Oostvogels and G. A. Wesselink, this issue, p. 347. Deviations from cylindrical symmetry are also discussed in this article.

[12] See for example J. Polman, H. van Tongeren and T. G. Verbeek, this issue, p. 320, and:
J. J. de Groot, J. A. J. M. van Vliet and J. H. Waszink, this issue, p. 334.

elastic collisions. These deviations in the collision frequencies can be accounted for reasonably well without great mathematical effort by dividing the electrons into two groups, depending on whether their energy is greater or smaller than E_1, each group with its own Maxwell-Boltzmann distribution but with different temperatures [13].

In expressions like eq. (3) the balance for a particular kind of particle is obtained by considering all the processes by which particles appear in a volume element $dxdydz$ or disappear from it. As already mentioned, this requires a knowledge of $f(E)$. If $f(E)$ is not known, a rigorous treatment of the whole problem can be made by setting up the Boltzmann equation for $f(E)$ and solving it as well as possible. Here all the processes that cause a particle to change its velocity, and hence its energy, have to be taken into account. In other words, a detailed description is given of how particles appear in or disappear from a volume element of the *phase space* (x, y, z, v_x, v_y, v_z) [14].

So far we have assumed the discharges to be in a steady state, and in general this is a permissible assumption. Although nearly all lamps are operated from 50 Hz a.c., the relaxation times of the plasma are usually less than a hundredth of a second, which means that the plasma conditions are fairly close to those of the steady state. Studies of the dynamic behaviour of plasmas at higher frequencies, however, have led to interesting conclusions. The relaxation time of the electron gas is found to be much shorter than that for the plasma as a whole, so that there are frequencies at which the current variations are followed by the electron temperature T_e but not by the whole plasma. It is perhaps interesting to note that the luminous efficacy of fluorescent lamps is greater in this frequency range (a few kHz) than at 50 Hz [15].

Technological aspects

Even when we have succeeded in providing a new experimental discharge lamp with a high efficacy and good colour rendering, there remains an important question. How long will these good qualities last? The technological problems that face us here are many and various: many chemicals, some of them very reactive, are used in discharge lamps, and the operating temperatures are often extremely high. If proper precautions are not taken, disastrous reactions may take place between gas, wall and electrodes. In actual fact it is important to ensure that the composition of the gas, the light transmission of the envelope, the electron emissivity of the electrodes and, for fluorescent lamps, the properties of the fluorescent powder, will remain reasonably constant for a long period of time.

Changes in the composition of the gas may be caused by the release of hydrogen. This increases the ignition voltage, since on collision with a hydrogen molecule the energy of an electron is easily lost in translational, rotational or vibrational energy of the molecule. Furthermore, the high thermal conductivity of the hydrogen gas may upset the heat balance inside the lamp. For these reasons the quartz-glass or glass wall of a lamp must be baked at a sufficiently high temperature during manufacture. If this is not done the wall will later give off water: this dissociates, the oxygen combines with material such as the tungsten of the electrodes, and the highly undesirable hydrogen remains behind in the gas. The space between the inner and outer envelopes of a high-pressure lamp must also be free from hydrogen, which would otherwise diffuse very quickly into the discharge volume because of the high temperature of the gas-discharge wall. The outer enclosure of a low-pressure sodium lamp must in fact be evacuated to give good heat insulation. Since it is almost impossible to make the outer envelope completely free from water, the hydrogen produced in the space between the gas-discharge tube and the outer envelope is removed by getters [16].

Sodium readily reacts with glass or quartz glass. Although such a reaction would remove sodium from the gas, this would not matter in practice, since there is always a surplus of sodium (condensate) present. However, sodium compounds are formed in the wall, and in the long run these cause impermissible blackening. For this reason it was not possible to produce sodium lamps before sodium-resistant wall materials had been developed. The low-pressure sodium lamp is made with soda-lime glass coated with a thin borate-glass layer that sodium ions cannot penetrate. In the high-pressure sodium lamp vacuum-tight walls of translucent sintered alumina (Al_2O_3) are found to be satisfactory. The sodium in the low-pressure lamp, incidentally, itself acts as a getter, in particular for the water given off by the wall.

A light-absorbing layer may gradually form *on* the wall as a result of the transport of electrode material, e.g. tungsten, to the wall. This transport is often caused by electrode evaporation or sputtering, but it may also be due to chemical reactions resulting from contaminants in the gas.

Finally, the light transmission of the wall may also be impaired by recrystallization. The required vitreous

[13] L. Vriens, J. appl. Phys. 44, 3980, 1973.
[14] See for example I. P. Shkarofsky, T. W. Johnston and M. P. Bachynski, The particle kinetics of plasmas, Addison-Wesley, Reading Mass. 1966.
[15] J. Polman, H. van Tongeren and T. G. Verbeek, this issue, p. 320.
[16] G. Kuus, this issue, p. 354.

state is metastable, and in principle the transition to the crystalline state is a question of time. Normally the recrystallization process is negligible, but in the super-high-pressure lamps, for example, the very high temperatures prevailing may make recrystallization the life-determining factor. In some halogen lamps the process is catalysed by constituents of the gas.

The most important requirement for the electrodes, which act alternately as cathode and anode when connected to an a.c. supply (the usual case), is to maintain the electron emissivity. For this purpose it is essential to prevent chemical corrosion of the emitting surface. Here again, water is disastrous, causing the formation of tungsten oxide, which ultimately covers the surface. To keep the electrode emissive while at the same time avoiding evaporation, it is necessary to keep the temperature within a fairly narrow range. The temperature is determined by Joule heating in the electrode itself, by electron and ion bombardment, thermal conduction and radiation, and is therefore sensitive to the geometry. Chemical corrosion may alter the geometry and in this way indirectly impair the electrode operation. A change in the geometry of the electrodes may also affect the gas filling by raising or lowering the temperature of the coldest spot in the tube, which of course determines the vapour pressure of the components that partly remain condensed when the lamp is operating.

Summary. After a brief description of the low-pressure and high-pressure sodium lamp, the high-pressure mercury lamp, the super-high-pressure mercury and xenon lamp, fluorescent discharge lamps and metal-halide lamps, an account is given of research now in progress in this field. Considered in detail, gas discharges are highly complex phenomena, which are studied by means of many diagnostic aids combined with model calculations. High-pressure-discharge models presuppose local thermal equilibrium (LTE). A considerable problem in the theoretical treatment is the effect of radiative transport in an optically dense plasma. In low-pressure discharges there is no LTE, and the nature of the energy-distribution function of the electrons becomes important. If this is a Maxwell-Boltzmann distribution, a temperature can be assigned to the electrons, which, in low-pressure discharges, will be much higher than the gas temperature. At high frequencies it may happen that the electron gas is in a state of equilibrium at every instant whereas the plasma as a whole is not. Technological research aimed at improving the useful life of the lamps is primarily concerned with preventing interactions between filling gas, wall and electrodes from causing changes in gas composition, the light transmission of the wall and the electron emissivity of the electrodes.

Carbon-filament lamps with a chemical transport cycle

At the beginning of this century nearly all electric lamps had a carbon filament, but by about 1910 the carbon filament was almost entirely superseded by the tungsten filament. Carbon-filament lamps of the traditional type are now only produced on a limited scale, for special purposes. The main reason for the switch to tungsten was that it has a much slower rate of evaporation than carbon at the same temperature. This means that under the same operating conditions, the bulb of a tungsten-filament lamp does not blacken nearly as quickly as that of a carbon-filament lamp. The difference between carbon and tungsten can also be seen from *fig. 1*, which shows their vapour-pressure curves.

Nevertheless the carbon-filament lamp does have its virtues. In principle a carbon filament can operate at much higher temperatures than a tungsten filament, because carbon sublimes at about 4100 K [1] whereas tungsten melts at a temperature four hundred degrees lower. *Fig. 2* shows the temperature dependence of the luminance and luminous efficacy for both carbon and tungsten. At the same temperature, the efficacy of the carbon-filament lamp is rather worse than that of the tungsten lamp, but the carbon filament has a greater luminance. For applications requiring high luminance, carbon would therefore seem to have the advantage over tungsten. Another reason for taking a new look at the carbon-filament lamp is that much more information is now available about the various kinds of carbon and the ways in which they can be manufactured and shaped. Finally there is the question of possible shortage of materials; there is some risk of this with tungsten, but not with carbon.

A primary object of the investigation — to which we shall confine ourselves here — was to find a way of preventing the bulb from blackening at filament temperatures considerably higher than those used in the traditional incandescent lamp (2100 K) while maintaining an acceptable life.

We tried to see whether this could be done by filling the bulb with a gas that reacts with carbon and functions in much the same way as iodine or bromine in

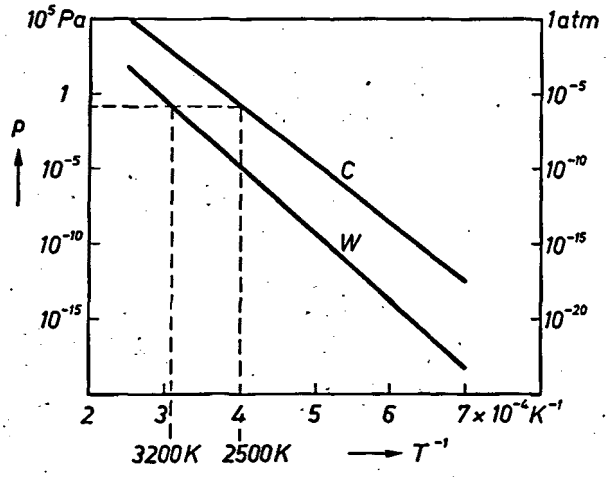


Fig. 1. The vapour pressure p of carbon and of tungsten, as functions of the reciprocal of the absolute temperature T. The vapour pressure of carbon is a summation over the vapour pressures of the molecular types C_2, C_3, C_4, C_5 and atomic carbon [1]. The vapour pressure of tungsten at 3200 K is approximately equal to that of carbon at 2500 K. The evaporation rate of the materials is directly proportional to their vapour pressure.

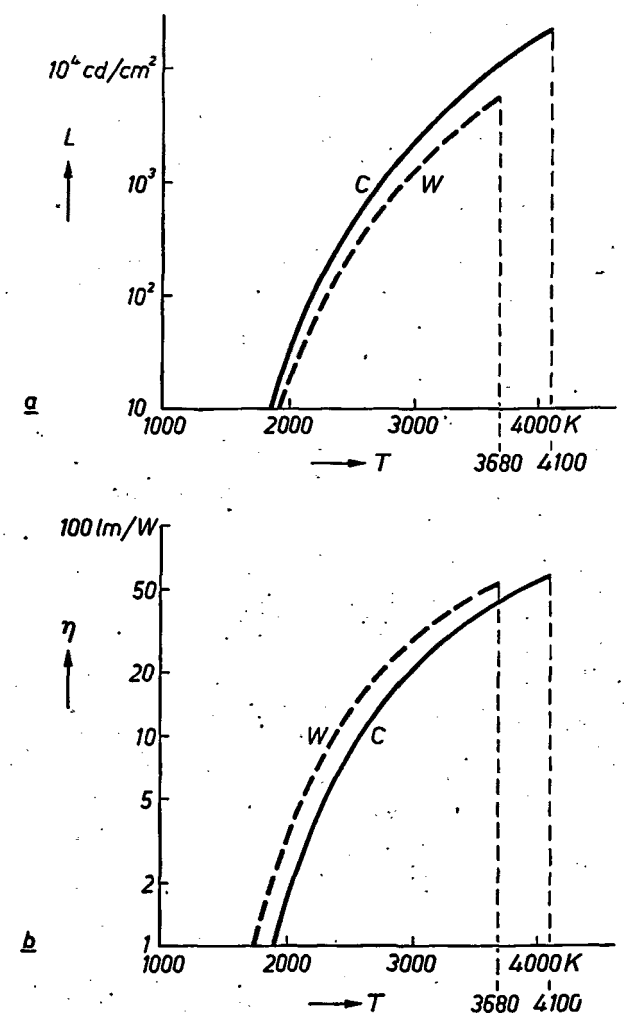


Fig. 2. *a*) The luminance L for carbon and for tungsten as a function of absolute temperature T. Carbon sublimes at about 4100 K. Tungsten melts at 3680 K. *b*) As (*a*), for the luminous efficacy η.

halogen lamps [2]. A gas of this kind would have to convert carbon into stable gaseous compounds at the bulb wall and close to it. These compounds would then be transported by diffusion and convection to the filament, where the carbon would be released again because of dissociation due to the high temperature — and in the ideal case would completely regenerate the filament. The 'transport gas' released on dissociation would then diffuse to the bulb wall again. A lamp in which a mechanism of this type is operative can operate at higher temperatures, giving a higher luminance and efficacy.

The occurrence of a chemical transport cycle and its properties can in principle be predicted for any gas

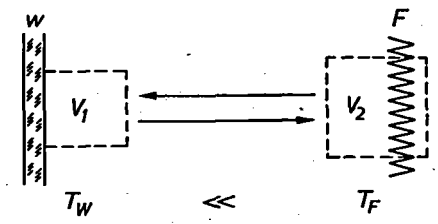


Fig. 3. Model of a (chemical) transport reaction in a carbon-filament lamp. *F* filament. *W* bulb wall. The two volumes V_1 and V_2 each contain solid carbon and gas. The temperature in both volumes is assumed to be uniform (T_W is about 500 K, T_F is 2500-3000 K). It is assumed that both volumes contain equal numbers of bonded or free atoms of the reactive element. After exchange of the gas in the two volumes the original composition of the gas phase is restored. In this process an effective transport of carbon takes place from the solid phase in the one volume to that in the other.

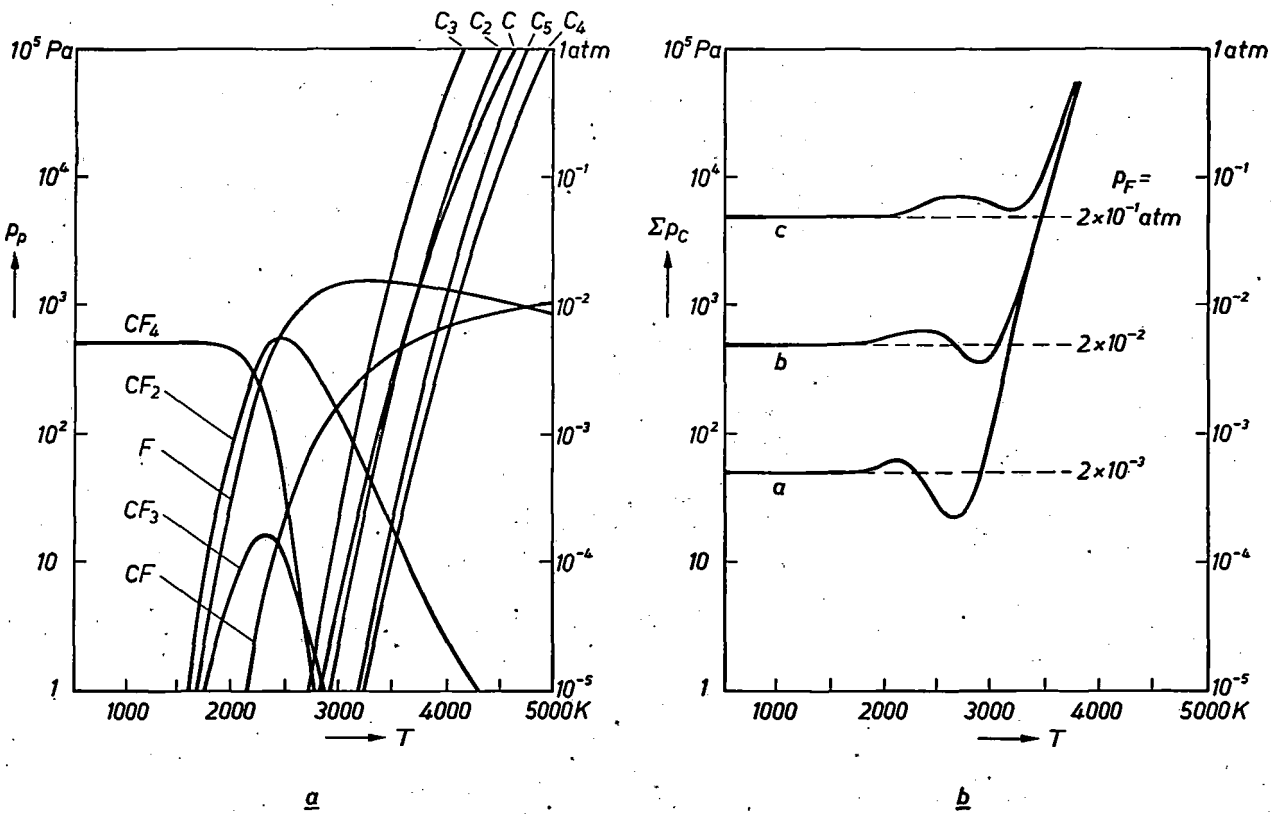


Fig. 4. *a*) Calculated composition of the gas in the heterogeneous system carbon-fluorine as a function of absolute temperature *T*. The partial pressure p_p of the compounds formed is plotted vertically. $C_2, \ldots, C_5$ are carbon molecules; *C* and *F* are atomic carbon and atomic fluorine respectively. The calculations were based on a total quantity of fluorine corresponding to a constituent pressure of 2×10^3 Pa. No data relating to kinds of molecule at pressures lower than 10 Pa are included. *b*) Calculated carbon total vapour pressure Σp_C of the heterogeneous system carbon-fluorine as a function of *T*. The parameter used is the total number of bonded or free fluorine atoms, characterized by the constituent pressure p_F. Curve *a*: a lamp with this variation of Σp_C will still have a clear bulb, at least at filament temperatures between 2300 and 2900 K, since Σp_C is then lower near the filament than at the wall; for curve *b* somewhat higher temperatures are possible, up to 3000 K. In the case of curve *c* carbon transport goes towards the bulb wall at all filament temperatures above 2000 K.

from chemical thermodynamics. The model we used is shown in *fig. 3*. It is assumed that there is a film of solid carbon on the bulb wall. The essential assumption is that there is thermodynamic equilibrium both at the wall and at the filament. The composition of the gas at

[1] The various aspects of the evaporation of carbon are dealt with by H. B. Palmer and M. Shelef in P. L. Walker, Jr. (ed.), Chemistry and physics of carbon 4, 85-135, Dekker, New York 1968.

[2] See for example J. H. Dettingmeijer, G. Dittmer, A. Klopfer and J. Schröder, this issue, p. 302.

a given temperature can then be calculated from the available thermodynamic data relating to the (heterogeneous) equilibrium between the solid carbon and the reactive transport gas.

If there are many kinds of molecule the composition of the gas will be rather complex. For instance, in the case of the heterogeneous system carbon-fluorine, which we shall now consider, we have to contend with the molecules C_2, C_3, C_4, C_5, CF, CF_2, CF_3, CF_4, C_2F_2, C_2F_4, C_2F_6 and F_2, as well as the atoms C and F. The composition of the gas phase of this system is given as a function of temperature in *fig. 4a*.

From the curves in fig. 4*b* the direction of carbon transport — towards the bulb or away from it — can be determined for a given temperature distribution in the lamp. Here we assume that two volumes of gas with equal numbers of bonded or free fluorine atoms are exchanged between the region near the relatively cool bulb and the region near the much hotter filament. (In reality the exchange takes place by diffusion and convection.) After the exchange the original equilibrium is restored, which means that the composition of the gas phase is now the same as it was before the exchange. In proportion to the difference in total vapour pressure Σp_C that can be read off in fig. 4*b* between the two exchanged volumes, there must have been an effective transport of solid carbon from the one volume to the other. The transport always goes from a region where the carbon total vapour pressure is high to a region where it is low, and carbon is therefore deposited from the gas phase in this region.

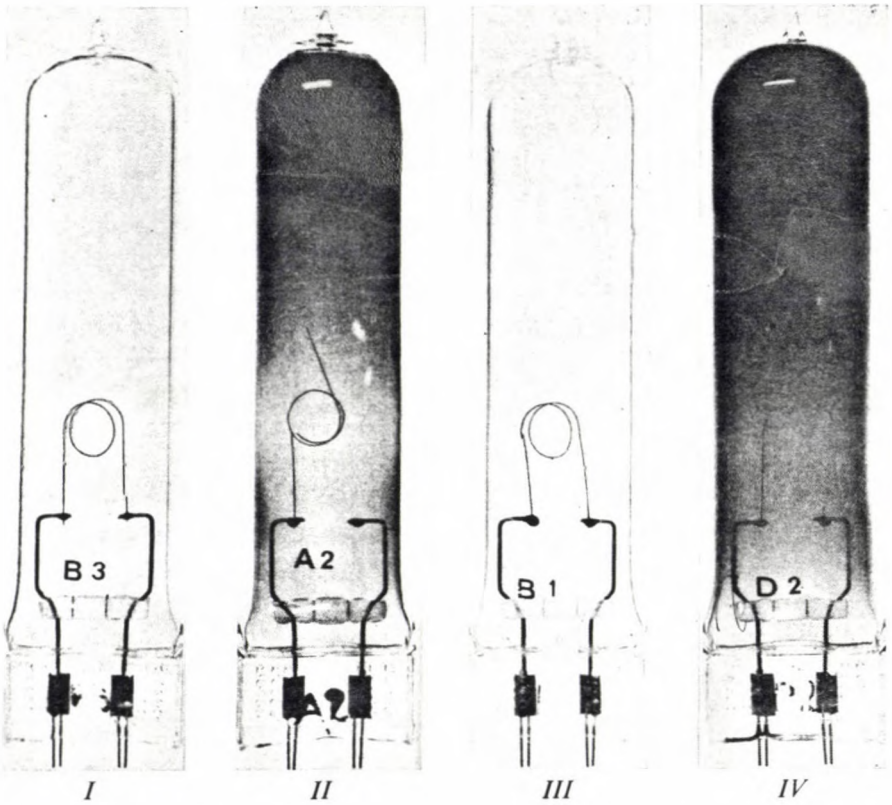


Fig. 5. Blackening of the bulb in four experimental carbon-filament lamps with a carbon-fluorine transport cycle. *I* gas filling (at room temperature) 65×10^3 Pa of Xe + 1.3×10^2 Pa of CF_4 (see curve *a* in fig. 4*b*). *II* gas filling 65×10^3 Pa of Xe *without* CF_4. When the photograph was taken, both lamps had operated for six hours at a filament temperature of 2750 K. *III* gas filling as in *I*. *IV* gas filling 53×10^3 Pa of Xe + 1.3×10^4 Pa of CF_4 (see curve *c* in fig. 4*b*). Lamp *III* has operated for fifty hours at 2560 K, lamp *IV* for only two minutes.

The extent to which a gaseous medium takes up the carbon in a bonded or free form may be expressed by a kind of total vapour pressure (Σp_C). This is done by summing the partial pressures of the gaseous carbon compounds, taking into account the number of carbon atoms present in the various kinds of molecule. Fig. 4*b* shows a plot of Σp_C for the carbon-fluorine system as a function of temperature. The parameter p_F is the constituent pressure, a quantity that characterizes the total number of bonded or free fluorine atoms.

A lamp with a gas filling corresponding to curve *a* in fig. 4*b* will still have a clear bulb at filament temperatures between 2300 and 2900 K, i.e. at temperatures

near the minimum in the curve. If the fluorine content is high enough for curve *c* to apply, the mechanism will transport carbon to the bulb, and the lamp will not be usable. Our experimental results agree with these predictions (*fig. 5*) [3].

A life of several hours has already been achieved with experimental carbon-filament lamps operated at a filament temperature of 3000 K, and a life of the order of several tens of hours has been reached with a filament temperature of 2500 K. These are promising results.

The calculations have also shown that a transport cycle that keeps the bulb clear is feasible not only with fluorine but also with chlorine or hydrogen. This has been confirmed in the meantime by experiments [4].

W. J. van den Hoek
W. A. Klessens

[3] W. J. van den Hoek and W. A. Klessens, High Temp. Sci. 7, 215, 1975.
[4] W. J. van den Hoek and W. A. Klessens, Carbon 13, 429, 1975.

Dr Ir W. J. van den Hoek and Ing. W. A. Klessens are with the Philips Lighting Division, Eindhoven.

Philips tech. Rev. 35, 320, 1975, No. 11/12

Open-air laboratory for road lighting

A short piece of roadway at a Philips establishment at Acht, near Eindhoven, has been arranged as a lighting laboratory. The road surface is 250 m long by 17 m wide, and the equipment provided can be used for investigating the performance of experimental lanterns and the associated optical equipment. The 'lamp posts' can be moved on rails. The lanterns can be mounted at any height up to 16 m above the road surface, and their position in the horizontal plane can also be varied.

Low-pressure gas discharges

J. Polman, H. van Tongeren and T. G. Verbeek

The low-pressure gas discharge, whose behaviour we shall describe here, is the basis of two efficient and therefore widely used light sources: the low-pressure sodium lamp and the fluorescent lamp. Some data relating to these two lamp types are shown in *Table I*.

The radiation from the sodium lamp consists almost entirely of the resonance radiation of the sodium atom (see *fig. 1*), whose wavelength is close to that of the maximum in the eye-sensitivity curve (the spectral-luminous-efficiency curve). It is partly for this reason that the sodium lamp is the light source with the highest luminous efficacy [1]. Since it emits light of virtually a single wavelength, it gives no colour rendering.

In the fluorescent lamp a mercury discharge is used. This discharge yields mainly ultraviolet radiation (fig. 1), which is converted into visible light by a coating of fluorescent powder on the tube wall [2]. By carefully choosing the wavelengths at which the powder fluoresces, very good colour rendering can be obtained [3].

In this article we shall confine ourselves to discussing the region of the gas discharge that is responsible for light generation: the positive column [4]. Other regions of the discharge are discussed elsewhere in this issue, as are also the problems connected with the electrodes [5].

The number of processes that come into play in the positive column and the extent to which these processes interact with one another make the whole picture rather unclear. Numerical calculations based on a sufficiently simple theoretical model, however, allow studies to be made of the effects of changes in the various quantities of interest. This not only provides a greater understanding of the properties of a discharge, but also enables predictions to be made about the consequences of changing one or more of its parameters. Estimates can also be made of the likely behaviour of discharges that differ considerably from those now used.

First of all we shall describe the theoretical model we have used for the positive column [6-11]. Then we shall look more closely at the development of the model and at comparisons with experiment, both for the fluorescent lamp and for the low-pressure sodium

Dr J. Polman is with Philips Research Laboratories, Eindhoven; Dr Ir. H. van Tongeren, formerly with Philips Research Laboratories, is now with the Philips Electronic Components and Materials Division (Elcoma), Eindhoven; Ir T. G. Verbeek is with the Philips Lighting Division, Eindhoven.

Table I. Some data for two types of low-pressure discharge lamp: the 40-W fluorescent lamp and the 180-W low-pressure sodium (SOX) lamp.

	Fluorescent 40 W	SOX 180 W
Diameter of discharge tube	36 mm	19 mm
Wall temperature	about 40 °C	about 260 °C
Current (r.m.s.)	430 mA	940 mA
Buffer gas	75% argon 25% neon	99% neon 1% argon
Buffer-gas pressure	350 Pa (2.5 torr)	600 Pa (4.5 torr)
Luminous flux	3200 lm	33 000 lm
Efficacy	80 lm/W	183 lm/W
Colour temperature	4200 K	—
General colour rendering (for 'colour 33')	65	—

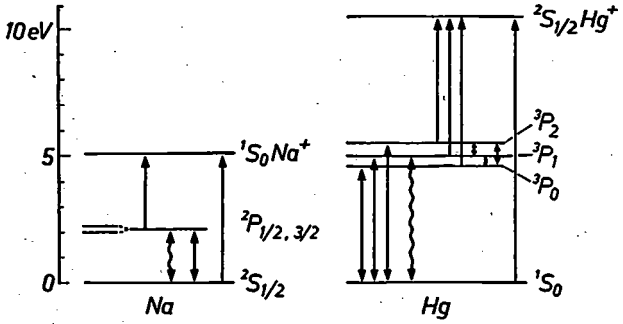


Fig. 1. Energy-level diagrams for sodium and mercury; only the principal levels for the low-pressure discharges are shown. The transitions indicated by arrows are due to the interaction of atoms with electrons in the discharge. A return from the upper level of the ground state (recombination of ion and electron) only takes place at the wall of the discharge tube. The two D levels of sodium are so close together that they may be regarded as a single level for the treatment of the collision processes in model calculations. The differences in population between the two levels do have to be taken into account, however, in describing the radiative decay. The wavelengths of the emitted radiation are 589 and 590 nm for sodium, and 254 nm for mercury.

[1] See Table I in the article by R. Bleekrode, M. Koedam and L. Rehder, this issue, p. 308.
[2] G. Blasse and A. Bril, Philips tech. Rev. 31, 304, 314 and 324, 1970.
[3] J. J. Opstelten, D. Radielović and J. M. P. J. Verstegen, this issue, p. 361.
[4] A general survey of the physics and technology of discharge lamps is given in J. F. Waymouth, Electric discharge lamps, M.I.T. Press, Cambridge, Mass. 1971. A review of recent research on low-pressure gas discharges is given by J. Polman in Proc. 12th Int. Conf. on Phenomena in ionized gases, Eindhoven 1975, to be published shortly.
[5] A. Bouwknegt, H. Nienhuis, D. J. Schipper and P. A. W. Tielemans, this issue, p. 356.
[6] M. A. Cayless, Brit. J. appl. Phys. 14, 863, 1963.
[7] J. Polman and P. C. Drop, J. appl. Phys. 43, 1577, 1972.
[8] J. Polman, J. E. van der Werf and P. C. Drop, J. Physics D 5, 266, 1972.
[9] P. C. Drop and J. Polman, J. Physics D 5, 562, 1972.
[10] H. van Tongeren and J. Heuvelmans, J. appl. Phys. 45, 3844, 1974.
[11] H. van Tongeren, Thesis, Eindhoven 1975 (also published as Philips Res. Repts. Suppl. 1975, No. 3).

lamp. It will be shown that the model gives a satisfactory description of the discharge in some cases. Finally we shall indicate some prospects for the further development of low-pressure discharge lamps.

Model for the positive column

The positive column of a low-pressure gas discharge contains electrons, ions and uncharged particles in an excited or non-excited state. The concentrations of electrons and ions are equal and much lower than the concentrations of the uncharged particles. In our model we assume that all the concentrations are constant in the axial direction and radially symmetrical. The electrons are mainly responsible for charge transport; the contribution from the ions is small because of their low mobility. Owing to the low pressure the interaction between the electrons and the heavier particles is insufficient, and the electric field is also too high to permit an energy equilibrium to be established between the two groups. Equilibrium does exist, however, within each group, and we may therefore speak of a gas temperature and an electron temperature. The electron temperature is much higher than the gas temperature; the difference may be a factor of ten or more. The pressure in the discharge, however, is sufficiently high for the mean free path of all particles to be much smaller than the dimensions of the discharge vessel.

Collisions of electrons with uncharged particles can give rise to ionization, mostly via an intermediate excited state, thus generating new electrons and ions. These particles disappear as a result of recombination; in the cases considered here this is assumed to take place solely at the wall of the discharge tube. Ions and electrons reach the wall by diffusion; the two groups of particles are not independent of each other but are linked by electrostatic forces. This diffusion process may therefore be described by a single parameter, the ambipolar diffusion coefficient. It will be shown that the behaviour of the discharge is to a great extent determined by ambipolar diffusion followed by recombination at the wall. In the steady state the ionization and recombination must be in equilibrium with one another. Ambipolar diffusion and the diffusion of uncharged particles also come into play here. The velocity distribution of the electrons must therefore be such that the ionization exactly compensates the loss of charge carriers by diffusion. This velocity distribution can be described to a good approximation by a Maxwell distribution. A distribution of this type is characterized by a single temperature, which means that this temperature, the electron temperature, is partly determined by the ambipolar diffusion. The approximation for the energy distribution of the electrons can be improved by introducing a modification on the high-energy side [12]. A more exact description of the energy distribution can be obtained by solving Boltzmann's equation [13]. This is no easy matter, however, and numerical calculations require a great deal of computer time, particularly for the time-dependent case.

The important role of diffusion in this whole picture implies that the properties of a discharge can be radically changed by altering the diffusion. This is usually done by adding an inert gas to the filling of a discharge lamp. Because of collision processes the inert gas reduces the speed of the diffusion process without itself taking part in the discharge. The inert gas added also reduces the mobility of the electrons in the discharge and consequently has an effect on the lamp voltage required for a particular current. In this way a discharge lamp can be adapted to the available supply voltage. The usual inert-gas addition in a sodium lamp is a mixture of 99% neon and 1% argon [14]; in a fluorescent lamp a mixture of 75% argon and 25% neon would be used.

A complication is that the resonance radiation cannot emerge directly from the discharge, but only after repeated absorption and re-emission. This radiation transport is a very complex process [15]; it is one of the main problems in model calculations, and as yet it has not been completely solved.

The transport of particles and energy in a discharge is effected by diffusion, ionization, excitation and radiation. The model for the positive column consists of a number of equations that describe these processes. To begin with, there are continuity equations for the various particle concentrations, then there is the energy-balance equation for the electrons and finally the equation that gives the relation between current and field-strength in the column. The equations are found to contain one unknown more than the number of equations. The equations can be solved, however, by imposing an extra boundary condition, e.g. by specifying the value of the discharge current.

Solving the set of equations gives the voltage across the column and the concentrations of electrons, ions and excited and non-excited uncharged particles. From the concentration of the excited particles we can calculate the radiation output of the discharge as a function of the power supplied to it.

If we confine ourselves to the description of a sodium discharge and assume only one excitation level in the calculations, the model outlined above may be expressed by the equations given below. To start with, we have the continuity equations for the concentrations of uncharged atoms (n_0), excited atoms (n_1), ions (n_i) and electrons (n_e). The last two equations are identical since the plasma is assumed to be neutral ($n_i = n_e$). The equations

giving the dependence of these quantities on the time t are then:

$$\frac{\partial n_0}{\partial t} = D_0 \nabla^2 n_0 - k_{01} n_0 n_e - k_{01} n_0 n_e + k_{10} n_1 n_e + \frac{n_1}{\tau_{eff}}, \quad (1)$$

$$\frac{\partial n_1}{\partial t} = D_1 \nabla^2 n_1 + k_{01} n_0 n_e - k_{11} n_1 n_e - k_{10} n_1 n_e - \frac{n_1}{\tau_{eff}}, \quad (2)$$

$$\frac{\partial n_i}{\partial t} = \frac{\partial n_e}{\partial t} = D_a \nabla^2 n_e + k_{01} n_0 n_e + k_{11} n_1 n_e. \quad (3)$$

In these expressions D_0 is the diffusion coefficient for atoms, D_1 the diffusion coefficient for excited atoms and D_a the ambipolar diffusion coefficient. Owing to the radial symmetry, all that is left of the operator ∇^2 is the expression

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right).$$

The parameters k_{jk} are the reaction-rate coefficients for the transition from state j to state k resulting from the collision of a particle in state j with an electron. They can be calculated by integrating the appropriate collision cross-section σ_{jk} over all relevant electron velocities:

$$k_{jk} = \int v\sigma_{jk} f(v)\, dv,$$

where $f(v)$ is the distribution for the electron velocities. In solving the continuity equations the following boundary conditions are taken: for all particles $\partial n/\partial r = 0$ for $r = 0$ and all values of t owing to the radial symmetry; also, $n_i(R,t) = n_e(R,t) = 0$, and $n_0(R,t) = n_w$, where R is the radius of the discharge vessel and n_w the concentration of the uncharged atoms at the wall. Finally, τ_{eff} is the effective lifetime of the excited state. The effective lifetime may be many times greater than the natural lifetime of the excited state, because the resonance radiation is 'imprisoned' in the plasma by the repeated absorption and re-emission. The extent to which this is the case, and hence the radiative transport in the discharge, depends on the population in the ground state (n_0) and in the excited state (n_1). All this leads to rather complicated calculations [15].

The equation for the conservation of electron energy is:

$$\int_0^R \frac{\partial}{\partial t}[n_e(r,t)\tfrac{3}{2}kT_e]\,2\pi r dr =$$

$$\tfrac{5}{2}kT_e D_a \left(\frac{\partial n_e}{\partial r}\right)_{r=R} 2\pi R + EI - \int_0^R (P_{el} + P_{inel}) n_e(r,t)\, 2\pi r dr,$$

where k is Boltzmann's constant and P_{el} and P_{inel} are the elastic and inelastic losses per electron and per second. This equation, together with the relation between current and field-strength:

$$I(t) = E(t)\, e \int_0^R n_e(r,t)\mu_e(t)\, 2\pi r dr,$$

permits the power taken up by the column to be calculated, always provided, as mentioned earlier, that an additional boundary condition is specified, e.g. in the form of a given discharge current. The product $\mu_e E$, where μ_e is the electron mobility [13], in the last equation gives the drift velocity v_d for the electrons in the discharge. The concentration n_1 of the excited atoms and the effective lifetime τ_{eff} of this state then allow the light output for this power to be determined.

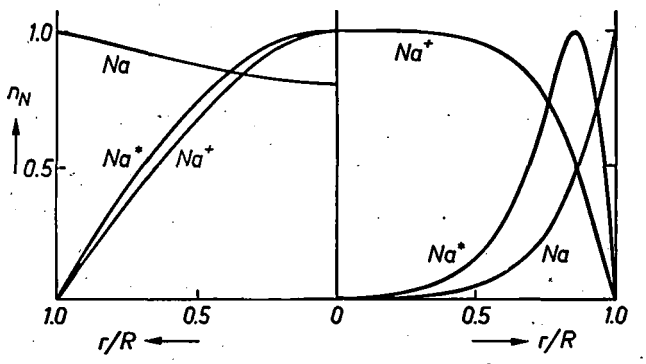


Fig. 2. The normalized particle concentrations n_N as a function of the normalized radius r/R in a sodium discharge. The curve *Na* gives the concentration of the uncharged atoms, *Na** that of the excited atoms and *Na+* that of the ions. The concentrations have been normalized by dividing by the maximum concentrations of the appropriate particle so that all three curves can be shown on the same figure. The left-hand half of the figure shows the situation for a low discharge current, the right-hand half the situation for a high discharge current. The diameter of the discharge tube is 19 mm, the wall temperature is 260 °C. See also Table II.

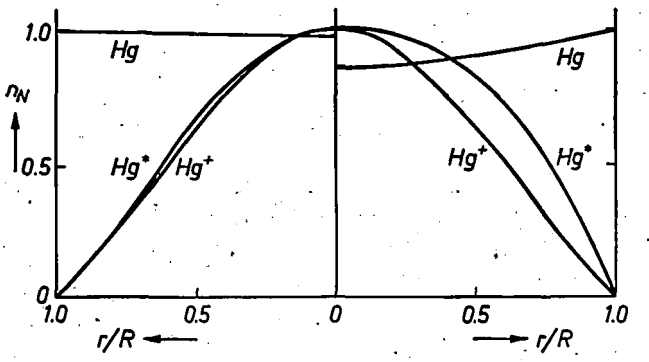


Fig. 3. Normalized particle concentrations in a mercury discharge. The representation is the same as in the previous figure. The diameter of the discharge tube in this case is 36 mm, the wall temperature 42 °C. See also Table II.

Results of model calculations

Results of numerical calculations based on the model described above are given in *fig. 2* and *fig. 3*, which show the concentrations of the various kinds of particle as a function of the radius in a sodium and in a mercury discharge under practical conditions for these discharges. The figures give relative values for the various quantities; some numerical values are listed in *Table II*.

The differences in behaviour between the two discharges are mainly explained by the differences in current density. It can be seen in the case of the sodium discharge that at higher current densities the concentration of Na atoms at the axis drops virtually to zero. The decrease in the concentration of the uncharged

[12] L. Vriens, J. appl. Phys. 44, 3980, 1973.
[13] I. P. Shkarofsky, T. W. Johnston and M. P. Bachynski, The particle kinetics of plasmas, Addison-Wesley, Reading Mass. 1966.
[14] The use of an inert-gas mixture causes the ignition voltage to drop below the value for a pure inert gas. See F. M. Penning, Z. Physik 46, 335, 1928.
[15] T. Holstein, Phys. Rev. 72, 1212, 1947, and 83, 1159, 1951. C. van Trigt, Phys. Rev. 181, 97, 1969, and A1, 1298, 1970. See also M. J. C. van Gemert, J. appl. Phys. 46, 4899, 1975.

Na atoms near the axis of the discharge is referred to as 'depletion' [16].

In a mercury discharge the concentration distributions show little variation when the current density is varied. For the mercury discharge it is therefore a reasonable approximation to take the concentration of atoms in the ground state as constant; the concentrations of ions and excited atoms as a function of the radius can be described to a good approximation by Bessel functions of order zero [6]. When introducing these approximations the model equations can simply be integrated over the radius, which then leaves equations that only have a time dependence. This makes it easier to study time-dependent processes and reduces the computer time required for the numerical calculations. A typical calculation of this kind will be presented in the next section.

It is only when $2\pi/\omega$ lies between these values that the current modulation can produce a periodic change in the electron temperature T_e, thus causing, as we shall show, a distinct change in the behaviour of the discharge. If $2\pi/\omega$ is much smaller than the characteristic times, the discharge is too slow to follow the current variations and T_e will adjust to a level corresponding to the average current I_0. If, however, $2\pi/\omega$ is much greater than the characteristic times, we then have what is in fact a slowly varying d.c. discharge. Here again there will be hardly any variation in T_e, since only slight changes in T_e can cause current variations, because of the exponential relation existing between T_e and the ionization rate.

In the intermediate region, i.e. for $\tau_a > 2\pi/\omega > \tau_e$, a modulated discharge current may be expected to cause a modulation of the electron temperature, for the charge-carrier concentration in the plasma will not then follow the current modulation, since $2\pi/\omega$ is less than τ_a. This implies that current variations can only occur as a result of variation of the drift velocity of the electrons, which in turn means that the field-strength in the plasma must also be subject to these variations. In this case the electron temperature will follow the field-strength variations, because $2\pi/\omega > \tau_e$.

Table II. Some numerical results of calculations on low-pressure discharges in a mixture of sodium and inert gas and a mixture of mercury and inert gas.

	Mercury discharge		Sodium discharge	
Current	140 mA	1000 mA	70 mA	700 mA
Electron concentration on the axis	2×10^{18} m^{-3}	2×10^{19} m^{-3}	5×10^{18} m^{-3}	4.5×10^{19} m^{-3}
Tube diameter	36 mm		19 mm	
Wall temperature	42 °C		260 °C	
Concentration of uncharged atoms at the wall	2.2×10^{22} m^{-3}		0.45×10^{20} m^{-3}	
Buffer gas and buffer-gas pressure	Argon 400 Pa (3 torr)		Neon 730 Pa (5.5 torr)	

The time-dependent behaviour of a low-pressure mercury-inert-gas discharge

The time-dependent behaviour of the mercury-inert-gas discharge used in a fluorescent lamp is not only relatively easy to describe but is also of practical importance, since lamps of this type are almost invariably operated from an a.c. supply. In order to continue to benefit from the agreement between the theoretically and experimentally known d.c. behaviour, we started by studying a modulated d.c. discharge in which the current I through the discharge can be described as a function of time t by $I = I_0(1 + \beta \sin \omega t)$. Here I_0 is the value of the d.c. current, β the modulation coefficient and ω the angular frequency of the modulation.

The behaviour of the discharge is determined by the duration of the period $2\pi/\omega$ of the modulation in comparison with the two characteristic times that enter into the description of the discharge [7] [8] [17]: the ambipolar diffusion time τ_a and the relaxation time τ_e for the electron energy. Representative values of these times are about 1 ms for τ_a and 5 μs for τ_e.

The far-reaching consequences that result from such a periodic variation of T_e may be understood from the reasoning summarized in *fig. 4* together with the results of numerical calculations and experimental measurements. There is reasonable agreement between measurements and calculations as long as the modulation coefficient β does not exceed 0.5 (*fig. 5*). The calculations indicate that a further increase in β should give a further decrease in the concentration of excited atoms, but the measurements show that the concentration in fact increases. An explanation for this disparity between theory and experiment can be found by studying the effects of time-dependent deviations in

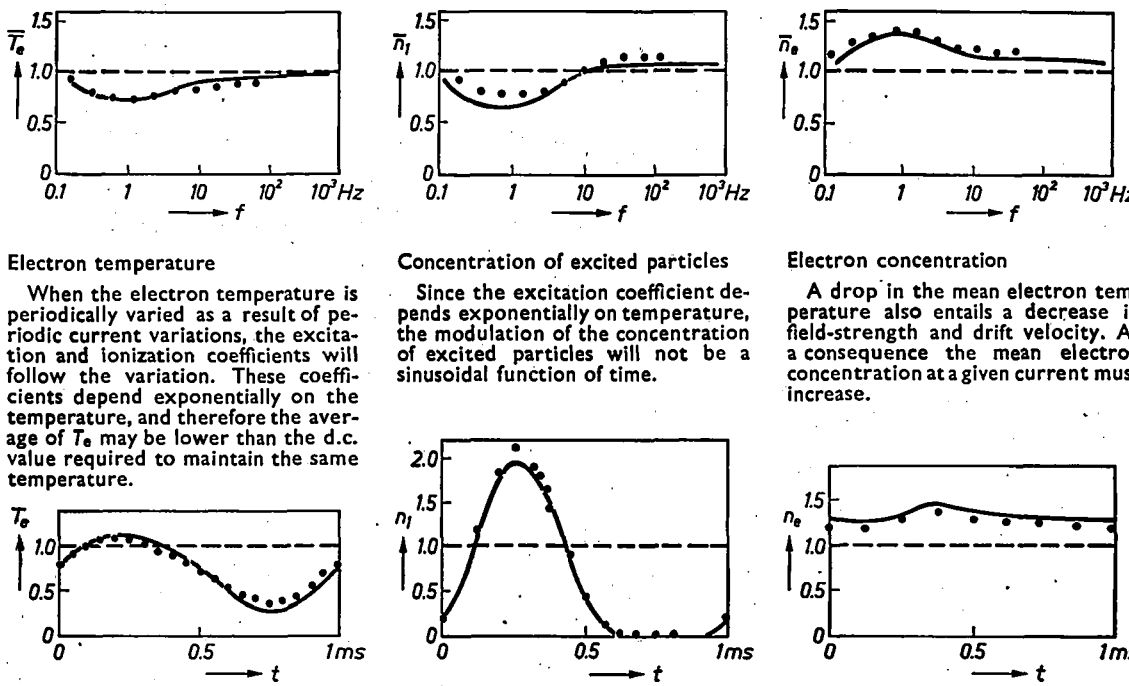


Fig. 4. Periodic variation of the electron temperature T_e and its consequences in a modulated mercury-neon discharge (modulation coefficient $\beta = 0.5$). The solid lines were calculated, the points give the results of measurements. *Left:* above, the average value $\bar{T}_e$ of the electron temperature as a function of the modulation frequency f of the discharge current; below, a period of T_e at a modulation frequency of 1 kHz. In both cases the dashed line, as in the following figures, gives the value of the parameters for a d.c. discharge at the same mean current. *Centre:* above, average concentration $\bar{n}_1$ of the excited atoms as a function of supply frequency; below, curve of the value n_1 of this concentration during one period at a frequency of 1 kHz. *Right:* above, average value $\bar{n}_e$ of the electron concentration as a function of the frequency; below, curve of the value n_e of this concentration during one period at 1 kHz. The diameter of the discharge tube was 5.2 cm, the current in the d.c. discharge 0.4 A; the particle densities in this discharge were $n_{0Hg} = 2 \times 10^{20}$ m^{-3}, $n_1 = 6 \times 10^{17}$ m^{-3}, $n_{Ne} = 2 \times 10^{23}$ m^{-3} and for each of the various $^3P_1$ states about 5×10^{17} m^{-3}.

the velocity distribution of the electrons [18]. Calculations with the time-dependent Boltzmann's equation have shown that deviations of this kind may occur when the electron temperature increases rapidly as a result of a rapid increase in the electric field-strength. This may have the effect of altering the velocity distribution in such a way that there are relatively more electrons of high energy than there would be in a d.c. discharge with the same average electron energy. This will clearly cause an increase in both excitation and ionization. The calculations show that this effect increases with the modulation of the electron temperature. Experiments on a pulsed gas discharge have revealed that this pulsation may in fact have a considerable influence on the behaviour of the discharge

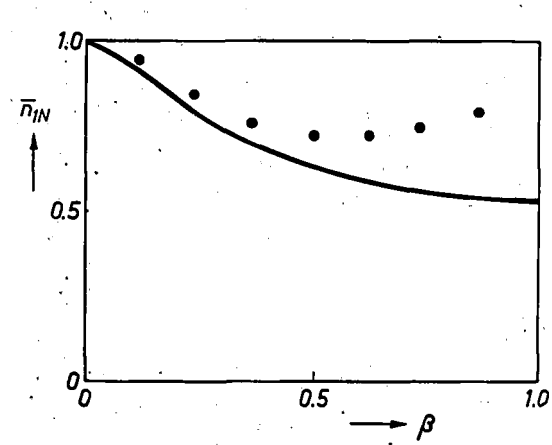


Fig. 5. Normalized value $\bar{n}_{1N}$ of the concentration of the excited atoms in the $^3P_1$ state as a function of the modulation coefficient β in a mercury-neon discharge carrying a current described by $I = I_0 (1 + \beta \sin \omega t)$; the modulation frequency is 1 kHz. The points give experimental values, the solid curve was calculated from the model. Up to $\beta = 0.5$ good agreement is obtained between the theoretical and experimental results. The value of I_0 was 0.4 A, the diameter of the discharge was 5.2 cm.

[16] R. Bleekrode and J. W. van de Laarse, J. appl. Phys. 40, 2401, 1969.
J. H. Waszink and J. Polman, J. appl. Phys. 40, 2403, 1969.
[17] J. Polman and J. E. van der Werf, Phys. Lett. 42A, 153, 1972.
[18] J. Polman, Physica 54, 305, 1971.

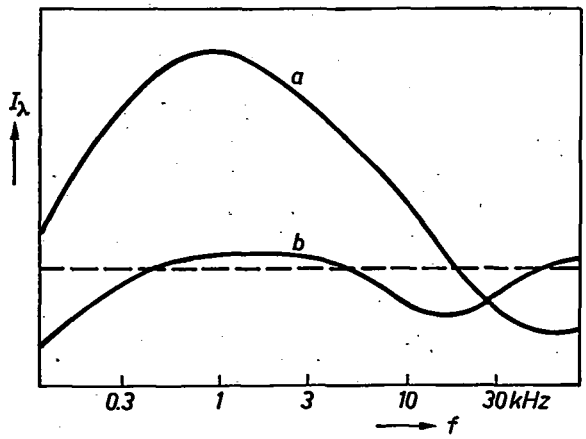


Fig. 6. Radiation output I_λ of a pulsed mercury discharge as a function of pulse frequency f. The outputs have been normalized to the output of a d.c. discharge at the same power (this level is indicated by a dashed line). Curve *a* gives the output at a wavelength of 436 mm, curve *b* the output at a wavelength of 254 nm.

(*fig. 6*). A relative increase in the number of fast electrons has the greatest influence on the populations at the higher levels, which explains why the effect on the 254-nm emission is less than the effect on the 436-nm emission. (The 436-nm emission originates from a transition terminating at the $^3P_1$ level.)

Similar arguments can also be applied to discharges operated from a normal a.c. supply, and these have been confirmed experimentally. The efficiency of such a discharge is highest at supply frequencies of about 100 kHz, and lowest at 500 Hz. The standard 50-Hz supply is not ideally suitable here [9].

The efficiency measurements and calculations referred to above relate only to the positive column; what is more, they were carried out on discharges with dimensions not normally used in practice. It is also found that the electrode losses are smaller for a high-frequency discharge current [19]. At the present state of the technology the overall improvement of efficiency, however, is not very substantial. This means that, after allowance has been made for the losses arising in the conversion of the 50-Hz mains voltage into a 100-kHz supply voltage, only a slight improvement of efficiency remains. However, fluorescent lamps in trains and aircraft are operated from higher-freqency a.c. supplies.

From what we have said above it can be seen that experiments and analysis of the properties of the theoretical model provide a good understanding of the behaviour of a low-pressure discharge in a mixture of mercury vapour and inert gas. The efficiency with which electrical energy is converted into the required ultraviolet radiation in the positive column may closely approach 80%. This can even be improved slightly by modification in the mercury pressure, the inert-gas pressure or the inert-gas composition [20].

Behaviour of a low-pressure sodium-inert-gas discharge

As we have already seen (fig. 2), depletion occurs in a sodium discharge at higher currents. At higher current densities the ionization reaches a level at which the uncharged atoms can only just reach the central part of the discharge, and excitation is then confined to a zone along the wall.

From the equations (1), (2) and (3) forming part of the model discussed above for the positive column, we can derive the relation

$$n_0(r) = n_w - \frac{D_a}{D_0} n_e(r) - \frac{D_1}{D_0} n_1(r), \qquad (4)$$

which describes the depletion in the steady-state case. Under the normal conditions in a low-pressure sodium-inert-gas discharge the last term amounts to no more than 10% of the total. Since D_a is about 7 to 10 times greater than D_0, and $n_e(r)$ increases towards the axis of the discharge, n_0 will decrease towards the discharge axis, the rate of decrease increasing with the current and hence the value of n_e.

Fig. 7 shows a number of current-voltage characteristics for a sodium-inert-gas discharge [11]. When these characteristics were taken a heat pipe was used to keep the wall temperature of the discharge tube and hence the sodium density constant [21]. With this arrangement the wall temperature could be adjusted as required and kept independent of the power taken by the discharge. It can be seen that, as the current increases, there is initially a decrease in the voltage as a result of stepwise ionization. At the same time the sodium-ion concentration increases, so that near to the axis the concentration of atoms in the ground state decreases, as indicated by equation (4). This concentration will gradually become so low that any further increase in the ionization of Na atoms can only be produced by an increase in voltage. Once all the sodium atoms have been ionized, a further increase in current is only possible through the ionization of inert-gas atoms. Their ionization energy is much higher than that of the Na atoms, which means that the voltage across the discharge must increase sharply to produce a further rise of current, as can clearly be seen from the characteristics. The figure also shows that the rise in the lamp voltage as a result of depletion occurs at a current value that increases with the wall temperature and hence the concentration n_w of atoms at the wall. The consequences of depletion may also be recognized

[19] M. Koedam and W. Verweij, Proc. 7th Int. Conf. on Phenomena in ionized gases, Belgrade 1965, Vol. III, p. 392.
[20] T. G. Verbeek and P. C. Drop, J. Physics D 7, 1677, 1974. M. Koedam and A. A. Kruithof, Physica 28, 80, 1962.
[21] C. R. Vidal and J. Cooper, J. appl. Phys. 40, 3370, 1969. A general survey on heat pipes can be found in G. A. A. Asselman and D. B. Green, Philips tech. Rev. 33, 104 and 138, 1973.

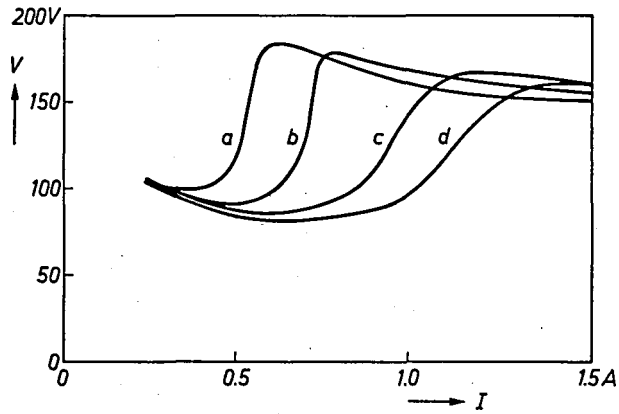


Fig. 7. Current-voltage characteristics of a d.c. sodium discharge for constant wall temperature.
Curve *a* wall temperature 248.2 °C ($n_w = 2.7\times10^{19}$ m^{-3}),
curve *b* wall temperature 255.6 °C ($n_w = 3.3\times10^{19}$ m^{-3}),
curve *c* wall temperature 257.4 °C ($n_w = 4.1\times10^{19}$ m^{-3}),
curve *d* wall temperature 260.0 °C ($n_w = 4.6\times10^{19}$ m^{-3}).
The diameter of the discharge tube was 2 cm. The particle density of the 99% neon-1% argon mixture added as buffer gas is about 10^{23} m^{-3}.

from measurements and calculations of other parameters of the discharge as a function of the current (*fig. 8*). All the calculations were carried out at currents below the critical value I_{crit}, where very considerable depletion causes a marked increase in the electric field-strength. It should be recalled here that a basic assumption for the model was that the inert gas in a discharge is not ionized, a condition that is not satisfied at currents above I_{crit}. The effects shown in fig. 8 are in complete agreement with equation (4) and with the description of the depletion given above. Thus, when I_{crit} is approached there is an increase both in field-strength and in electron temperature. According to eq. (4) the electron concentration at the axis in the event of complete depletion, i.e. where $n_0(0) = 0$, must be approximately equal to $(D_0/D_a)n_w$, since in that case $n_1(0) \approx 0$. If sodium is the only source of ions, which is the case up to I_{crit}, and since n_w is constant as long as the wall temperature is constant, this implies that $n_e(0)$ is determined by the ambipolar diffusion coefficient D_a, which characterizes the total diffu-

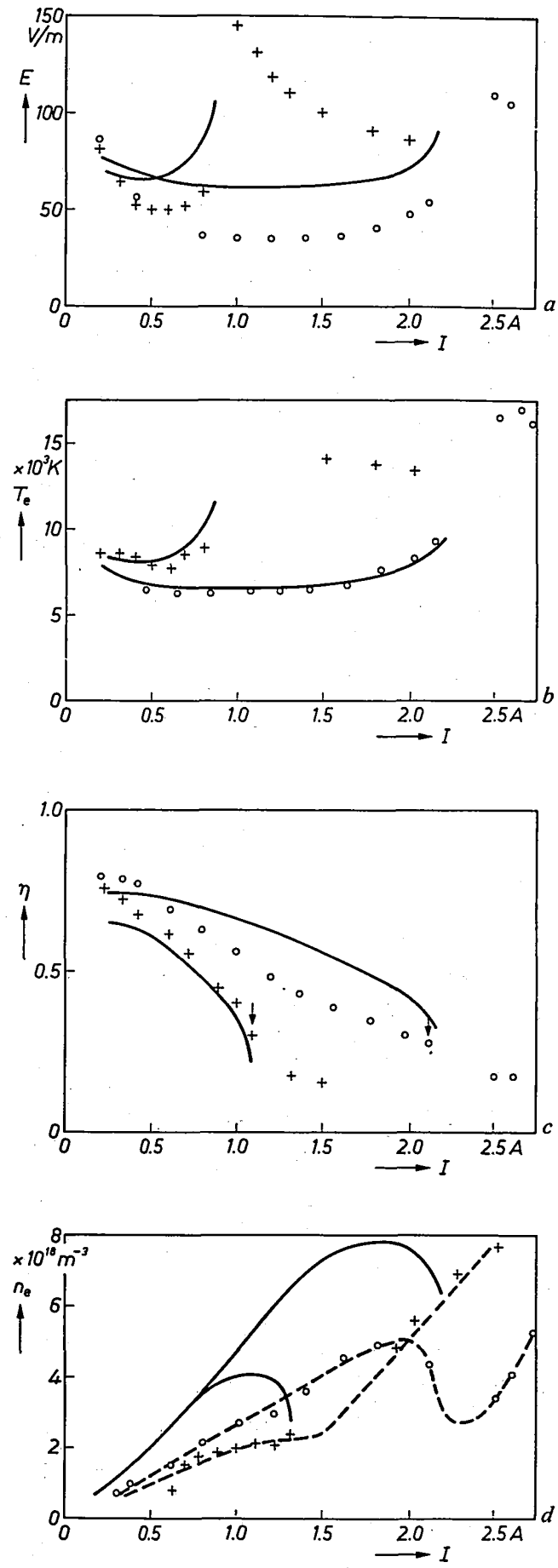


Fig. 8. Effect of depletion of a sodium discharge on the behaviour of various parameters of the discharge as a function of the current I, at constant wall temperature. Measurements at a wall temperature of 236.1 °C are indicated by +, and those at 254.3 °C by O. The solid lines give the model calculations at these two temperatures (only for currents below the critical value). The diameter of the discharge tube was 2.8 cm. *a*) The field-strength E in the positive column. *b*) Electron temperature T_e. *c*) Luminous efficacy η; the results of measurements and calculations were set equal at a single point, since absolute measurements of efficacy are very difficult with the arrangement of the discharge tube in a heat pipe. The arrows indicate the critical current values. *d*) Electron concentration n_e. In all cases there is at least a reasonable qualitative agreement between measured and calculated behaviour.

sion of electrons and ions. This coefficient is proportional to the *ion* mobility and the *electron* temperature: $D_a = \mu_i k T_e/e$. An increase in the electron temperature with depletion will therefore initially indicate a decrease in the electron concentration at the axis. Here again, there is at least a quantitative agreement between theory and experiment. To improve the quantitative agreement between measurements and calculations it is necessary to refine the model used in a number of ways. Thus, the way in which the radiative transport is taken into account and also the description of the atomic processes should be examined in greater detail. The higher energy levels of the sodium atom should also be included in the calculations.

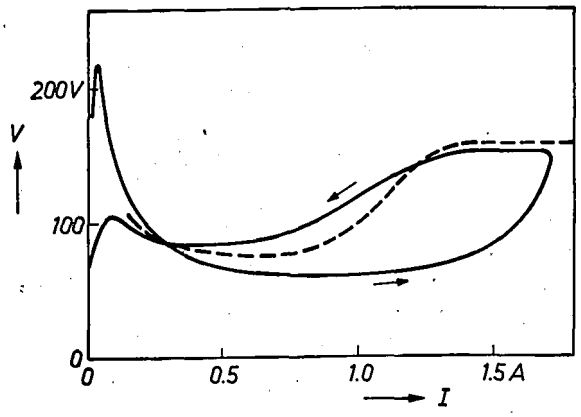


Fig. 9. Hysteresis in the current-voltage behaviour of a sodium discharge with a 50-Hz supply. V is the instantaneous value of the voltage, I that of the current. The arrows indicate the direction in which the point moves along the curve. The dashed curve is the d.c. voltage characteristic.

The sodium lamp

In practice low-pressure sodium lamps are almost invariably operated from an a.c. supply. The above description of the d.c. behaviour, although contributing to a better understanding of the discharge, is not directly applicable to practical situations. The main reason for this is the occurrence of hysteresis in the discharge. The behaviour during a decrease in current is not the same as that during an increase in current, as can be seen from *fig. 9*. We have therefore made a closer study of the differences between d.c. and a.c. discharges in a low-pressure sodium lamp. Calculations indicate that the maximum depletion in a sodium lamp occurs about 2 ms after the maximum of the 50-Hz current. This effect is one of the causes of the hysteresis.

The characteristics of an a.c.-fed lamp, plotted while the wall temperature was kept constant by external means (*fig. 10*), show substantial agreement with the d.c. characteristics [22]. The lamp characteristic plotted for a lamp with the normal heat insulation [23], whose wall temperature was thus not kept constant, shows that under standard conditions the effective wall temperature is about 258 °C and that in this case depletion effects do indeed come into play. This also appears from *fig. 11*, where the behaviour of the field-strength in the discharge is plotted as a function of time for half a period of the 50-Hz a.c. current. The figure also gives the value of the field-strength in a lamp operated with a d.c. current equal to the r.m.s. value of the a.c. current. It can be seen that the current here remains below the critical value. From fig. 8*c* it can be seen that the efficiency of the discharge as a light source decreases with increasing depletion. At currents above the critical value some of the power dissipated in the discharge is taken up by the inert gas, which is very inefficient as a light producer. Since a large part of the power in an a.c. discharge is dissipated near the current

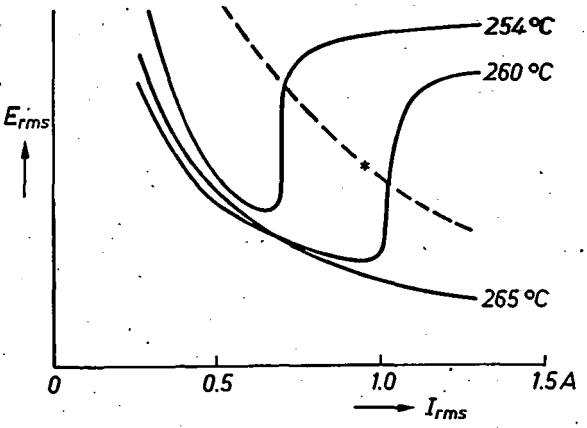


Fig. 10. R.M.S. field-strength E_{rms} in an a.c. sodium discharge for three values of wall temperature (solid curve) as a function of the r.m.s. value I_{rms} of the discharge current. The diameter of the discharge tube was 1.9 cm and the particle density of the buffer gas was about 10^{23} m^{-3}. The dashed curve gives the operating points for a lamp whose wall temperature adjusts to an equilibrium value corresponding to the dissipated power. The operating point of a normal SOX lamp with the same parameters is indicated by *; it can be seen that the wall temperature is about 258 °C and that depletion in this lamp is not negligible.

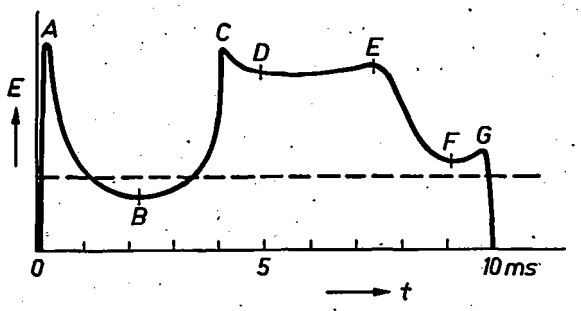


Fig. 11. The field-strength E in the same SOX lamp as in fig. 10, as a function of the time t for an r.m.s. discharge current of 1250 mA at a wall temperature of 260 °C, during a half-cycle of the 50-Hz supply. A is the re-ignition peak; while the curve goes from A to B the field decreases because of an increase in the stepwise ionization in the plasma. The rise BC is due to the occurrence of depletion; the field-strength then decreases to D (current maximum) owing to increasing ionization of the inert gas. From then on, as the current decreases, the complete process takes place in the region DG, but in the reverse sequence, and with a phase shift.

maximum, which is just where the maximum depletion occurs, it is to be expected that the efficacy of a sodium lamp will be improved by d.c. operation. *Fig. 12* shows the conditions under which d.c. is more efficient than a.c. as an energy source for the discharge, and the conditions under which the opposite applies.

Although a d.c. discharge would be preferable in certain cases, there are a number of problems with d.c. operation, such as the stabilization of the discharge.

Instead of a d.c. supply, a circuit can be used that clips the peaks of a normal a.c. current, giving an approximation to a 'square wave'; the r.m.s. and average values of the lamp current are then much closer.

The functions of such a circuit, and indeed of any ballast for a discharge lamp, are to ignite the discharge, to stabilize the discharge and to ensure re-ignition at every half-cycle of the supply. For an SOX 180-watt lamp the ignition and re-ignition voltages are so much higher than the operating voltage, which is 245 V, that the voltage of the unloaded ballast — the open-circuit voltage — has to be 660 V to ensure ignition and re-ignition under all circumstances. This implies that the impedance of the ballast must be high compared with that of the lamp, and therefore the losses will be relatively high. The ballast will then deliver a current that is virtually independent of the load. It has been found that the operating voltage of the lamp increases during life, so that the power taken by the whole circuit will also increase, which is not usually very desirable.

In a ballast designed by J. Rottier [*] an electronic circuit supplies the ignition and re-ignition peaks. The open-circuit voltage of the ballast then does not have to be much higher than the maximum operating voltage of the lamp. *Fig. 13* shows the basic circuit diagram for a ballast of this type. The inductance L_2 is so designed that its iron core is saturated for about one third of the time near the maximum value of the lamp current. This gives a voltage waveform with a strong third harmonic (*fig. 14*). The use of such a waveform has the effect of reducing the losses mentioned above due to power dissipation in the inert-gas discharge, and increases the luminous efficacy of the lamp by about 5%. A further advantage is that the ballast impedance can now be low since the ignition voltage is supplied by a separate circuit. This means that when the operating voltage rises during the life of the lamp, the delivered current will decrease, so that the power taken will remain approximately constant. The lower impedance also reduces the losses in the ballast, increasing the overall efficiency of the circuit by more than 10%.

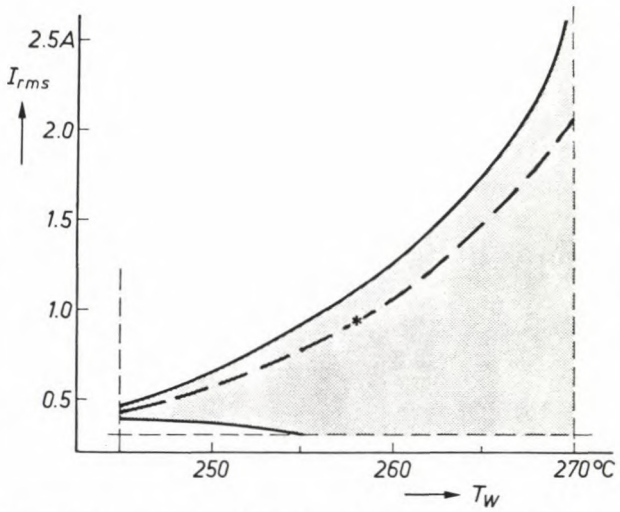


Fig. 12. The region in which, depending on the r.m.s. discharge current I_{rms} and the wall temperature T_w, a d.c. sodium discharge (grey area) or an a.c. sodium discharge (white area) is more efficient as a light source. The dashed line gives the conditions for which the difference between the two forms of discharge is greatest. The asterisk denotes the operating point of the 90-W SOX lamp.

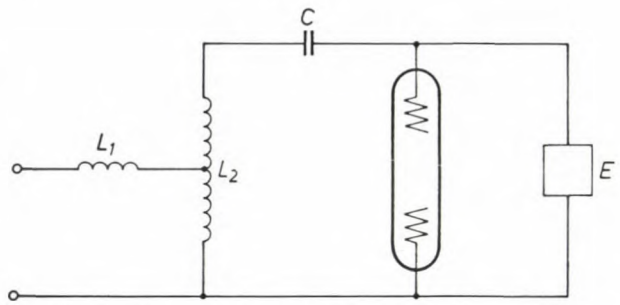


Fig. 13. New ballast proposed for the supply of a low-pressure sodium lamp. L_1 linear inductance. L_2 saturable reactor connected as an autotransformer. C capacitor. E electronic circuit that supplies ignition and re-ignition voltage. The inductances of L_1 and L_2 are given a ratio of 1 : 4, which ensures that the current through the circuit and hence the magnetic flux is mainly determined by L_2. Near the current maximum, L_2 becomes saturated, causing the lamp voltage to drop. This limits the power dissipated during depletion in the inert-gas discharge and thus improves the luminous efficacy.

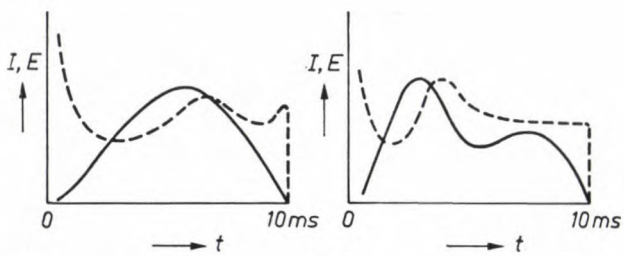


Fig. 14. Current I and field-strength E for a low-pressure sodium lamp (current : solid curve, field-strength: dashed curve). *a)* For a lamp connected to the conventional autoleakage transformer. *b)* For a lamp connected to the new ballast shown in fig. 13.

[22] M. Koedam, R. L. C. de Vaan and T. G. Verbeek, Lighting Design & Application **5**, No. 9, p. 39, 1975.

[23] The conventional low-pressure sodium lamp (SOX lamp) consists of a U-shaped discharge tube surrounded by an evacuated outer envelope, coated on the inside with an indium-oxide layer. This layer improves the heat insulation because of its high infrared reflectivity. See for example H. J. J. van Boort and R. Groth, Philips tech. Rev. **29**, 17, 1968.

[*] Electrical Laboratory, Philips Lighting Division, Eindhoven.

Prospects for low-pressure discharge lamps

Any substitute for sodium or mercury in discharge lamps would have to meet the following specifications. The metal must have a particle density of about 10^{20} m^{-3} (a vapour pressure of a few mtorr) at a temperature lower than 450 °C. If the temperature for this required density in a discharge lamp were to be higher, the luminous efficacy would be reduced by heat losses. In addition, the atoms of the substituting element must give a high-efficiency conversion of the energy dissipated in the discharge into visible light or radiation in the near ultraviolet. These specifications indicate that cadmium, zinc and magnesium might be possible substitutes. Cadmium, whose energy-level diagram is comparable with that of mercury, requires by far the lowest working temperature (about 250 °C), and in this respect is thus comparable with sodium. The radiant emission of cadmium lies in the ultraviolet (λ = 228.8 and 326.1 nm), which means that, as in the case of mercury, phosphors are necessary for the conversion into visible light.

A considerable amount of work has been done on the use of cadmium in discharge lamps. Model calculations have shown [17] that the UV radiation can be generated very efficiently: 80-100 lm/W. Measurements have confirmed these calculations [24]. The only serious problem is to find suitable phosphors that are not attacked by the cadmium. Since no such phosphors have yet been found, it looks for the time being as if the low-pressure sodium lamp and the low-pressure mercury lamp are the only lamps of their kind.

Summary. Two low-pressure gas discharges are used for lighting purposes: the mercury discharge, in fluorescent lamps, and the sodium discharge. For a proper understanding of the behaviour of these discharges it is desirable to have a theoretical model to describe them. The model given here consists of the equations for the particle densities, the energy-balance equation and the equation giving the relation between current and field-strength, with added boundary conditions.

Study of the time-dependent behaviour of a mercury discharge shows that modulation of the electron temperature occurs at supply-voltage frequencies in the region of 100 kHz, resulting in a higher luminous efficacy. At practical current densities the behaviour of the sodium discharge is determined by diffusion effects. At high current densities a sodium deficit occurs in the centre of the discharge, the buffer gas then participates in the discharge and the efficacy decreases. A specially designed ballast helps to prevent this decrease. This article concludes with a brief review of possible substitutes for mercury or sodium in low-pressure discharge lamps.

[24] J. E. van der Werf, Verh. Dtsch. Phys. Ges., Frühjahrstagung Stuttgart 1974, p. 502.
H. J. F. G. Smeets and T. G. Verbeek, Bull. Amer. Phys. Soc. (2) 20, 254, 1975.

One of the three bays of the 'Jan Dellaert' hangar at Schiphol (Amsterdam Airport), specially built for maintenance work on Boeing 747 B aircraft. Each bay is illuminated by 199 high-pressure sodium lamps (400 W).

Temperature profiling of tungsten filaments in incandescent lamps by a chemical transport reaction

The use of incandescent lamps in projectors, floodlights, automobile lamps, etc. requires a high and uniform luminance from the filament. Most of these lamps are low-voltage types with a relatively high filament temperature, and nowadays usually with a halogen filling. The filament in such lamps typically consists of a coil of fairly thick tungsten wire with relatively few turns. This has the drawback that the temperature of the turns is considerably lower at the ends of the coil (200 to 400 °C) than at the centre, because of the relatively high thermal conduction of the wire and the heat dissipation through the thick lead-in wires. Added to this, temperature differences along the coil cause an axial tungsten transport from turn to turn, since the vapour pressure of tungsten is highly temperature-dependent. In low-voltage gas-filled lamps, in which marked axial temperature gradients occur, this effect may exceed the radial evaporation and can even determine the life of the lamp; the filament burns out at the place where the mass loss due to axial tungsten transport is greatest (*fig. 1*) [1].

Clearly, the quality of such lamps could be substantially improved by reducing the temperature differences within the filament coil. A completely flat temperature distribution is not, however, the best solution of this problem. The net mass transport from turn to turn would certainly be zero then, and the luminance would be the same all along the coil, but at the free ends of the coil, where there is a greater radiating surface, the radial tungsten losses would be more than twice as large as at the centre of the coil. Such a measure would therefore lead to faster burn-out at the end turns of the coil or at the lead-in wires. The best temperature distribution for long life is one in which the net axial mass transport from turn to turn is constant and exactly compensates the greater tungsten losses at the free ends of the coil. In fig. 1*a* the solid curve shows the characteristic steep temperature gradient along the coil of a 24-V projection lamp; the hatched area indicates where the optimum temperature-distribution curve should be.

This optimum temperature profile can be obtained in a simple manner by running the tungsten filament very briefly in a fluorine-containing atmosphere before normal operation [2]. The effect of this pretreatment may be explained with reference to *fig. 2* [3]. The graph shows the tungsten 'solubility' $\Sigma p_W/\Sigma p_{F_2}$ as a function of temperature in a gas consisting of argon to which 10^3 Pa (10^{-2} atm) of fluorine has been added. Between 2600 K and 3100 K the solubility decreases steep with rising temperature. For a hot tungsten filame this means that tungsten will be transported fro regions of lower temperature to regions of high

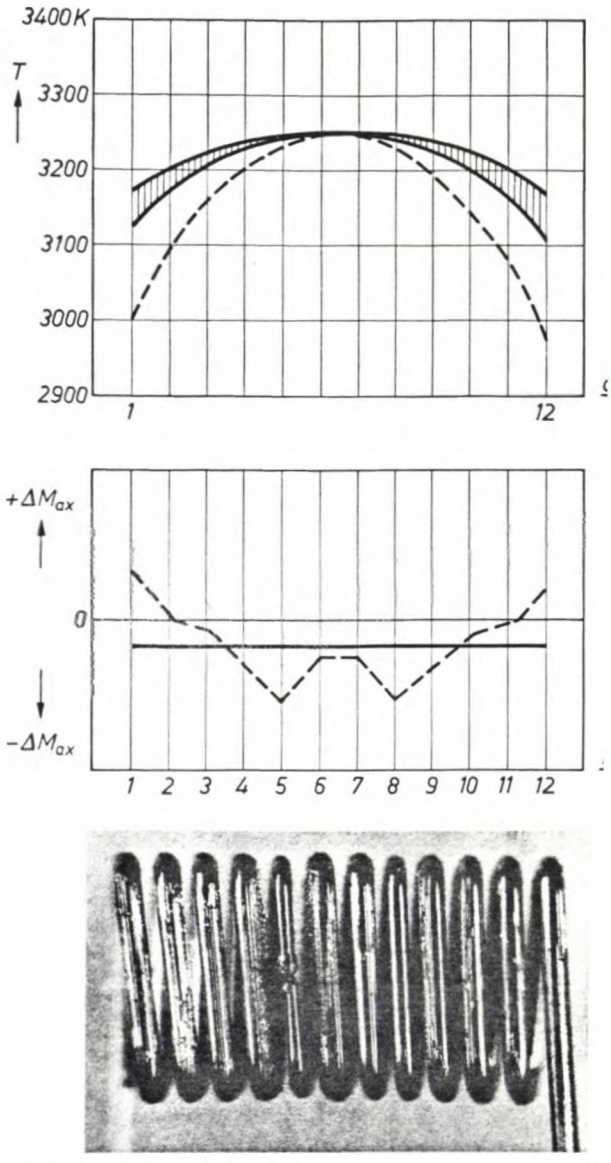


Fig. 1. Photograph of the filament in a projection lamp (12 V 100 W) after about 80 hours of operation. Because of axial ma transport the greatest loss of material in the filament of this kir of lamp is not found at the centre, but in two turns symmetrical located on either side of it.
a) Temperature-distribution curve of a non-profiled filame (dashed curve), and the region containing the optimum temper ture profile obtained after a preliminary treatment. This pretrea ment is carried out in a gas atmosphere containing fluorine. Tl coils have the number of turns shown in the photograph.
b) Typical curve for the effective tungsten transport ΔM_{ax} in tl axial direction of a filament in which all the turns are equal thick (dashed curve), and in an ideally profiled coil (solid curv

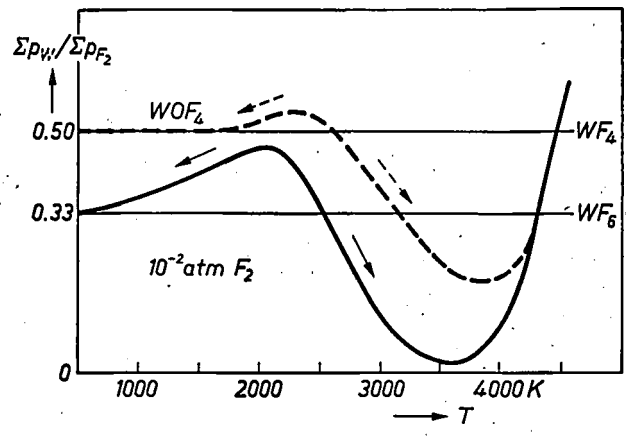


Fig. 2. Relative tungsten solubility $\Sigma p_W/\Sigma p_{F_2}$ in an atmosphere consisting of an inert gas with 10^3 Pa (10^{-2} atm) of fluorine (solid curve) and in an atmosphere to which oxygen has also been added, producing WOF_4 (dotted curve). The transport direction changes at 2500 K and 2600 K respectively. At temperatures below that limit the transport is towards the cooler parts of the system, whereas above it the transport is towards the hotter parts. The maximum in the curve of the system with WOF_4 is somewhat flatter, which means that there is only a slight tendency for tungsten to be transported towards the cool bulb wall.

temperature; the transport stops as soon as all parts of the coiled filament have the same temperature. In this pretreatment with fluorine the lamp is switched off as soon as the temperature profile appears to be of the optimum type.

Although in principle such transport reactions are also possible in an atmosphere containing chlorine, the processing is faster in a fluorine atmosphere; the profile is also somewhat smoother and is a better approximation to the optimum curve.

The filament can be profiled either in a separate vessel or in the lamp bulb itself. The cold pressure of the inert gas is $1\text{-}5 \times 10^4$ Pa (100-500 torr). We do not add free fluorine to the inert gas but NF_3, which is non-toxic and is easily handled; the filling pressure of the NF_3 is of the order of 10^3 Pa. During the profiling process the filament is at a temperature of between 2650 and 2800 K. The whole operation takes no more than about forty seconds.

The experimental results so far obtained with automobile lamps (12 V, 55 W, filled with CH_2Br_2) show that the pretreatment increases life about 120% for the same luminous flux or, for the same life, gives a 10% increase in luminous efficacy.

When filaments in small lamps are profiled, black spots may appear on the bulb in the immediate vicinity of the lead-in wires, which are cool. This happens because the transport goes in the opposite direction at temperatures below about 2650 K; i.e. it goes from the hot tungsten to the cool bulb wall. The black deposit can be returned to the filament, however, by 'flashing' the lamp in the same atmosphere for a few seconds at 3100 K. Chemical reaction between the fluorine and the quartz glass of the bulb can best be reduced by bringing the bulb into contact with cold water during the pretreatment. Preliminary experiments have demonstrated that the deposition of reaction products can be completely avoided by adding a little oxygen to the reactive gas filling. This results in the formation of WOF_4, which gives a flatter maximum in the solubility curve and a reduced mass transport to areas of lower temperature, as can be seen in the dashed curve in fig. 2. The details of this process are at present under investigation.

J. Schröder

[1] H. Hörster, E. Kauer and W. Lechner, Philips tech. Rev. 32, 155, 1971.
[2] For a general treatment of fluorine applications, see J. Schröder, Philips tech. Rev. 26, 111, 1965.
[3] See also G. Dittmer, A. Klopfer, D. S. Ross and J. Schröder, J. Chem. Soc. chem. Comm. 1973, 846.

Dr J. Schröder is with Philips GmbH Forschungslaboratorium Aachen, Aachen, West Germany.

The high-pressure sodium lamp

J. J. de Groot, J. A. J. M. van Vliet and J. H. Waszink

In the visible spectrum the high-pressure sodium lamp, like the low-pressure sodium lamp, mainly emits the resonance radiation of the sodium atom. In the high-pressure lamp (Na pressure 7-40 $\times 10^3$ Pa or 50-300 torr) this radiation is spread over a much wider range of wavelengths than in the low-pressure lamp (Na pressure about 0.4 Pa or 3×10^{-3} torr) [1]. This broadening of the spectrum (*fig. 1*) is caused by collisions between the particles, which appreciably broaden the excited level. Furthermore, the radiation absorption, which has a maximum at the central wavelengths of the yellow sodium lines, is very large because of the high density of the sodium vapour. This means that the radiation can more easily leave the plasma the farther its wavelength is removed from the central wavelength.

Because of the broadening of its spectrum, the light from the high-pressure sodium lamp permits discrimination between colours, which makes the lamp suitable for town lighting. The monochromatic light emitted by the low-pressure sodium lamp, on the other hand, allows very little colour discrimination and this lamp is therefore mainly used for lighting major roads. In a high-pressure discharge considerably more power can be dissipated per unit volume than in a low-pressure discharge. For this reason the high-pressure lamp is smaller than a low-pressure lamp and its luminance very much higher. Its luminous efficacy, however, is lower than that of a low-pressure lamp. The main reason for this is that an appreciable part of the energy input in a high-pressure lamp is emitted as spectral lines in the infrared. The visible radiation is also spread over a wider range of wavelengths. The spectrum of the high-pressure lamp is not so well adapted to the sensitivity curve of the human eye (the spectral-luminous-efficiency curve) as that of the low-pressure lamp, so that the same radiant power corresponds to a lower luminous flux.

The temperature of the discharge tube in a high-pressure sodium lamp (*fig. 2*) is fairly high. Sodium vapour is so chemically active at these temperatures that glass or quartz glass cannot be used as the material for the discharge tube. In the present state of the technology the only material suitable for this purpose is alumina, either in a densely sintered polycrystalline form or in the form of single crystals (sapphire). Wall temperatures of up to about 1500 K are permissible with this material.

As well as sodium the lamp also contains xenon, for ignition, and mercury, to reduce the heat-conduction losses and give a better match between the lamp and its ballast. The most important characteristics of the high-pressure sodium lamp are summarized in *Table I*.

In this article we shall survey the main physical aspects of a high-pressure sodium lamp and consider some questions that arise in the design of such a lamp. We shall not be concerned here with the ballast, necessary for the stable operation of the lamp, or with the technological problems associated with the manufacture of the various types of high-pressure sodium lamp [2] [3].

The physical model for a high-pressure discharge

In research on high-pressure discharges it is desirable to have a simple model for the discharge, partly to improve understanding of the processes involved, and partly to restrict the number of experiments needed for

Table I. Characteristics of the SON 400W high-pressure sodium lamp.

Gas pressures (in the operating lamp)	
Na	8 000- 20 000 Pa (60-150 torr)
Hg	55 000-110 000 Pa (400-800 torr)
Xe	20 000- 27 000 Pa (150-200 torr)
Material of the discharge tube	Al_2O_3, polycrystalline
Diameter of the discharge tube	7.5 mm
Electrode spacing	82 mm
R.M.S. current	4.4 A
R.M.S. lamp voltage	105 V
Electrode losses	24 W
Radiant power	
ultraviolet	2 W
visible	118 W
infrared	80 W
total	200 W
Luminous efficacy	120 lm/W
Colour-rendering index	23
Colour-temperature	2100 K

Dr Ir J. J. de Groot and Ir J. A. J. M. van Vliet are with the Philips Lighting Division, Eindhoven; Dr J. H. Waszink is with Philips Research Laboratories, Eindhoven.

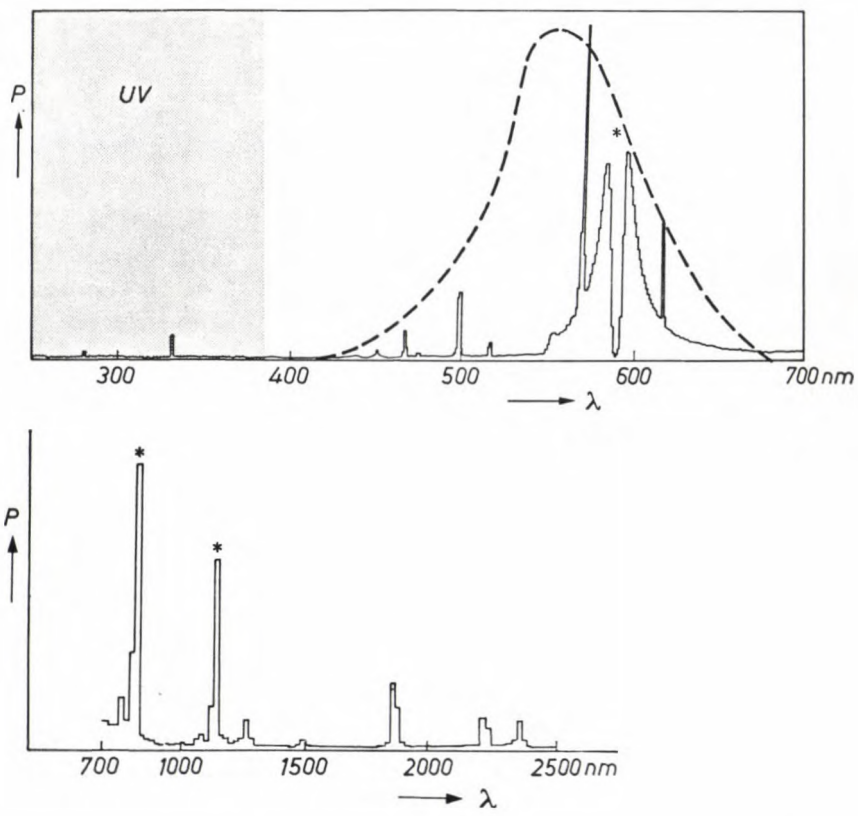


Fig. 1. The spectrum of a high-pressure sodium discharge. The relative power P for the ultraviolet and visible regions (*above*) has in all cases been integrated over wavelength intervals of 1 nm, and for the infrared (*below*) over intervals of 20 nm. The main spectral lines, corresponding to the energy-level transitions indicated in fig. 3, are marked with an asterisk. The line broadening and also the line reversal due to self-absorption can clearly be seen from the yellow sodium lines (wavelengths 589.0 and 589.6 nm). The dashed curve gives the spectral luminous efficiency (eye sensitivity) as a function of wavelength λ.

optimizing lamp design and developing new types. It is particularly important to restrict the number of experiments in cases where experimental parameters cannot easily be varied — a typical example here is the diameter of the discharge tube.

The dominant processes in a high-pressure sodium discharge are elastic and inelastic collisions between the particles in the plasma, and also the emission and absorption of radiation. Because of the high gas density and the low density gradients in the plasma, the diffusion currents are so small as to be negligible. This is quite different from low-pressure discharges, where diffusion is the dominant process.

In elastic collisions between electrons and atoms the electrons transfer to the atoms the energy they have acquired from the electric field, and this process has the effect of heating the gas. These collisions also determine the drift velocity of the electrons in the discharge and consequently the charge transport. The velocity distributions of the atoms and electrons may be described by a Maxwell distribution, so that the atoms may be said to have a temperature T_g and the electrons a temperature T_e. The difference between the two temperatures proves to be small in the case of a high-pressure discharge, as we shall see shortly.

As a result of inelastic collisions, atoms are excited or ionized, or one of these processes occurs in the reverse direction, giving rise to de-excitation or volume recombination.

The behaviour of the high-pressure sodium lamp is mainly determined by the radiative processes that take place in the positive column. About 80 % of the power emitted in the form of radiation by a high-pressure sodium discharge originates from transitions indicated in the simplified energy-level diagram in *fig. 3*. A model that only takes these three transitions into account will give a good description of the energy transport in the discharge. Inelastic processes involving mercury or xenon atoms have very little effect under the conditions considered here and we shall therefore disregard them.

In addition to the temperatures T_e and T_g we need various other parameters to describe the plasma in the positive column. These are the densities of all the particles occurring in the discharge, the radiation density and the electric field-strength. It is also necessary to know the effective cross-sections of all the collision processes. It can be assumed that the plasma is homogeneous along its axis, that the partial pressures of Na, Hg and Xe are constant over every cross-section of the discharge, and that the plasma has cylindrical symmetry.

We can now postulate a number of relations between the various parameters, which together form the

[1] K. Schmidt, C.R. 6e Conf. Int. sur les Phénomènes d'ionisation dans les gaz, Paris 1963, Vol. III, p. 323.
[2] See for example B. de Vrijer, Elektrotechniek **43**, 512, 1965.
[3] See for example L. B. Beijer, H. J. J. van Boort and M. Koedam, Lighting Design & Appl. **4**, July 1974, p. 15.

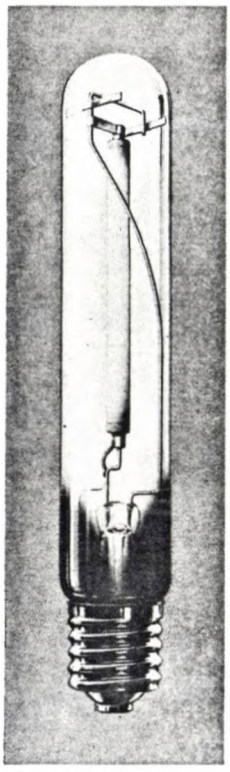


Fig. 2. A high-pressure sodium lamp. The central cylinder is the discharge tube, made from densely sintered alumina. The outer envelope, which is cylindrical here, is evacuated. Another type of envelope is shown in the introductory article [4].

physical model of a steady-state discharge. We begin with conservation equations for the particle densities of the sodium atoms in the various states. Next come the energy-balance equations, one for the electrons and one for the atoms. The first of these equations states that the power supplied to the electrons per unit volume is equal to the power transferred to the atoms by elastic and inelastic collisions. The power transferred by inelastic collisions is equal to the power radiated per unit volume. The second equation states that the power transferred to the atoms at elastic collisions corresponds to the losses due to heat conduction. In the case of local thermodynamic equilibrium it is assumed that electron temperature and gas temperature are identical. There is then only one energy-balance equation, which applies to the system of electrons and atoms collectively. This equation is the sum of the two already mentioned, and is given in the introductory article [4] as equation (1). The power radiated per unit volume is the difference between the emitted and the absorbed power. The absorbed power depends on the radiant flux density, which in turn depends on the particle densities elsewhere in the plasma. This second dependence is described by an equation that gives the radiative transport in the discharge, an equation that presents appreciable difficulties in calculations on discharge models, as can be seen from the discussion in the introductory article [4]. Because of these difficulties we have to use relatively simple models for calculations on discharges. The set of equations collectively forming the model for the high-pressure discharge is completed by the equation of state for the discharge gas. Simultaneous solution of these equations yields the values of the discharge parameters at every point in the column.

If we now compare the high-pressure sodium discharge with the low-pressure discharge described elsewhere in this issue, we see the following differences. In the high-pressure discharge a substantial part of the power taken up by the electrons from the electric field is transferred to the atoms by elastic collisions. In this way the gas becomes very hot, and as a result — unlike the situation in a low-pressure discharge — there is only a small difference between T_e and T_g. Furthermore, as we saw earlier, diffusion currents in the high-pressure discharge are low, whereas in the low-pressure case they determine the properties of the discharge. Among the inelastic processes, volume recombination at low pressure is negligible, whereas at high pressure it is in fact the principal recombination process. The other inelastic processes discussed — excitation, de-excitation, ionization, and the emission and absorption of radiation — play an important role in both cases. It follows from what has been said that there will be significant deviations from local thermodynamic equilibrium in a low-pressure discharge but only minor ones in a high-pressure discharge.

Complete thermodynamic equilibrium can never exist, however, in a plasma that emits radiation. The actual deviation from this equilibrium is not easy to estimate, but nevertheless it is important to know something of the deviations. If they are small, it is a reasonable approximation to assume that local thermodynamic equilibrium exists, which makes the calculations very much simpler. To get some idea of the

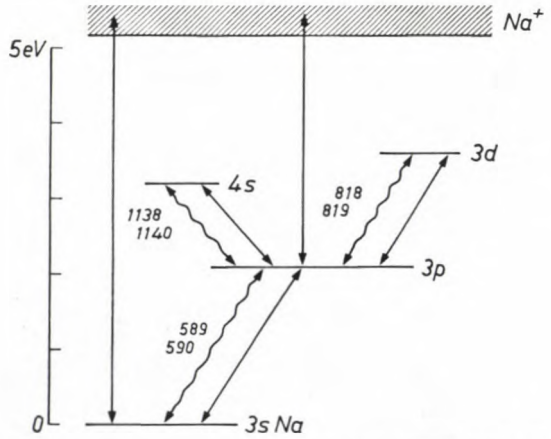


Fig. 3. Simplified energy-level diagram for sodium. Only those levels are shown that appear in the model used here for the theoretical description of a high-pressure discharge. The ordinary arrows indicate non-radiative transitions, the wavy arrows the radiative transitions. The wavelength of the radiation is given in nm.

difference between T_e and T_g we have calculated the temperature of electrons and atoms as a function of the radial coordinate, without assuming local thermodynamic equilibrium. *Fig. 4* shows a result of these calculations [5]. Near the axis the difference between the two temperatures is small, about 10 K, but near the outside of the discharge the difference increases to about 600 K. This means that there is effective local thermodynamic equilibrium in the hottest part of the plasma, and there is therefore no objection to assuming this equilibrium in calculations of the plasma temperature, field-strength and total radiant flux. However, the spectral intensity distribution of the resonance radiation originating from the outer layers of the discharge can only be accurately described by taking into account the deviations from local thermodynamic equilibrium in these layers. We shall deal with this subject when we discuss the various spectra.

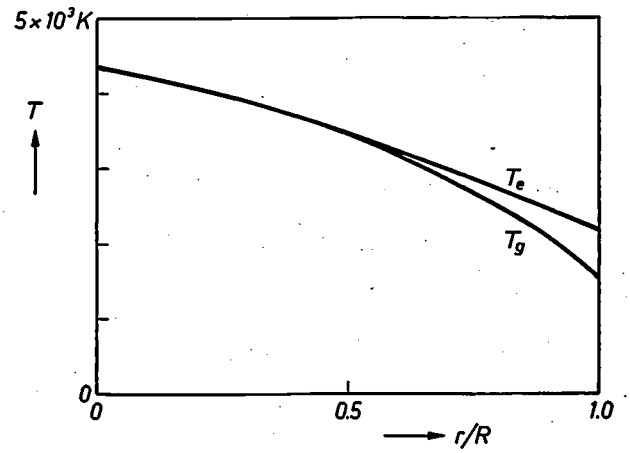


Fig. 4. Calculated curve for the gas temperature T_g and the electron temperature T_e in a high-pressure sodium discharge as a function of the normalized radius r/R. (R is the radius of the discharge tube, which was 3.75 mm.) The calculations were carried out for a discharge in which the sodium pressure was 36 000 Pa (270 torr) and the xenon pressure 23 000 Pa (170 torr). In these calculations a model with only three levels was used; the 4s and 3d levels were not taken into account. Local thermodynamic equilibrium was *not* assumed.

Spectroscopic methods of measurement

For the model calculations referred to above, and for their experimental verification, and also for designing a particular lamp, it is necessary to know various quantities relating to the discharge, such as the vapour density, the temperature distribution in the discharge and the wall temperature of the discharge tube. We measure these quantities by the spectroscopic methods commonly used in plasma physics [6].

As we have seen, line reversal takes place in a plasma of sufficient optical density (fig. 1). This is due to a lack of uniformity in the radial temperature distribution, which gives rise to inhomogeneity in the population densities of the various energy levels. The vapour pressure can be derived from the difference in wavelength $\Delta\lambda$ between the two maxima in the profile of a self-reversed resonance line (*fig. 5*) [7]. From the maximum radiance of a non-resonant spectral line observed in a particular direction it is possible to derive the maximum temperature in that direction. (Radiance is defined as the radiant flux per unit surface emitted by a radiating object per unit solid angle and per unit wavelength.)

Infrared pyrometry, which is the most obvious method of determining the wall temperature of the discharge tube, can only be used by including in the outer envelope a special window that transmits infrared radiation. It has been found, however, that the wall temperature can be determined simply and reliably by measuring the electron temperature after switching off the discharge [8]. The electron temperature, which is higher than the wall temperature T_w in an operating lamp (fig. 4), decreases to the value T_w within a few milliseconds after switching off the lamp. The time

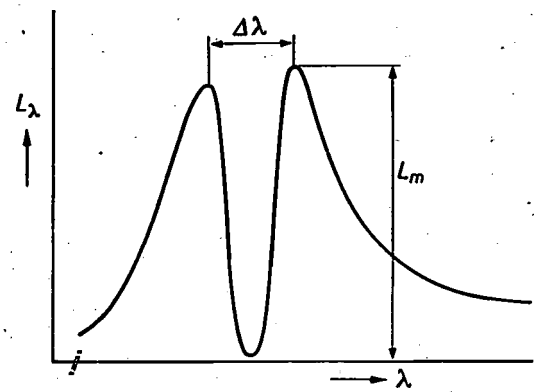


Fig. 5. Profile of a self-reversed spectral line. The difference in wavelength $\Delta\lambda$ between the two maxima in the case of a resonance line is a measure of the particle density of the atoms in the ground state. In the case of a non-resonant line the maximum radiance L_m is a measure of the maximum temperature in the direction of observation.

constant for the cooling of the wall is a few seconds, and there is therefore ample time to determine the electron temperature, which has fallen to T_w, from the radiance at the centre of an optically thick spectral line like the sodium D line. The wall temperature of the discharge can then be found by extrapolating back to the time at which the discharge was switched off.

[4] R. Bleekrode, M. Koedam and L. Rehder, this issue, p. 308.
[5] J. H. Waszink, J. Physics D 6, 1000, 1973, and J. appl. Phys. 46, 3139, 1975.
[6] See for example H. Zwicker, in: W. Lochte-Holtgreven (ed.), Plasma diagnostics, North-Holland, Amsterdam 1968, Chap. 4.
[7] Teh-Sen Jen, M. F. Hoyaux and L. S. Frost, J. quant. Spectr. rad. Transfer 9, 487, 1969.
[8] J. J. de Groot, Proc. 2nd Int. Conf. on Gas discharges, London 1972, p. 124.

Measurement and calculation of the macroscopic parameters

The physical and electrical properties of the high-pressure sodium discharge are determined by the choice of the sodium vapour pressure, the nature and pressure of the buffer gas, and the power dissipated in the discharge. In calculating these quantities it may be assumed, as we have seen, that there is local thermodynamic equilibrium in the discharge. As we noted earlier, this considerably simplifies the model.

The densities of the excited atoms and of the electrons follow from the equations of Boltzmann and Saha. The energy-balance equation for the time-varying discharge is now [9]:

$$\varrho c_p \frac{\partial T}{\partial t} = \sigma E^2 - \mathrm{div}(-\kappa \,\mathrm{grad}\, T) - U(T).$$

The corresponding equation for a discharge that is constant in time has been dealt with in the introductory article [4]. The relation between the discharge current I and the field-strength E is given by

$$I = 2\pi E \int_0^R \sigma r dr.$$

In the above equations ϱ is the mass density and c_p the heat capacity of the gas, T is the temperature, t the time, σ the electrical conductivity and κ the thermal conductivity. U is the radiant power per unit volume. In calculating U we have assumed that the sodium D lines are optically thick and that the infrared lines are optically thin [10]. From these two equations, together with the equation of state for the discharge gas, we can calculate the temperature T as a function of time and position. The boundary conditions are the wall temperature T_w, the condition that $\partial T/\partial r$ is at all times zero at the axis, and the condition that the temperature distribution is $T(r,t_0)$ at any given time t_0. Finally, the current I must be given as a function of time. In the case where the current is assumed to be constant as a function of time, this model with local thermodynamic equilibrium [9] [10] yields values for the gas temperature, the field-strength and the radiation output that compare well with the results of a model in which thermodynamic equilibrium is not assumed.

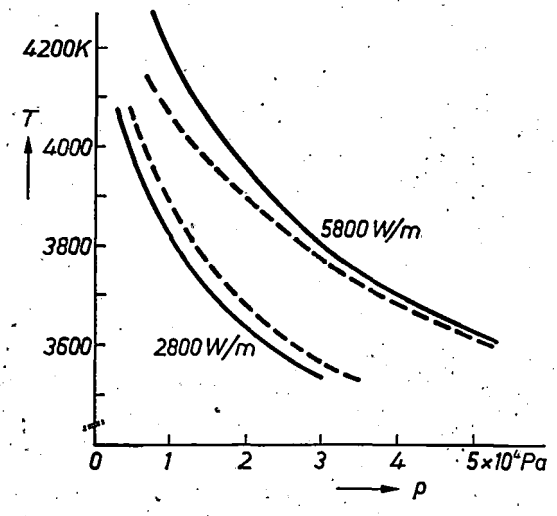


Fig. 6. Measured values (solid curves) and calculated values (dashed curves) for the temperature T at the axis of a sodium discharge as a function of the sodium pressure p for two values of the power dissipated in the discharge. The diameter of the discharge tube was 7.6 mm.

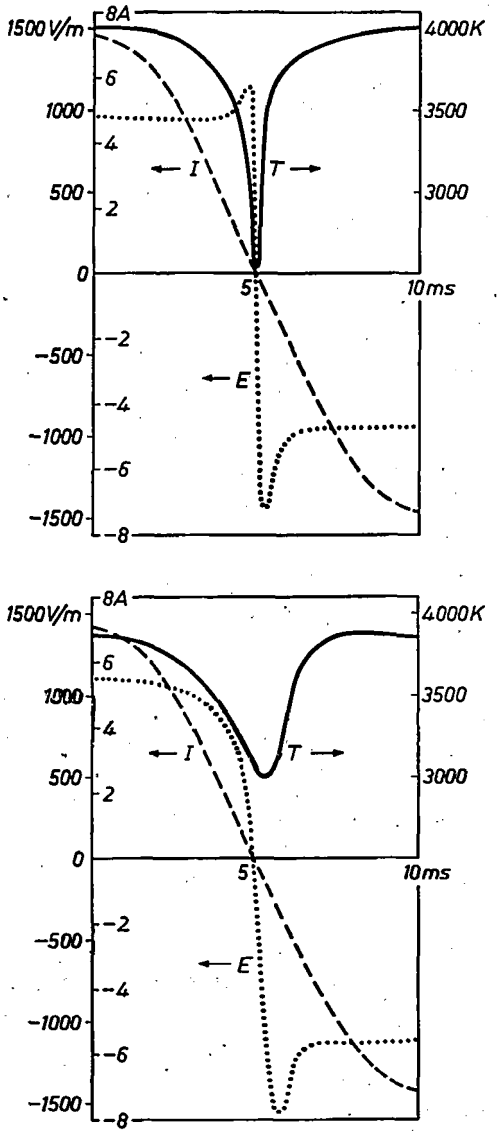


Fig. 7. Variation of the field-strength E and the temperature T at the axis as a function of time in a high-pressure sodium discharge operated from an a.c. supply (50 Hz). The dashed curves show the variation of the discharge current I. *a*) Behaviour of a pure sodium discharge at a pressure of 17 500 Pa (130 torr). *b*) Behaviour of a discharge to which xenon at a pressure of 230 000 Pa (1700 torr) was added as a buffer gas. The diameter of the discharge was 7.6 mm. The increased thermal inertia due to the addition of the buffer gas results in smaller temperature variations. In the field-strength curve the extinction peak has completely disappeared and the ignition peak is reduced.

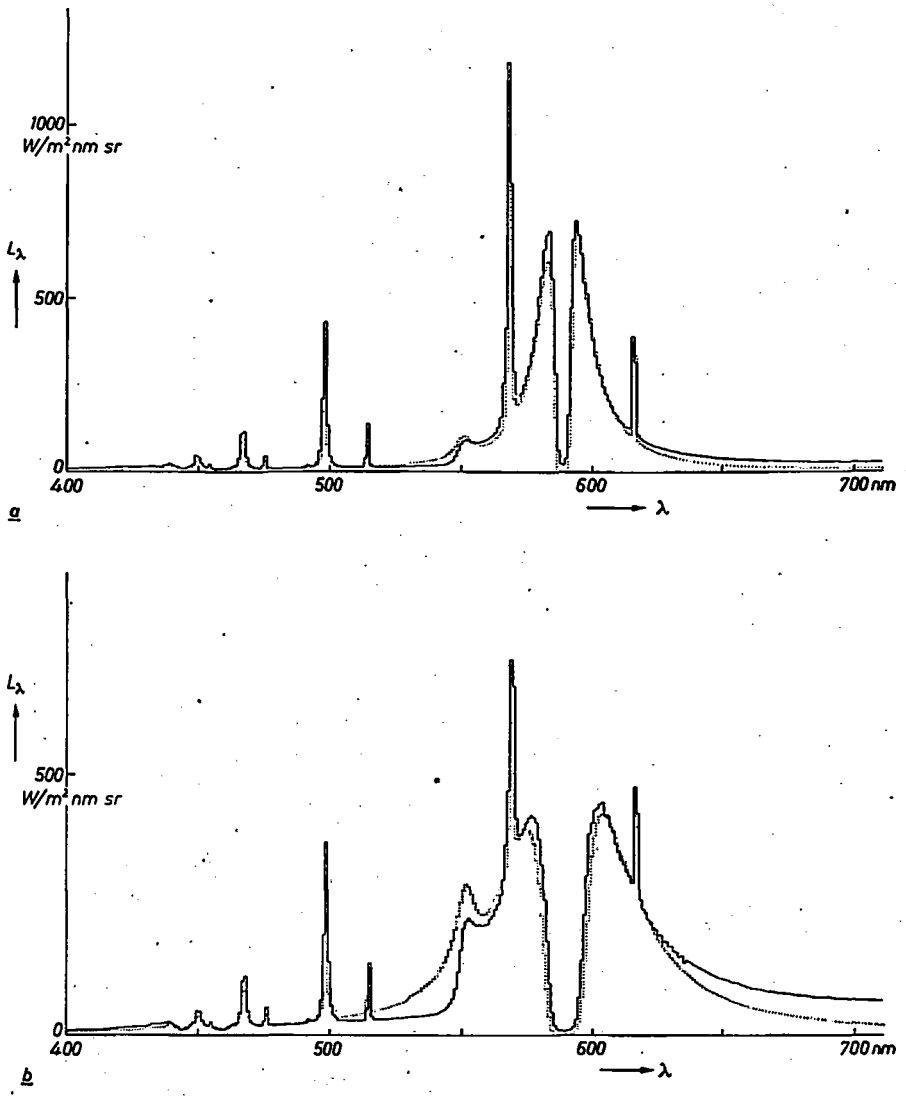


Fig. 8. Measured and calculated emission spectra of a high-pressure sodium lamp. The spectral radiance L_λ (in W/m² nm sr) has been integrated over wavelength intervals of 1 nm and represented as a function of the wavelength λ. Solid curves give the measured values, dashed curves the calculated values. *a*) Sodium pressure 17 000 Pa (125 torr). *b*) Sodium pressure 40 000 Pa (300 torr). For the band visible at 551.5 nm, particularly in case (*b*), the absorption coefficient is proportional to the square of the sodium concentration. The diameter of the discharge tube was 7.6 mm, the wall temperature in both cases was 1600 K and the dissipated power 4700 W/m. The calculations were corrected for reflection from the wall of the discharge tube.

Our calculations of the plasma temperature [10], both as a function of sodium pressure and of the dissipated power, are in good agreement with the results of the measurements, as can be seen from *fig. 6*. *Fig. 7* shows the effect, also calculated, of the buffer gas xenon on the field-strength and also the temperature at the axis in a sodium discharge operated from an a.c. supply (50 Hz). Without buffer gas there is a sharp decrease in the plasma temperature while the current passes through zero. As a result there is a temporary decrease in the degree of ionization, giving rise to sharp peaks in the field-strength: the extinction and re-ignition peaks. Raising the heat capacity and lowering the heat conduction of the plasma, which can be done by the addition of a heavy buffer gas like xenon or mercury, will have the effect of increasing the thermal inertia of the discharge. In this way the sharp drop in the plasma temperature during the zero crossing of the current can be avoided and hence the sharp extinction and re-ignition peaks.

The spectrum of the high-pressure sodium lamp

Calculation of the emission spectrum of a high-pressure sodium discharge is necessary to allow the effect of various parameters on the luminous efficacy of a discharge to be predicted without making measurements. The spectrum can be derived from the calculated temperature distribution, the sodium pressure and the equation for the radiative transport in the plasma [10]. In addition to the transitions shown in fig. 3, we have included in the calculations all other lines that make any significant contribution to the light output. Both for calculating and measuring the spectra shown in *fig. 8* the intensities in all cases were integrated over wavelength intervals of 1 nm. The measurements were made with a.c. discharges (50 Hz), with time-averaged radiance values. We adopted this procedure because we had found that time-averaged

[9] J. J. Lowke, J. quant. Spectr. rad. Transfer 9, 839, 1969.
[10] J. J. de Groot and J. A. J. M. van Vliet, J. Physics D 8. 651, 1975. See also J. J. de Groot, thesis, Eindhoven 1974.

values compared well with the results obtained from measurements on d.c. discharges for the same dissipated power [10].

The agreement between measurements and calculations is in general very good. Two remarks should be made, however. To begin with, the radiance measured for the long-wavelength 'wing' of the spectrum of the sodium D lines is greater than the calculated value, whereas the opposite is the case for the short-wavelength wing. When calculating the spectrum we assumed that the line had a Lorentzian profile determined by resonance broadening. This assumption is justified at relatively low pressures, because the linewidth is then small. In our case, however, we have a line considerably broadened by high pressure, and calculations now have to be carried out for wavelengths so far from the centre that a Lorentzian profile is obviously no longer a good approximation.

The second remark concerns the radiance at the centre of the sodium D lines. As we have seen, a model in which thermodynamic equilibrium is assumed will give a good description for the spectrum of radiation originating from the central part of the discharge. At the central wavelengths of the D lines the absorption coefficient of the plasma is so large, however, that the emitted radiation originates from a thin layer near the wall of the discharge tube. Our calculations lead us to expect a difference of about 600 K between T_e and T_g in this region (fig. 4). The model in which thermodynamic equilibrium is not assumed will therefore yield a higher radiance in the centre of a line than the model that does assume thermal equilibrium. *Fig. 9* shows that the non-equilibrium model yields a result in good agreement with the measurements, whereas the equilibrium model yields a radiance value that is too low. This confirms the calculated temperature curves in fig. 4.

Influence of the buffer gas

The addition of a buffer gas makes it possible to influence both the electrical and the spectral properties of a high-pressure sodium discharge. The effect of a xenon addition on the electrical behaviour has already been shown in fig. 7. It changes the mobility of the electrons only very slightly, and consequently there is little change in the field-strength. The addition of mercury, on the other hand, has a very considerable effect on electron mobility, and can therefore be used to increase the effective field-strength in the plasma. From the results of calculations presented in fig. 7 it also follows that increasing the thermal inertia of the plasma has a good effect on the extinction and re-ignition peaks. Owing to the poorer thermal conductivity of the plasma after the addition of xenon or mercury, the heat losses are reduced and the light output is increased. To examine the effect of the buffer gas on the luminous efficacy we must first determine the effect it has on the spectral energy distribution.

The main effect of the addition of mercury is to raise the intensity in the long-wavelength wing of the sodium D lines. The addition of mercury or of xenon both give rise to bands in the spectrum whose intensity, at constant temperature, is found to be proportional to the density of the sodium and also to that of the buffer gas. The band observed on the addition of xenon almost coincides with the maximum of the spectral-luminous-efficiency curve; the addition of mercury gives bands on the red wing of the sodium D lines.

By lowering the heat conduction losses, both buffer gases have the effect of increasing the luminous efficacy. Xenon produces a greater increase than mercury, because it improves the spectral intensity distribution. The high xenon pressures required, however, give too high an ignition voltage for the discharge.

The high-pressure sodium discharge as a light source

The designer of a high-pressure sodium lamp is free to choose such parameters as the sodium pressure, the type of buffer gas, the buffer-gas pressure and the length and diameter of the discharge tube in such a way as to produce an optimum light source for the

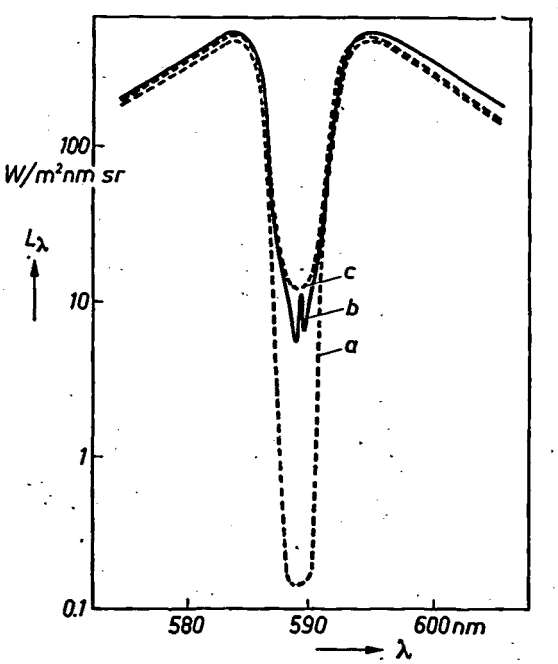


Fig. 9. The spectral radiance at the centre of the sodium D lines. *a* contour calculated on the assumption of local thermodynamic equilibrium. *b* measured contour. *c* contour calculated without assuming local thermodynamic equilibrium. The two separate emission minima revealed by the measurement may be a result of the effect of the tube wall on the plasma.

available supply voltage. The compromise he has to find between colour rendering and luminous efficacy is illustrated in *fig. 10*, where efficacy and colour-rendering index [11] are shown as a function of the sodium pressure.

A high-pressure sodium discharge has a negative, or very weakly positive, current-voltage characteristic. This means that the lamp must always be used in series with a ballast for stable operation. An optimum design is therefore not based on the efficacy of the discharge alone but must also include the ballast. A complication here is that the lamp voltage rises during the life of the lamp. One reason for this is that electrode material settles on the wall, causing blackening of the discharge tube around the electrode and hence locally increasing the temperature of the wall. Since the coolest parts of the wall are situated behind the electrodes and there is a surplus of both sodium and mercury, this also means that the sodium and mercury pressure will increase, thereby increasing the lamp voltage and the height of the re-ignition peaks, while at the same time reducing the luminous efficacy (see fig. 10). Another factor is the upper limit of the wall temperature. When densely sintered alumina is used the wall temperature should not exceed about 1500 K. Above that limit there is sublimation of the alumina, and this has the effect of reducing the transmission of the outer envelope.

We shall now show what the consequences are when a lamp with a given power has to be designed to meet the following specifications: maximum light output, an r.m.s. lamp voltage that to some extent is predetermined by the ballast and the mains voltage, and a wall temperature lower than 1500 K. In doing so we shall consider the influence of the parameters already discussed: sodium pressure, the type and pressure of the buffer gas and the dimensions of the discharge tube.

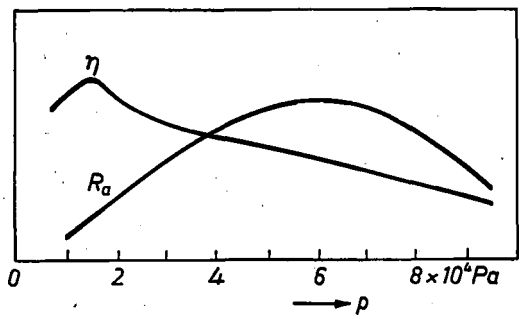


Fig. 10. Luminous efficacy η and general colour-rendering index R_a of a high-pressure sodium lamp (SON 400 W) as a function of the sodium pressure p [3].

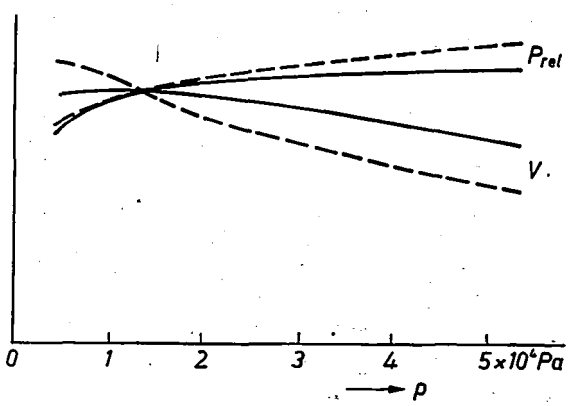


Fig. 11. Calculated (solid) and measured curves (dashed) of the relative radiant power P_{rel} and the luminous efficiency [11] of the radiation V as a function of the sodium pressure p in a high-pressure sodium discharge. The diameter of the discharge was 7.6 mm; the dissipated power was 3600 W/m. All curves have been normalized by dividing by the value at $p = 1.3 \times 10^4$ Pa (100 torr). The calculated curve for the luminous efficiency V differs from the measurements because of the discrepancy between the calculated and measured intensity in the long-wavelength wing of the sodium D line (see fig. 8).

The sodium pressure

As shown in fig. 10 a sodium pressure exists at which the luminous efficacy has a maximum. As the pressure decreases, the radiant efficiency decreases. This may be explained as follows. As the pressure decreases the temperature of the gas rises (fig. 6) and the temperature gradient towards the wall becomes steeper. As a result, losses due to heat conduction increase and the total radiant efficiency decreases (see *fig. 11*). When the sodium pressure increases, the total radiant efficiency increases slowly. The broadening of the D line, however, is so marked that the spectrum becomes badly matched to the spectral-luminous-efficiency curve. As a result the luminous efficacy decreases at higher pressures. The optimum sodium pressure is that at which the distance between the two intensity maxima of the self-reversed sodium D lines is about 10 nm, as found from measurements on various tube diameters and buffer-gas pressures.

Wall temperature

The wall temperature of the discharge tube in an evacuated outer envelope results from a steady state of the power supplied by conduction from the plasma and the power radiated from the wall in the form of heat. The heat-radiation term depends on the radiation-emission coefficient of the wall material of the discharge tube.

Measurements with the 'switch-off method' described earlier have shown that the wall temperature at optimum sodium pressures is mainly determined by the power dissipated in the discharge per cm² of wall area and depends very little on the tube diameter. For a wall temperature of 1500 K the power dissipation is

[11] See J. J. Opstelten, D. Radielović and J. M. P. J. Verstegen, this issue, p. 361.

between 15 and 20 W per cm^2. A lower wall loading, and hence a lower wall temperature, results in a lower luminous efficacy for all diameters and gas compositions.

This condition, implied by the maximum wall temperature, yields a relation between the length and diameter of the discharge tube for different values of the power dissipated in the lamp.

If so much xenon or mercury is added that the buffer gas determines the heat conduction in the lamp, the thermal conductivity of the plasma decreases to such an extent that the wall temperature drops by 50-75 K. The buffer gas thus allows a higher wall loading than could have been tolerated without buffer gas.

Lamp voltage

The electrical characteristics of a high-pressure sodium lamp are subject to two preconditions: the r.m.s. lamp voltage is predetermined, typically at half the value of the mains voltage. The re-ignition peak must not project too far above the plateau shown in fig. 7, so that re-ignition of the lamp is guaranteed throughout its life.

The re-ignition peak is found to be lower with increasing r.m.s. current, and the height of the peak also depends, as we have seen, on the pressure of the buffer gas. The lamp voltage is the sum of anode fall, cathode fall and column voltage. The electric field-strength in the column does not depend on its length; the r.m.s. value of the current, provided it is not too low, does not affect the field-strength either. The sole factors affecting the field-strength are the diameter of the discharge tube, the buffer-gas pressure and the sodium pressure. The value of the sodium pressure required for maximum luminous efficacy is a fixed quantity. The xenon pressure has very little effect on the field-strength, unlike the mercury pressure, which is determined by the composition of the sodium amalgam used for the dosage of sodium and mercury in the lamp.

The parameters available to us for fixing the lamp voltage are thus the length and diameter of the discharge tube and the composition of the amalgam. This yields, for a given amalgam composition, a second relation between the length and diameter of the discharge tube. We can now draw two sets of curves giving the length as a function of the diameter. In the first set the curves are drawn for given values of lamp power, and in the other for various mixing ratios of the vapour. This mixing ratio follows in turn from the composition of the amalgam. The optimum length and diameter of the discharge tube are now found from the points of intersection of these curves.

Summary. The discharge in a high-pressure sodium lamp can be described on the basis of a model consisting of the continuity equations for the particle densities, the energy-balance equation and the equation of state for the gas. Because of the high pressure it is sufficient for most calculations to assume that the gas temperature and electron temperature are identical. Most problems are encountered in calculating the radiative energy transport. The buffer gas — mercury and xenon — added to the discharge to control the thermal and electrical properties of the plasma is shown to have some influence on the spectral distribution of the emitted radiation.

The insight into the operation of a high-pressure sodium lamp, obtained from experiment and theory, can be very useful in designing high-pressure sodium lamps for which the lamp voltage, the wall temperature and the power dissipation have been specified.

A traffic intersection in Cherbourg, illuminated by 96 metal-iodide lamps (HPIT 1000 W). The average illuminance at the road surface is 60 lux.

Investigation of gas discharges by light scattering

In investigations of gas discharges [1], it is desirable to make measurements of particle concentrations and temperatures *as a function of position*, e.g. to verify a theoretical model or to follow changes in gas composition. The most suitable conventional methods are optical spectroscopy and interferometry: good results can be achieved with beams about 1 mm in diameter. These methods have, however, the disadvantage that only *integrated* results are obtained; e.g. when absorption is measured, the quantity obtained is the absorption of all of the gas in the cylindrical region described by the beam. The local effect can in principle be calculated as a function of the radius r by applying an Abel inversion to the integral effect, measured as a function of the transverse coordinate x (*fig. 1*). However, this is only true for gas columns of perfect cylindrical symmetry. The inversion can also turn small errors of measurement into serious errors in the final result. In particular it is virtually impossible to obtain information about the innermost parts of an optically dense plasma in this way. Measurements with beams parallel to the axis are in theory an alternative, but in practice are not much help because of the presence of the electrodes or other end effects.

We have achieved a substantially better spatial resolution by measuring the light scattering caused by the gas. A beam of laser light is focused on to the volume element under investigation, and the light scattered at right angles to it is collected by a second optical system (*fig. 2*). The intensity of the scattered light is proportional to the concentration of the scattering particles. The measurement is done with non-resonant light, for which the plasma is optically thin. In general the collected scattered light is an extremely small fraction (10^{-11} to 10^{-15}) of the incident light and is also only a small fraction of the continuum background light emitted by the discharge itself. To distinguish the measured signal from all the many forms of unwanted signals it is therefore necessary to use every available means (double monochromator, a chopper combined with phase-sensitive detection, and laser-beam focusing to about 25 μm).

There are various kinds of light scattering. 'Raman scattering', at temperatures up to 2000 K, from molecules in flames [2] and in gas discharges [3] and 'Thomson scattering' from electrons in strongly ionized plasmas [4] are examples of methods already used in plasma diagnostics. Our investigation relates to Rayleigh and Raman scattering from atoms in high-pressure discharges at temperatures up to 6000 K.

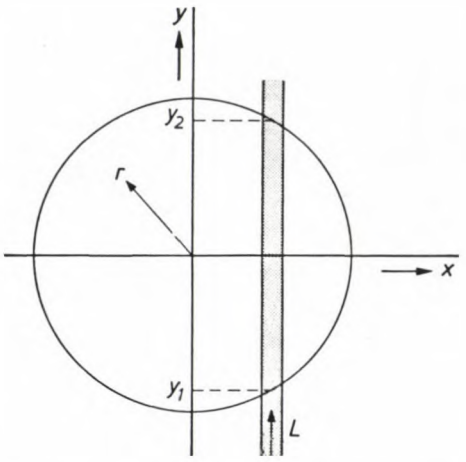


Fig. 1. Optical absorption measurement, schematic. The circle represents the gas discharge in cross-section. L laser beam. The quantity measured is the absorption integrated over y from y_1 to y_2. If this integral is measured accurately enough as a function of x, and given complete cylindrical symmetry, the absorption can be derived from it as a function of r by an Abel inversion. This procedure, however, may lead to considerable errors in the final result.

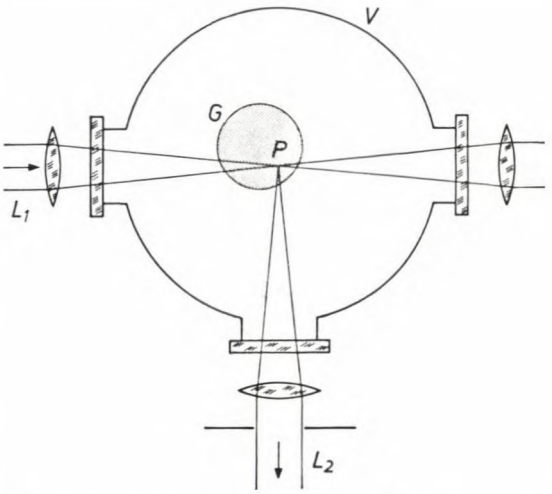


Fig. 2. Light-scattering measurement, schematic. G cross section of the discharge tube in the vessel V. The light of a laser beam L_1 is focused at the point P under investigation, and the scattered light L_2 is selected by a second optical system. The intensity of L_2 is proportional to the concentration of the scattering particles in P.

[1] R. Bleekrode, M. Koedam and L. Rehder, this issue, p. 308.
[2] M. Lapp, L. M. Goldman and C. M. Penney, Science **175**, 1112, 1972.
[3] L. Y. Nelson, A. W. Saunders Jr., A. B. Harvey and G. O. Neely, J. chem. Phys. **55**, 5127, 1971.
[4] D. E. Evans and J. Katzenstein, Rep. Progr. Phys. **32**, 207, 1969.
[5] L. Vriens, Phys. Rev. Lett. **30**, 585, 1973.
L. Vriens and M. Adriaansz, J. appl. Phys. **45**, 4422, 1974.
[6] L. Vriens, Optics Comm. **11**, 396, 1974.
L. Vriens and M. Adriaansz, Optics Comm. **11**, 402, 1974, and J. appl. Phys. **46**, 3146, 1975.

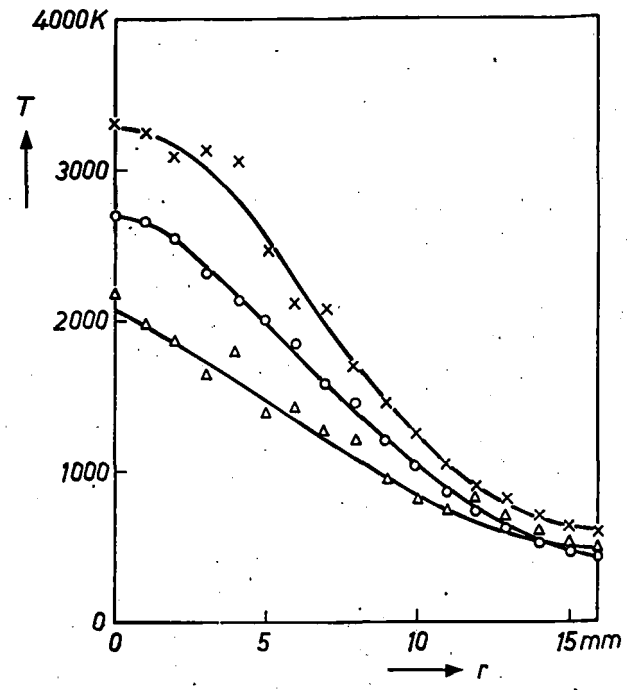


Fig. 3. The temperature T as a function of the radius r in discharges in neon of 10^5 Pa (1 atm), measured with Rayleigh scattering. The currents were 1 A (△), 2 A (○) and 3 A (×).

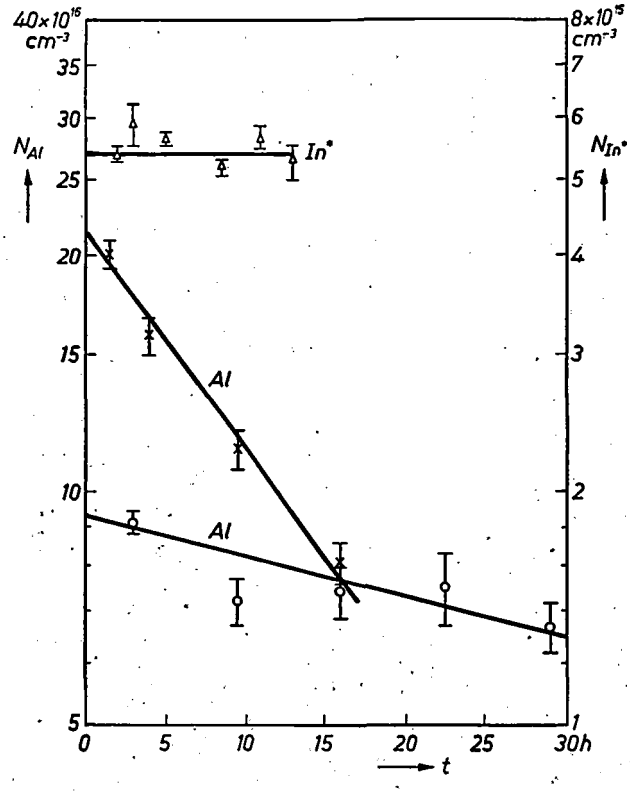


Fig. 4. The densities of aluminium atoms ($^2P^0_{1/2} + {}^2P^0_{3/2}$) and indium atoms ($^2P^0_{3/2}$) along the axis of metal-halide discharges as a function of operating time. The chemically active aluminium is removed from the discharge by reactions with the quartz-glass wall and possibly with the electrodes.

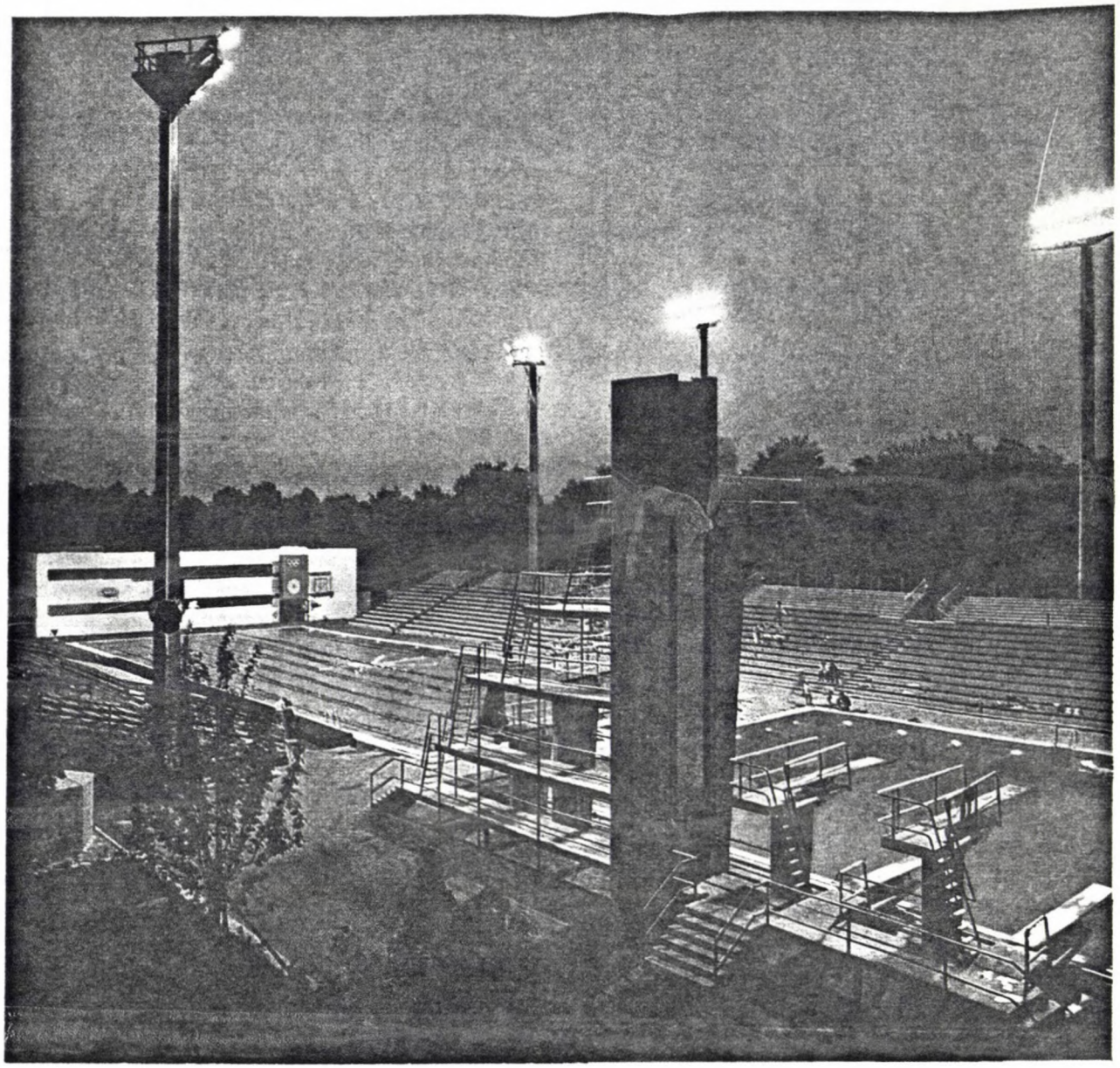
Every heavy particle gives rise to some Rayleigh scattering, in which the wavelength of the light remains unchanged. We have used Rayleigh scattering to study discharges in neon, argon and xenon [5]. Since the pressure is constant, the measured density distribution can be translated directly into a temperature distribution. *Fig. 3* shows a result for discharges in neon of 10^5 Pa (1 atm). Spatial variations in temperatures of thousands of degrees Kelvin are difficult to determine in other ways. Rayleigh scattering is not selective, and all scattering other than that caused by the gas must be carefully avoided. The conventional glass or quartz-glass envelope cannot therefore be used with these measurements. The laser beam and the scattered light pass unhindered through a gap at the centre of the quartz-glass tube. The divided quartz-glass tube is contained in a large vessel (see fig. 2), filled with the gas to be investigated.

In Raman scattering a transition takes place from a metastable state to the ground state or vice versa. This is accompanied by a change of energy which is equal and opposite to the difference between the energy of the scattered photon and that of the incident photon. When indium is irradiated with the non-resonant argon-laser light at 457.9 nm, the scattered light contains a component at 415.8 nm. This means that an In atom has changed from the metastable $^2P^0_{3/2}$ state to the $^2P^0_{1/2}$ ground state. If the monochromator is set to 415.8 nm, it therefore only measures scattering by the metastable $^2P^0_{3/2}$ In atoms. The wavelength shift in Raman scattering makes it possible to analyse discharges in the conventional envelope. Raman scattering seems to be a promising diagnostic aid, particularly for the otherwise relatively inaccessible metal-halide discharges. For absolute measurements of particle densities it is necessary to know the effective scattering cross-sections. In a combined experimental and theoretical study we have determined these cross-sections for the atoms Al, Ga, In and Tl [6].

We have used the method described for determining absolute densities of Al and In along the axis of metal-halide discharges. Some of the results are shown in *fig. 4*. It can be seen that the indium concentration remains constant, whereas the aluminium concentration in the lamps shows a decrease. This is attributable to the greater chemical reactivity of aluminium, which disappears as a consequence of chemical reactions.

L. Vriens
M. Adriaansz

Dr L. Vriens and M. Adriaansz are with Philips Research Laboratories, Eindhoven.



The Dante swimming stadium in Munich. During the 1974 Olympic Games it served as an open-air training pool for the swimmers and as a competition bath for water-polo. It is illuminated by 40 HPI 2000-W metal-iodide lamps mounted on five concrete masts at a height of 27 m. The average illuminance is 800 lux.

Metal-halide discharge lamps

P. C. Drop, E. Fischer, F. Oostvogels and G. A. Wesselink

For more than thirty years the high-pressure gas discharge has been the obvious light source to choose for applications requiring a high level of luminance. The most familiar example is the high-pressure mercury lamp used for street lighting. Its luminous efficacy, the colour of the light and the colour rendering leave much to be desired, however, since on the one hand a great deal of energy is emitted in the ultraviolet, and on the other the visible radiation consists of only a few spectral lines, none too ideally located in the spectrum (see *fig. 1a*). Some improvement was obtained by means of a fluorescent coating on the outer envelope, but it remained desirable to add to the discharge other metals that emit a better spectrum. The great problem here, however, was that most metals do not have a high enough vapour pressure or they react too strongly with the quartz-glass wall.

Some years ago lamps in which this problem has been solved appeared on the market. These lamps are not filled with the pure metal but with a metal compound, usually a halide. The compound is more volatile for many metals (e.g. Tl, Sn), and for others (the alkali metals) it is less reactive than the metal itself; for the alkaline-earth and rare-earth metals the compound is both more volatile and less reactive. The compound is stable near the relatively cool wall. At the hot centre, however, it is dissociated, so that the free atoms are released for the production of radiation. These lamps still contain mercury, but it acts mainly as a buffer gas; the spectrum is largely determined by the other metals.

Ir P. C. Drop is with the Philips Lighting Division, Eindhoven; Dr E. Fischer is with Philips GmbH Forschungslaboratorium Aachen, Aachen, West Germany; Ir F. Oostvogels is with the Philips Lighting Division, Turnhout, Belgium; Drs G. A. Wesselink is with Philips Research Laboratories, Eindhoven.

This principle has opened the way to an entirely new range of applications, and much research is being devoted to finding the best mixture of halides for each particular application. In this article we shall first briefly consider a number of metal-halide lamps on the basis of their spectra and applications. We shall then go somewhat more deeply into some special physical properties of discharges in metal halides.

Table I. Luminous efficacy η, colour temperature T_c and general colour-rendering index R_a for some metal-halide lamps. For comparison the table also gives the figures for the high-pressure mercury lamp (with fluorescent layer). The data relate to 400-W lamps.

Lamp	Gas filling	η (lm/W)	T_c(K)	R_a	Spectrum
High-pressure mercury lamp HPLN	Hg	55	4000	45	fig. 1*a*
Metal-iodide lamp HPI	NaI, TlI, InI, Hg	85	4000 or 6000	65	three lines, fig. 1*b*
Scandium-iodide lamp	NaI, ScI_3, Th, Hg	85	4000	65	many lines
Dysprosium-iodide lamp	TlI, DyI_3, Hg	75	6000	85	many lines, fig. 1*c*
Tin-halide lamp	SnI_2, $SnCl_2$, Li or In, Hg	60-70	3000-6000	90-95	continuum, fig. 1*d*

Types of lamp, spectra and applications

Since the metal-halide lamps offer a much wider range of spectral power distributions to choose from, they allow much better combinations of luminous efficacy η, chromaticity coordinates (x,y) and colour-rendering index R_a than the high-pressure mercury lamp; see *Table I*. If the choice were entirely free, the first thing would of course to be ensure that all the energy was emitted in the visible part of the spectrum. The actual distribution of the energy in this region, however, is a question of priorities. If the utmost in luminous efficacy is required, for example, it is usually at the expense of the colour rendering, and vice versa. In addition it is always necessary to specify the desired colour temperature T_c of the (white) light [1].

In practice metal-halide lamps may be roughly divided into two categories (see Table I): lamps with a *line spectrum* originating from atoms and lamps emitting a *quasi-continuous spectrum* from molecules. The first category includes the Philips HPI lamps (see *fig. 2*). These contain iodides of the elements sodium, thallium and indium. The dominant lines in the spectrum are the yellow line of sodium, the green line of thallium and the blue line of indium (fig. 1*b*). The

[1] This theoretical problem of optimum spectra for light sources is discussed in the article by J. J. Opstelten, D. Radielović and J. M. P. J. Verstegen, this issue, p. 361.

wavelengths and intensity ratios of the lines are chosen in such a way that these lamps outperform the high-pressure mercury lamp not only in luminous efficacy but also in colour rendering in spite of the large 'holes' in the spectrum. In lamps containing scandium or a rare-earth metal like dysprosium, the multiline spectra of these additives fill the holes in the spectrum between a number of resonance lines (fig. 1*c*). For ideal colour rendering the visible part of the spectrum should correspond to the continuous spectrum of the 'Planckian radiator' with the desired colour temperature. This can be approximated by the quasi-continuous radiation from molecules that remain undissociated at very high temperatures, such as the halides of lead, aluminium or tin. This is the principle underlying the 'tin lamp' (fig. 1*d*), which is at present in an advanced state of development at Philips and will be dealt with here in more detail.

The data shown in Table I give a first indication of the range of applications. However, when a lamp has to be chosen for a particular application, other factors have to be taken into account, of course. These include chromaticity coordinates, colour constancy, reproducibility, ignition devices, operating position, life and cost.

The Philips HPI lamps are used for lighting roads, streets, residential and shopping areas, and also for interior lighting in places such as industrial premises.

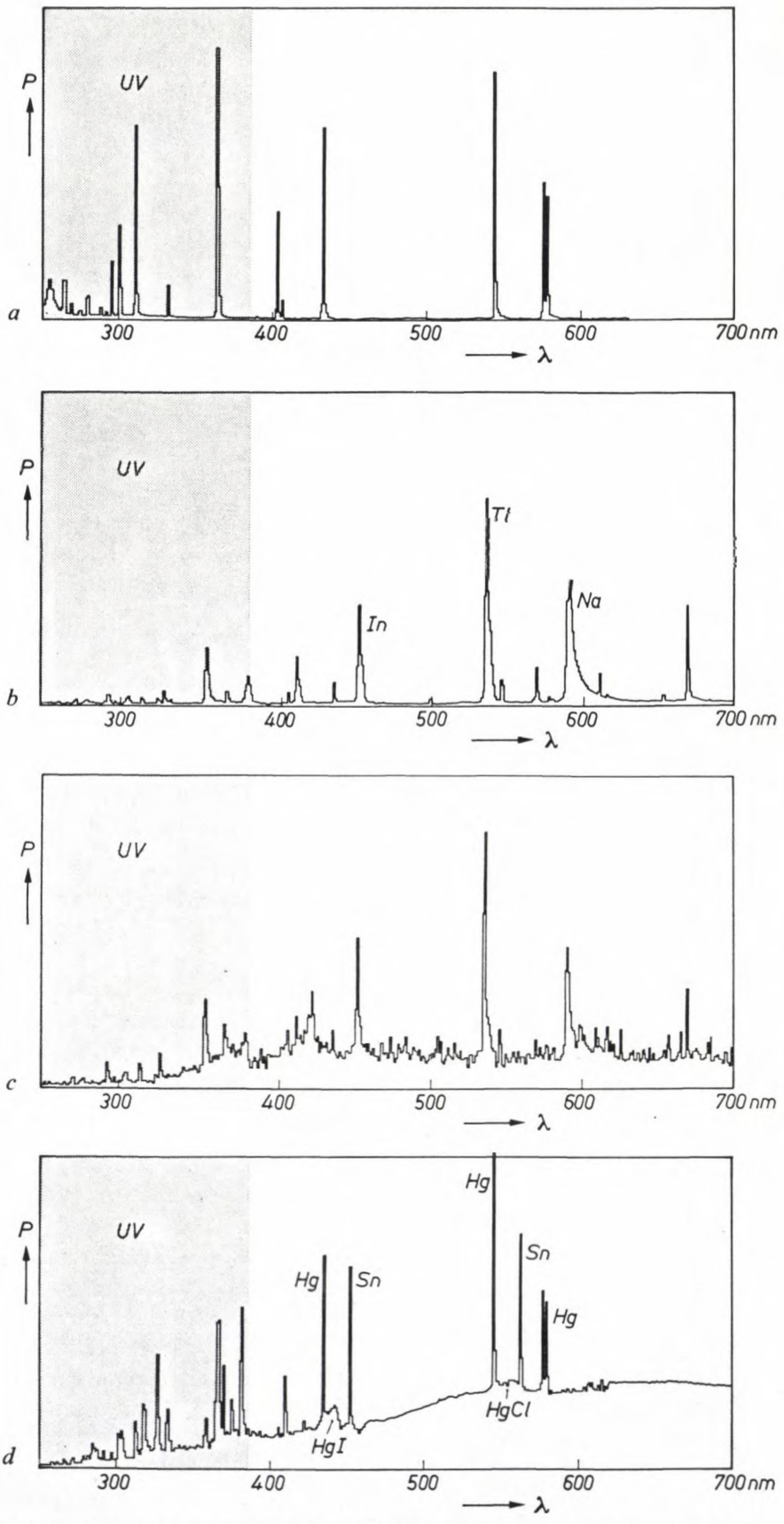


Fig. 1. Spectral power distributions of *a*) a high-pressure mercury lamp, *b*) a metal-iodide lamp, *c*) a dysprosium-iodide lamp, and *d*) a tin-halide lamp, all 400-W lamps. The vertical scales for the power are arbitrary and not the same for *a*, *b*, *c* and *d*. Tin-halide lamps usually contain lithium or indium as well (see p. 350); the spectrum (*d*) is from a lamp without these elements.

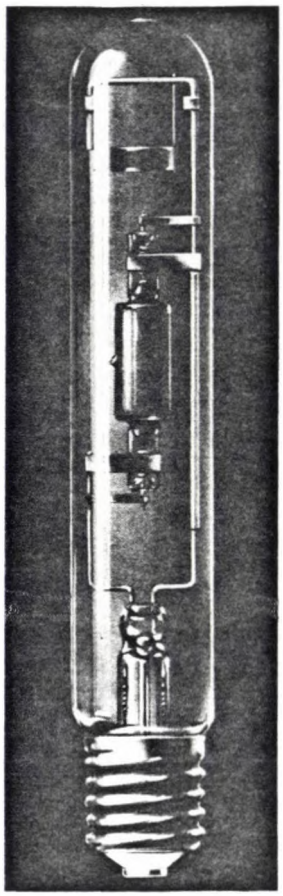


Fig. 2. High-pressure metal-iodide lamp (HPI). These lamps are made with a tubular clear envelope, as shown here, or with an elliptical outer envelope with a diffusing layer (see p. 310 of this issue).

The colour temperature used here is in the region of 4000 K. The HPI lamp is a great success in the lighting of sports fields, where it has made colour-television transmissions possible. For the average football stadium some 400 2-kW HPI lamps are used, giving the required level of 1500 lux in a plane perpendicular to the axis of the camera. In this application a colour temperature of 6000 K is sometimes used. Philips metal-halide lamps were also used at the Olympic Games in Munich in 1972. For interior lighting it is necessary to take into account the high luminance of the HPI lamp, which can cause glare from low ceilings. In places where colour rendering is particularly important, such as showrooms, the tin lamp seems to be the ideal solution, especially since the colour temperature of this lamp can be selected at will. Because of their high luminance and, compared with the incandescent lamp, their long life, HPI lamps are also used in professional overhead and episcope projectors. The tin lamp seems to be the most promising light source for professional slide projection.

After this brief survey we shall now take a closer look at the tin-halide lamp. HPI lamps have already been dealt with in some detail in this journal [2].

The tin-halide lamp

In most metal-halide lamps the halide molecules are only there to ensure that the metal is transported to the core of the discharge. Molecules like the monohalides of lead, aluminium and tin, however, are so stable that even at very high temperatures they are not fully dissociated, and make a significant contribution to the radiation from the discharge itself. The best results have been obtained with the tin-halide discharge, and we shall confine the discussion to that case.

The halide may be the iodide, bromide or chloride; the fluoride is too chemically active. Free chlorine would attack the tungsten of the electrodes and shorten the life of the lamp. There must therefore always be sufficient tin iodide present; with each free chlorine atom this produces tin chloride and iodine, both of which are less active. We have obtained the highest luminous efficacy with a mixture of tin chloride and tin iodide.

The spectrum of a tin lamp (fig. 1*d*) contains tin and mercury lines, and also two narrow molecular bands, at 445 and 555 nm, due to HgI and HgCl, but most of the radiation is emitted as a continuum that extends over the complete visible spectrum. A small part of this continuum is attributed to recombination of electrons with tin ions, but the major part is due to the molecular radiation of the tin halide. This radiation is associated with electron transitions whose initial and final levels are split by the many rotational and vibrational states of the molecule in such a way as to give rise to many bands each consisting of many lines. In a high-pressure discharge these are broadened to such an extent that a continuum is emitted.

In the case of tin iodide, observations with a *low*-pressure discharge, in which the separate bands could still clearly be distinguished, enabled us to identify the molecular radiation as originating from tin mono-iodide [3]. This radiation is emitted mainly from a cylindrical sheath around the axis; in the immediate neighbourhood of the axis the SnI is strongly dissociated.

It is interesting to consider how the excited SnI molecules come to be formed in this sheath. To some extent it is undoubtedly due to ordinary thermal excitation in this region, but it is probable that the following process also makes an important contribution. At the centre of the discharge excited tin atoms are formed. Now two of the excited states of tin are metastable and have a particularly long lifetime — according to our measurement of the order of milliseconds. Because of

[2] L. B. Beijer, C. A. J. Jacobs and T. Tol, Philips tech. Rev. **29**, 353, 1968.

[3] E. Fischer, R. Lorenz and L. Rehder, Lichttechnik **24**, 513, 1972.

this the excitation energy can be transported to the sheath, either by diffusion of these metastable tin atoms or by series of excitation collisions with tin atoms in the ground state. At the lower temperatures in the sheath the metastable tin atoms then combine with iodine to form excited SnI molecules.

We noted earlier that the colour temperature of the tin lamp can be chosen at will within a wide range — in fact from less than 3000 K to 6000 K. This particularly useful feature is due to the tin *di*iodide in the lamp. The diiodide is situated in a thin layer near the wall, where it acts as a filter that absorbs the blue more than it does the red (*fig. 3*). A higher concentration thus makes the colour temperature lower. This concentration can indeed be controlled by means of the lamp filling, because the components in this lamp, unlike those in many other halide lamps, are completely evaporated during operation. In our experience the colour temperature is entirely determined by the concentration of SnI_2. An increase in the $SnCl_2$ content has the effect of increasing the luminous efficacy, since more $SnCl_2$ produces more radiation but very little extra absorption, because the absorption due to $SnCl_2$ is much lower in the visible spectrum than that due to SnI_2 (see fig. 3).

Finally, we ought to mention that our tin lamps also contain lithium or indium, or both [4]. This is because the light from a pure tin-halide discharge has a rather greenish hue; the appropriate addition of red light from lithium (610 and 671 nm) and blue light from indium (410 and 451 nm) gives the desired white light — the colour temperature is lowered by lithium and increased by indium. The addition of this particular radiation also improves the colour rendering.

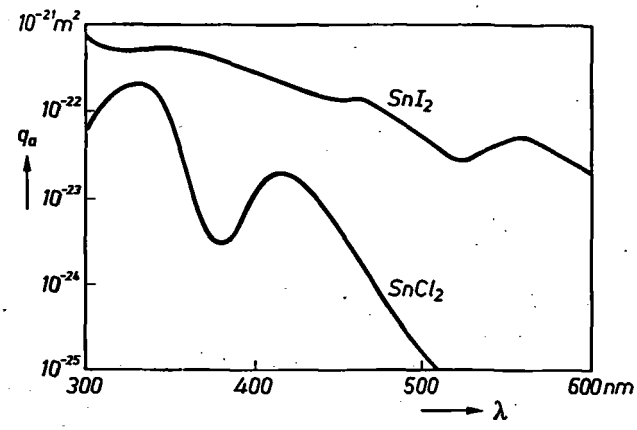


Fig. 3. Effective cross-section for photon absorption q_a of SnI_2 and $SnCl_2$ as a function of wavelength. The fact that the absorption of light by SnI_2 in the blue part of the spectrum is greater than in the red part is turned to advantage to give the colour temperature of the tin lamp a desired value.

Special physical properties

In this section we shall consider some characteristic physical differences between discharges in metal halides and those in mercury vapour. Our discussion will mainly be based on results relating to a model system containing only Sn and I, and which thus — unlike the tin lamp — contains no Hg and no Cl. This system has certain special properties that are characteristic of the tin lamp and also to a greater or lesser extent of other metal-halide discharges.

Apart from the inert gas, which we shall ignore for the moment, a mercury-vapour discharge contains particles of only three types: mercury atoms, mercury ions and electrons. In a pure Sn-I system, on the other hand, we have at least nine kinds of particle: Sn_2, I_2, SnI and SnI_2 molecules, Sn and I atoms, Sn^+ and I^- ions and electrons (e^-). All these particles react with each other and diffuse amongst each other. In spite of this complexity we have managed to draw up an adequate model for carrying out calculations on discharges in tin iodide and other metal halides [5]. *Fig. 4* shows one of the results obtained for the Sn-I system.

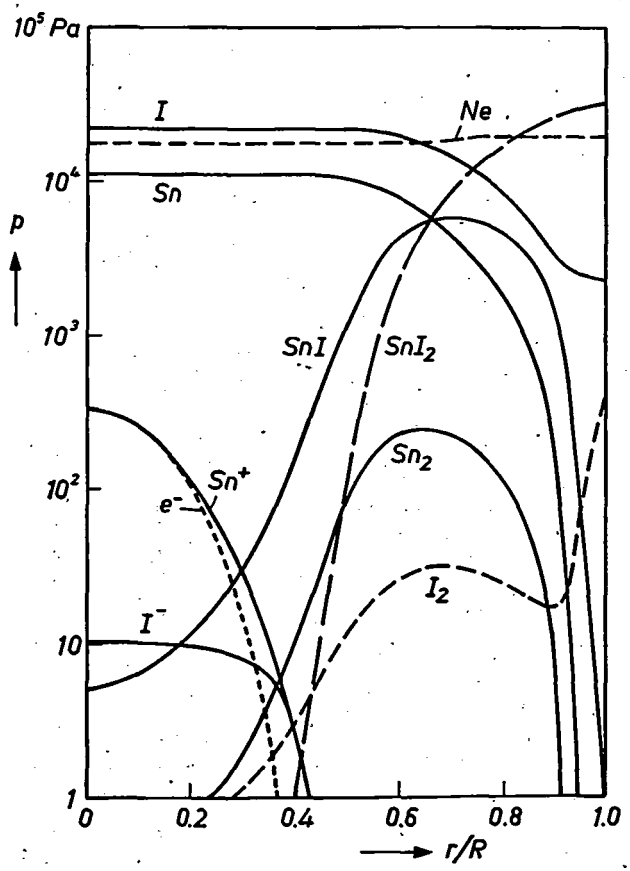


Fig. 4. Composition of the plasma in an Sn-I discharge determined from model calculations [5]. The partial pressure p of the components is plotted as a function of the distance r from the axis (R radius of the tube). The continuum of fig. 1*d* consists to a small extent of recombination radiation ($e^- + Sn^+ \rightarrow Sn$) from the centre of the discharge, but most of it is molecular radiation from SnI in a sheath around the axis.

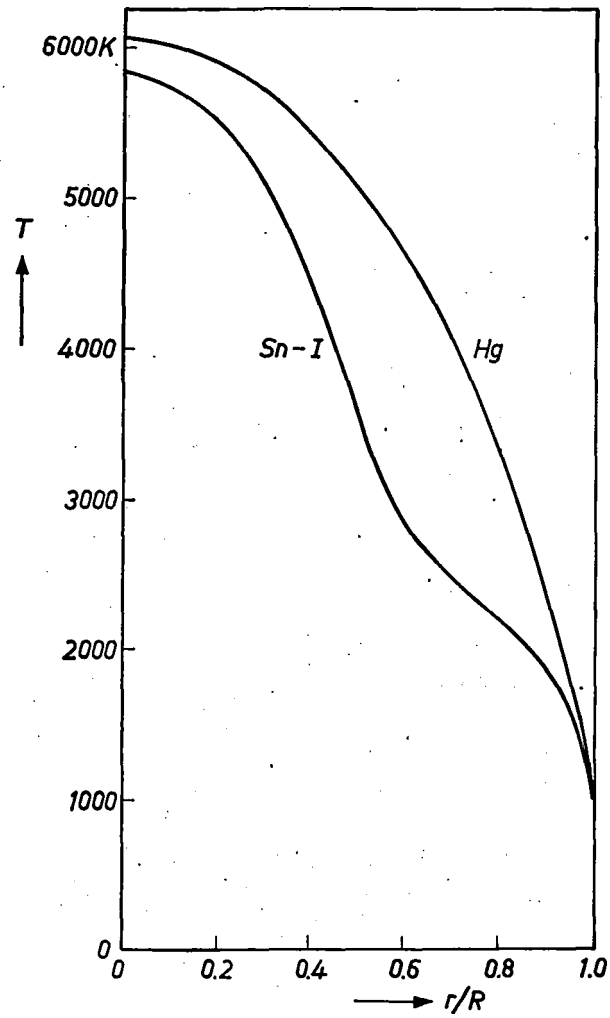


Fig. 5. The temperature profile in a tin-iodide discharge. The calculated temperature T is plotted as a function of the distance from the axis in a mercury plasma and in a pure tin-iodide plasma, for discharges of 60 W per cm in a tube with a radius R of 7.5 mm. The profile is narrower in tin iodide than in mercury.

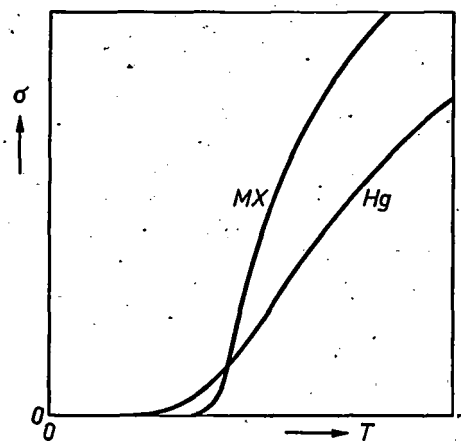


Fig. 6. Electrical conductivity σ as a function of temperature T for a metal-halide plasma (MX) and a mercury plasma (Hg), schematic.

In these calculations we proceeded from the hypothesis of local thermal equilibrium (LTE) [6]. It is thus assumed that in each small volume element there is one particular temperature T, which determines the reaction equilibria. The differences in T from one place to another cause shifts in the equilibria, giving rise to concentration gradients. The resultant diffusion currents prove to have a far-reaching influence, especially on the energy transport. In the case of tin iodide, for example, the SnI_2 and SnI molecules diffuse towards the centre of the discharge, where they dissociate into Sn and I atoms. These atoms then diffuse back towards the wall and recombine to form molecules. This cyclic process obviously transports dissociation energy outwards from the centre. In the pure tin-iodide discharge this energy flow is found at some places to be no less than 75% of the total outward energy flow. In a similar manner there is an outward flow of excitation and ionization energy. The energy-balance equation [7] must certainly therefore include a term for this kind of energy transport.

We shall not give a step-by-step account of these calculations, but merely discuss the principal result, which has also been experimentally verified, in more detail. This is the *narrowing of the current and temperature profile* compared with that of the mercury-vapour discharge. This narrowing can be seen in *fig. 5*: whereas the temperature profile for a mercury-vapour discharge has the familiar parabolic shape, the region of higher temperatures in a tin-iodide discharge is more closely concentrated. This may be explained in qualitative terms by the following two facts.

In the first place the curve of electrical conductivity as a function of temperature, $\sigma(T)$, is steeper for metal halides than for mercury (*fig. 6*). At temperatures below about 3500 K the electrical conductivity in the metal halides is very low because no electrons are formed, only negative ions. Owing to their low mobility these ions make very little contribution to the conductivity. At higher temperatures the electrons break away from the negative ions. Since the ionization potential of the metals in question is lower than that of mercury, the electrical conductivity now increases much faster with temperature than that of the mercury plasma.

In the second place the thermal conductivity κ in metal halides is much greater than in mercury vapour, especially at relatively low temperatures (*fig. 7*).

[4] P. C. Drop, J. J. de Groot, A. G. Jack and G. C. J. Rouweler, Lighting Res. Technol. 6, 212, 1974.
[5] E. Fischer, J. appl. Phys. 45, 3365, 1974.
[6] See the article by R. Bleekrode, M. Koedam and L. Rehder in this issue, particularly p. 312.
[7] See the article by R. Bleekrode, M. Koedam and L. Rehder in this issue, equations (1) and (2).

Whereas only two mechanisms contribute significantly to the heat conduction in mercury-vapour discharges, namely the transport of kinetic energy (classical heat conduction) and the transport of photons via resonance lines (radiative transport [8]), in metal halides the transport of dissociation energy via diffusion, as mentioned above, also makes a very important contribution. This gives a peak in the thermal conductivity at relatively low temperatures.

It is evident that these two effects lead to the narrowing we have mentioned. The steep increase in σ with temperature, and hence with the distance to the axis, implies a narrow current channel. The high heat conduction in the region of relatively low temperatures, i.e. in the zone around the current channel, leads to a small gradient in this zone: only a small gradient is necessary to remove the power generated at the centre.

The narrowing of the discharge has important practical consequences. When the temperature profile is parabolic the line of maximum temperature coincides with the axis of the discharge tube ('wall stabilization'). In a narrowed discharge, however, the zone with a small temperature gradient acts as a buffer separating the current channel from the wall. This means that the channel is much more sensitive to disturbances. We shall give two examples.

In the first place, interfering magnetic fields will cause a greater displacement in narrowed discharges than in an ordinary mercury discharge. In particular the magnetic field of the discharge itself may cause instability, especially since this field is higher in the narrowed discharge because of the higher current density. The Lorentz force resulting from this field is oriented towards the axis, and in a completely straight discharge will only give a small increase in gas pressure along the axis. Any curvature of the discharge, however, due for example to electrode effects or convection, will be reinforced by the magnetic field of the discharge, since this field, and hence the Lorentz force, is greater on the inside of the curvature than on the outside (*fig. 8*). This may give rise to a bowed or helical and possibly rotating discharge, which may almost come into contact with the wall at some places. Where this happens the envelope can become so hot that it fails. This effect sets an upper limit to the halide concentration that can be used, and consequently sets a lower limit to the colour temperature of a tin lamp operating in a vertical position (see p. 350).

Secondly, in a lamp operating *horizontally*, the natural convection can cause a large upward displacement of the current channel, and this can cause overheating of the wall material at the upper surface of the lamp. This means that tin-halide and similar discharges cannot be operated horizontally without taking special

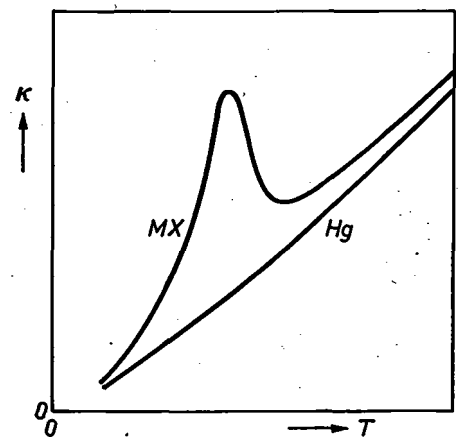


Fig. 7. Thermal conductivity κ as a function of temperature T for a metal-halide plasma (*MX*) and a mercury plasma (*Hg*), schematic.

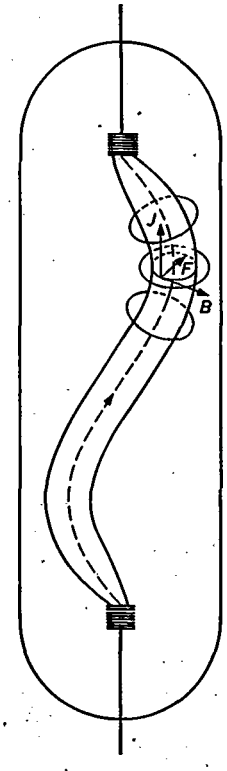


Fig. 8. Helical discharge due to magnetic instability. As a result of a magnetic field B caused by the discharge current the charge carriers at the edge of the current channel are subjected to a Lorentz force F, which is directed towards the axis. In a bowed discharge, both B and F are larger on the inside of the curve than on the outside, so that the discharge tends to become increasingly curved. This kind of instability is more likely to occur in metal-halide discharges than in the high-pressure mercury discharge because the current densities and magnetic fields are greater, but the wall stabilization is weaker than in the mercury discharge.

precautions. In our tin lamp for horizontal use (*fig. 9*) we have put the high sensitivity to a magnetic field to good use: an extra loop in the supply leads produces a weak horizontal magnetic field, which compensates for the upward force on the current channel. This solution has the additional advantage that the halide concentration is now no longer limited, so that very low colour temperatures can be obtained.

Finally, an effect that can occur in long vertical metal-halide discharges is the separation of the components of the discharge, accompanied by colour disproportionation or 'demixing'. To study such effects we made an Sn-I-Hg discharge tube that was three times longer than the normal tin lamp of about 4 cm. The

radiation emitted from the top of the tube consisted mainly of mercury light from the core, whereas the radiation from the lower part was mainly molecular SnI radiation from a broader sheath around the centre [9]. This is the result of a combination of convection and diffusion. At the centre of the tube there is an upward convection current of mercury, tin and iodine atoms. Along the wall a stream of mercury atoms and tin-iodide molecules flows downwards. The tin and iodine atoms diffuse out of the upward stream towards the wall where they recombine and enter the downward stream. The upward stream thus gradually loses its tin and iodine components. This applies to a much lesser extent to the downward stream, since the tin-iodide molecules, which are heavier, diffuse inwards at a much slower rate. The net result is that there is more tin iodide at the bottom of the tube than at the top. Clearly, this process can only occur if, in spite of this material transport, the total pressure can be kept constant throughout the tube by other gas components, in this case mercury.

The axial demixing can in principle be counteracted by separating the upward from the downward stream, for example by means of a coaxial quartz-glass cylinder, which will prevent radial diffusion between the two streams [10].

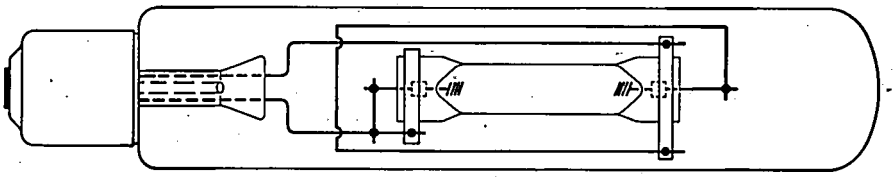


Fig. 9. Tin-halide lamp for horizontal operation. The supply lead forms a loop around the actual discharge tube. The resultant magnetic field gives a downward-directed Lorentz force on the charge carriers, which compensates for the upward force caused by convection.

Summary. Until recently metals other than mercury could not be used in high-pressure discharge lamps because of their low volatility and high reactivity with the quartz-glass wall. This problem has been solved by introducing halides of the metals. These are more volatile or less reactive, or both, than the metals. At the hot centre they dissociate, enabling the free metal to take part in the discharge. The resultant freedom in the choice of spectrum permits much better combinations of luminous efficacy and colour rendering than the mercury lamp. In the HPI lamp with a three-line spectrum the emphasis is on a high efficacy; in the tin lamp, which emits a continuous molecular spectrum, the emphasis is on good colour rendering. The spectrum emitted from the tin lamp originates mainly from SnI molecules in a sheath around the axis; the colour temperature can be freely selected through the choice of SnI_2 concentration. Various special physical properties of metal-halide discharges are connected with the fact that the current and temperature profile is narrower than that in a mercury lamp.

[8] See the article by R. Bleekrode, M. Koedam and L. Rehder in this issue, particularly p. 312.
[9] J. J. de Groot and A. G. Jack, J. Physics D 6, 1477, 1973.
[10] E. Fischer, shortly to be published in J. appl. Phys.

A getter for metal-iodide high-pressure mercury-vapour lamps

In the manufacture of high-pressure discharge lamps special precautions are necessary to prevent the excessive release of undesired gases in the operating lamp, whose wall temperature is about 800 °C. It is particularly important to limit the amount of hydrogen released, most of which is produced by the decomposition of water; nitrogen and carbon monoxide are also released, but in smaller quantities. All three gases increase the ignition voltage of the lamp. Hydrogen also has the effect of reducing the luminous efficacy as the time of operation of the lamp increases.

The discharge lamps in which these effects cause most concern are the high-pressure mercury-vapour lamps that use metal iodides, since these compounds are *hygroscopic*. The magnitude of the problem becomes apparent when we realize that as little as 10^{-4} mg of water in an HPI lamp with an inner-envelope volume of 10 cm^3, containing 35 mg of iodides of sodium, thallium and indium, will give an unacceptable increase in the ignition voltage.

At the Philips Lighting Laboratory we have made a study to find out the extent to which residual-gas content can be made acceptably low by the use of a getter, instead of outgassing during manufacture. Getters have been used in incandescent lamps for many years, and have more recently been used in the outer envelopes of high-pressure mercury-vapour lamps.

Introducing a getter into the *inner* envelope of a metal-iodide high-pressure lamp is a far more difficult problem, since there is no known getter material resistant to the atmosphere contained there. The solution described here was first tried out experimentally on 400-W HPI lamps, but is also suitable for other metal-iodide lamps.

Our approach is illustrated in *fig. 1*. A material (yttrium) with an affinity for hydrogen is surrounded by a material (tantalum) that is resistant to the lamp atmosphere and is sufficiently permeable to hydrogen, at least when the lamp is operating. These two metals have very little mutual solubility, even at the temperatures found at the wall of the operating lamp, and do not form intermetallic compounds.

Of all the metals with an affinity for hydrogen and a sufficiently low vapour pressure, yttrium forms the most stable hydrides. This means that a smaller quantity of yttrium than of any other metal is necessary to reduce the hydrogen pressure at the wall of the operating lamp to a value at which the hydrogen has no adverse effect on the ignition characteristics of the lamp. (This pressure is about 7 Pa, i.e. about 0.05 torr.) The quantity of hydrogen released in a 400-W non-outgassed HPI lamp at a wall temperature of 800 °C will be 5×10^{-7} mol at the most. It can be seen from *fig. 2* that 3.5×10^{-6} mol (0.3 mg) of yttrium is sufficient to prevent the hydrogen pressure from exceeding the value of 7 Pa in a lamp of this type.

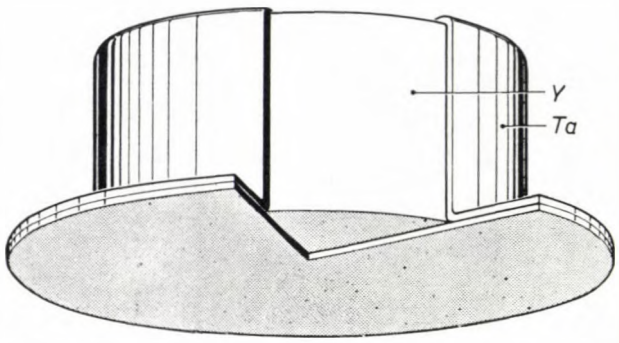


Fig. 1. The new getter for HPI lamps. A pellet of yttrium, which has an affinity for hydrogen, but would not be resistant to the atmosphere in the operating lamp, is surrounded by a protective shield of tantalum. The tantalum is sufficiently resistant and is permeable to hydrogen during the operation of the lamp.

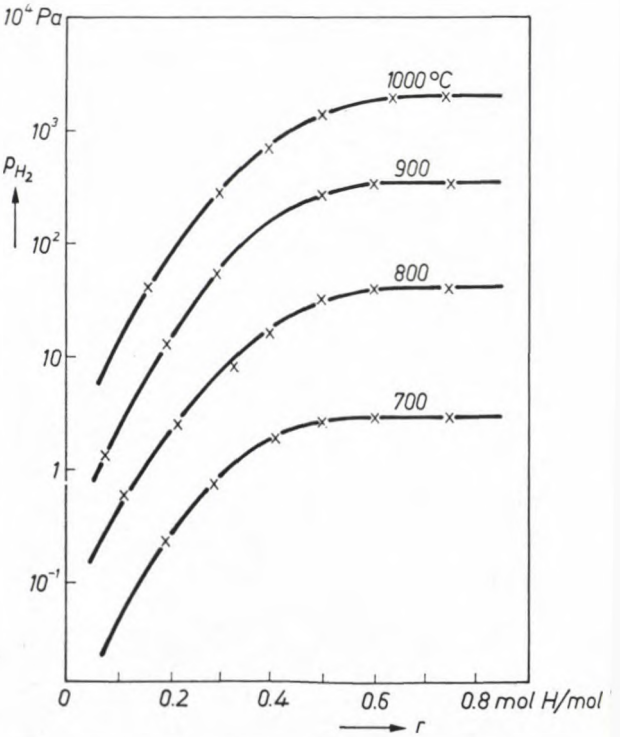


Fig. 2. The hydrogen pressure p_{H_2} as a function of the quantity r of hydrogen in contact with yttrium, at 700, 800, 900 and 1000 °C.

Other metals besides tantalum that are sufficiently resistant to be used for enclosing the yttrium are tungsten, molybdenum and chromium, but these are not so permeable to hydrogen as tantalum. An added advantage of tantalum is that it acts in itself as a getter for residual gases such as oxygen, nitrogen and carbon monoxide.

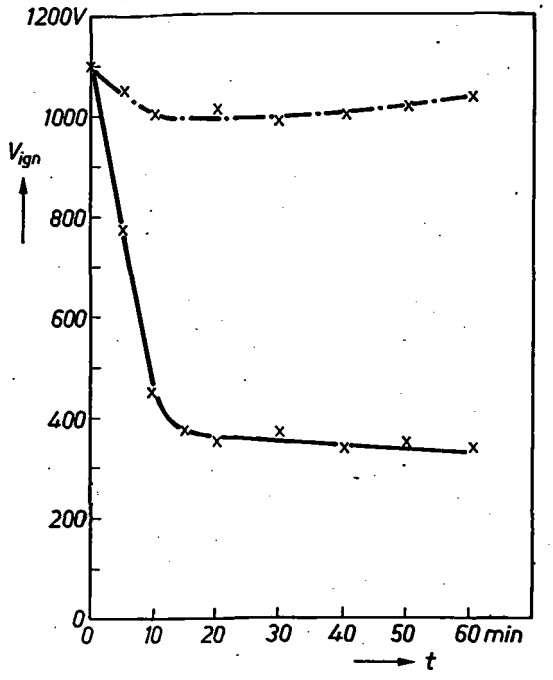


Fig. 3. The ignition voltage V_{ign} as a function of the operating time t for argon-filled 400-W HPI lamps made without outgassing, both with (solid curve) and without the getter described in the article. It can be seen that when the getter is used the ignition voltage decreases by a factor of three within 20 minutes.

The getter we have developed has been used to make HPI lamps that fully meet the specifications, yet without the use of any outgassing procedures. One of our new getters, 4×10^{-3} cm^3 in volume, in a discharge tube of volume 10 cm^3, will reduce a residual pressure of 130 Pa of hydrogen to less than 10^{-1} Pa in a few minutes at 1000 °C. This performance is found to be more than adequate.

Fig. 3 shows that the ignition voltage of experimental lamps made in this way decreases by a factor of between 3 and 4 within only 20 minutes of operation (solid curve), generally to an acceptable level. The chain-dotted curve shows, for comparison, the ignition voltage of lamps made with no outgassing and no getter.

As well as the decrease in ignition voltage, there is also a drop in the voltage at which the lamp re-ignites after each half-cycle of the a.c. mains supply (from 180 to 50 V).

G. Kuus

Drs G. Kuus is with the Philips Lighting Division, Eindhoven.

Electrodes in discharge lamps

A. Bouwknegt, H. Nienhuis, D. J. Schipper and P. A. W. Tielemans

The life of a discharge lamp is mainly determined by its electrodes. The electrodes also affect the luminous efficacy and the radio-frequency noise caused by the lamp. In spite of the vital significance of the chemical and physical processes associated with the operation of electrodes, however, our understanding of them is still incomplete. Some typical properties and requirements of electrodes will be considered in this article, which is largely based on the results of recent investigations in various Philips laboratories.

Function and properties of the electrodes

Most discharge lamps are operated from an a.c. voltage supply, and we shall only consider this kind of of supply in this article. In a.c. operation the electrodes act alternatively as cathode and anode during each half-cycle of the mains voltage. Both electrodes thus have the same function.

The principal task of the electrode in the cathode phase is to emit electrons. In a steady-state discharge this takes place mainly by *thermionic* emission. The required heating of the electrode comes from ion bombardment in the cathode phase and electron bombardment in the anode phase, and also from Joule heating. Owing to electron emission, thermal conduction and radiation, heat is extracted from the electrode. During steady-state operation of the lamp the sums of the positive and negative terms are equal.

A good electrode has to meet three requirements. It must guarantee a long lamp life (this relates to the ignition of the lamp and blackening of the bulb), it must not introduce appreciable losses and it must give the minimum of radio-frequency noise. We shall now take a closer look at what this implies.

For good ignition and long life the *electron* emission from the electrode is important: the emitter must have a low work function, which should not be unduly increased by chemical reactions, by sputtering due to strong ion bombardment, or by evaporation. Emitter materials that have a low rate of evaporation often have a high work function, however, and this implies higher electrode losses. The first two requirements are thus contradictory, and the emitter therefore has to be chosen very carefully to meet the conditions for the particular case in question.

Because of the thermal delay of the electrodes, the electron emission during a part of the half-cycle in which the electrode functions as a cathode is greater than is necessary for the discharge current (see sections *1* and *3* of the curve in *fig. 1a*). This gives a negative space charge at the cathode, and this space charge limits the electron current (fig. 1*b*). In section *2* of the curve, however, the discharge current requires more electrons than the cathode can readily supply. As a result the negative space charge disappears and a positive electric field takes its place (fig. 1*c*), causing an increase in the electron emission from the cathode. With the disappearance of the negative space charge the positive ions carry a greater part of the discharge current (the difference I_i between the lamp current and the emission current, see fig. 1*a*). As the space-charge region becomes thinner there is less chance that the ions will lose the energy they have taken up in the cathode fall by collision with gas atoms, and they are therefore able to give up more energy to the electrode Thus, the higher the value of I_i, the more intense the ion bombardment to which the electrode is subjected While a low ion bombardment may have an activating effect on electron emission — by improving the desirable physico-chemical properties of the cathode surface [1] — heavy ion bombardment seriously impairs the quality of the electrode by causing sputtering of emitter material. In this connection the *method of igniting* a discharge lamp is important. In fluorescen

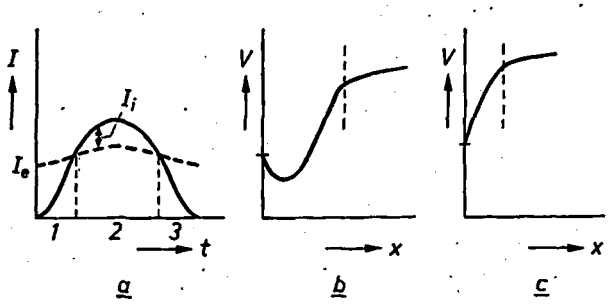


Fig. 1. *a*) Schematic waveform of the lamp current *I* in a ha cycle of the supply voltage. I_e electron-emission current from t cathode. I_i ion current. *b*) Voltage variation in front of the catho as a function of distance *x* during the parts *1* and *3* of the lam current curve, in which a negative space charge is built up due surplus emission. There is hardly any ion bombardment of t cathode. *c*) Voltage variation in part *2* of the lamp-current cur where the emission is intensified by an electric field. In this sitt tion the cathode is subject to the strongest ion bombardment.

Dr A. Bouwknegt is with the Philips Lighting Division, Roosendaal; Drs D. J. Schipper is with Philips Research Laboratories, Eindhoven; Ir H. Nienhuis and Ir P. A. W. Tielemans are with the Philips Lighting Division, Eindhoven.

lamps the electrodes are briefly preheated to facilitate ignition. SOX lamps and high-pressure lamps, on the other hand, are ignited with cold electrodes. In this case there is a heavy ion bombardment at first, which is greatly reduced as soon as the electrode reaches the temperature at which electrons are emitted. The electrodes therefore have to be designed to reach a high temperature quickly.

Electrode losses arise because heat is lost from the electrode by radiation, conduction in the electrode itself and conduction into the gas. Low losses are obtained with low electrode temperatures, which means that a low work function is required. The losses depend to a great extent on the geometry of the electrode [2]. To reduce heat conduction a small wire diameter must therefore be used, and although small dimensions give higher temperatures and hence greater radiation losses, the net effect — particularly with electrodes in high-pressure lamps — is a reduction of the losses. Since unduly high temperatures may impair the stability of the emitter, however, the dimensions of the electrode must be chosen in such a way that the heat supplied to the electrode — mainly in the anode phase — ensures a temperature level at which there is just the right amount of electron emission in the cathode phase.

Although a periodic negative space charge (fig. 1*a*, sections *1* and *3* of the curve) prevents heavy ion bombardment of the electrode, it also constitutes a source of *radio-frequency noise* [3] — the 're-ignition-extinction noise' — at frequencies in the region of 500 kHz, which often occurs in fluorescent lamps. The reason for this is that the space charge forms a potential well in which positive ions enter into oscillation (fig. 1*b*), and it is these oscillations that cause the noise. If the emission is reduced on this account, however, the life of the electrode is shortened by the resultant heavier ion bombardment.

Another form of radio interference, caused by anode oscillations, occurs when the active part of the electrode during the anode phase is too small to accumulate sufficient electrons. The anode potential then rises above that of the plasma (anode fall). The effect of this is to accelerate the electrons, causing additional ionization of gas atoms, with the result that the density of the electron current increases and the anode accumulates sufficient electrons. The anode potential then drops again to that of the plasma, and the cycle is repeated. The radio interference arises at these drops in anode potential.

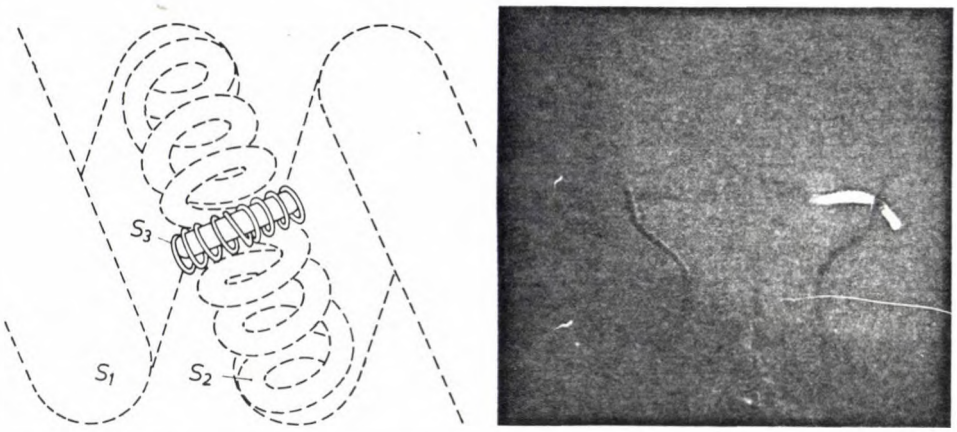


Fig. 2. Electrode of a fluorescent lamp. *Left:* detail of electrode construction (schematic, magnification about 30×; the emitter material is not shown). S_1, S_2 coiled-coil tungsten filament. S_3 separate coil of tungsten filament for better cohesion of the emitter material. *Right:* the electrode in operation.

The above brief description of possible causes of radio-frequency noise, which have recently been investigated, makes it clear that the composition and operation of the electrodes are important. A study of these effects [4] has shown that one way in which radio noise can be greatly reduced is to surround the electrodes with a cylindrical metal grid [5].

We shall now consider the various points raised here in a more detailed examination of the different kinds of electrodes used in discharge lamps.

[1] A. Bouwknegt and A. G. van der Kooi, J. Physics D **8**, 952, 1975.
[2] P. A. W. Tielemans, Measurement of electrode losses in high-pressure gas-discharge lamps, Conf. on Electrode phenomena in gas discharges, Bucharest 1974, paper 6.10.
[3] J. F. Waymouth, Electric discharge lamps, M.I.T. Press, Cambridge Mass. 1971, p. 93.
[4] D. J. Schipper and P. Zalm, unpublished.
[5] D. Kolkman and P. T. J. Piree, Dutch patent application 6514352.

Low-pressure discharge lamps

The electrode in the fluorescent lamp is a coil of tungsten wire coated with an emitter (*fig. 2*). The emitter consists of a mixture of oxides, mainly of Ba, Sr and Ca. The coating is applied by means of a suspension of carbonates which are converted into oxides by heating the coil during the lamp-manufacturing process:

$$MCO_3 \rightarrow MO + CO_2,$$

wnere M stands for Ba, Sr and Ca. This involves the following important subsidiary reaction:

$$3MCO_3 + W \rightarrow M_3WO_6 + 3CO.$$

The tungstate is important for the life of the electrode. It forms a layer between the oxides at the surface and the tungsten coil and regulates the production of free barium. The presence of free barium at the surface is necessary for a low work function. It can be formed from the reaction

$$6BaO + W \rightarrow Ba_3WO_6 + 3Ba,$$

and possibly also by the dissociation of BaO by ion bombardment.

The tungstate layer must have the right thickness and the right structure. If the layer is too thin or the structure too open, barium production will be excessive. As a result the barium quickly evaporates — so that barium spots form on the bulb wall near the electrode — and the electrode is very soon exhausted. If the layer is too thick or the structure too dense, barium production will be too low. This increases the work function, and the temperature of the electrode is then increased as a result of heavier ion bombardment, giving faster evaporation and sputtering. If the carbonate causes too much oxidation of the tungsten, there is a risk that too much oxygen will be released in the lamp during operation, resulting in the gradual formation on the wall of a light-absorbing layer consisting mainly of oxides of mercury.

Since the life of the electrode is partly determined by the amount of emitter, as much emitter material as possible must be stored in a suitable manner. A 'triple-coiled' filament is commonly used for this storage (see fig. 2): around the tungsten filament, which is itself a double helix (coiled coil) a third coil S_3 of very thin wire is included to improve the cohesion between oxide and tungsten base [6].

While the lamp is operating, the emission and collection of electrons do not take place uniformly over the whole surface of the electrode. A small bright spot in the filament appears at the place where the density of the emission current during the cathode phase is much greater than elsewhere. During the life of the lamp the position of this 'hot spot' usually shifts. During the anode phase the majority of the electrons arrive at the connection end of the filament.

The electrode losses differ from one type of lamp to another. The voltage drop at the electrodes, V_{el}, is generally about 15 V. The losses are then given by $V_{el}I$, and amount to about 6 W in the case of a 40-W fluorescent lamp.

The electrode in a low-pressure sodium (SOX) lamp, shown in operation in *fig. 3*, is generally required to supply a higher current than the electrode in a fluorescent lamp, and is therefore more rugged in construction. Its design is optimized for cold ignition. The emitter is the same as that on the fluorescent-lamp electrode, but in this case it is applied by electrophoresis. The electrode losses are 12 or 18 W, depending on the lamp current (0.6 or 0.9 A).

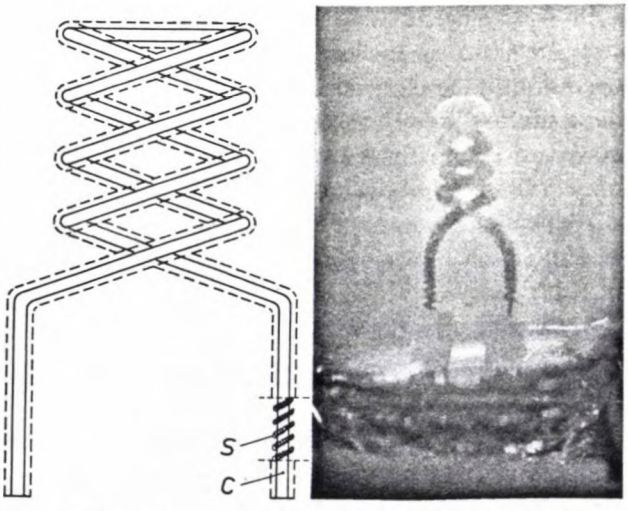


Fig. 3. *Left:* schematic representation of an electrode for an SOX lamp (magnification about 5×). Around a core *C* of tungsten wire, itself coiled, there is a coiled tungsten wire *S*. *Right:* the electrode in operation.

High-pressure discharge lamps

The electrode in high-pressure lamps is usually in one of the forms shown in *figs. 4* and *5*. Differences are found only in the case of electrodes in lamps for very low or very high powers. The electrode in fig. 4, which is used in high-pressure mercury (HP) and high-pressure sodium (SON) lamps consists of a tungsten rod *W* with a tungsten coil *S* wound around it in two layers. The emissive material is stored in the space between the turns, and consists in this case of $BaCO_3$, $CaCO_3$ and ThO_2. It has been found that ThO_2 acts as an emitter when the lamp is operating. The $BaCO_3$ and $CaCO_3$ are converted into Ba_2CaWO_6 during sintering; this tungstate is active on ignition. The emitter mixture used in the low-pressure lamp is not suitable

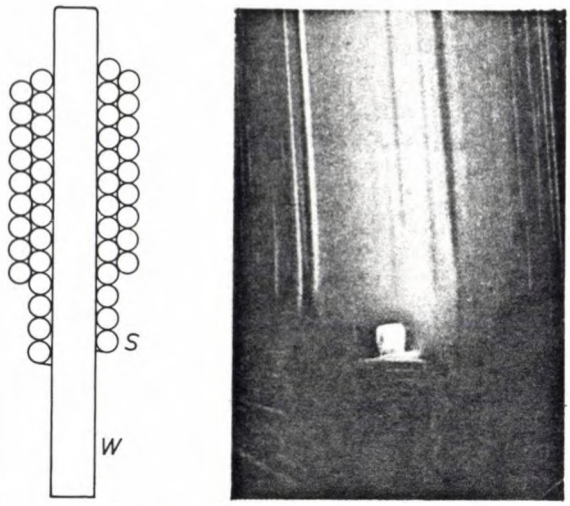


Fig. 4. Electrode of a 400-W HP lamp. *Left:* schematic representation of construction (magnification about 5×). *W* tungsten rod. *S* tungsten coil in two layers. *Right:* the electrode in operation.

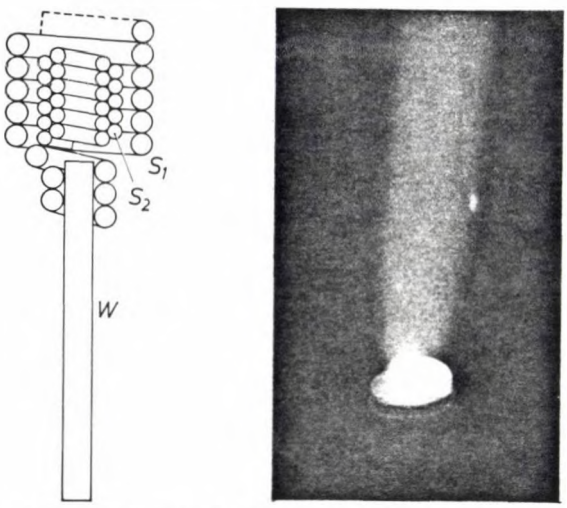


Fig. 5. Electrode of a 375-W HPI lamp. *Left:* schematic cross-section (magnification about 5×). *W* tungsten rod. S_1 tungsten coil. S_2 tungsten coil in two layers, located in the gap between *W* and S_1 and kept in place by the bent end of S_1. *Right:* the electrode in operation.

for high-pressure lamps, because the high gas temperature in such lamps (3000-6000 K) would make the electrode so hot, because of conduction, that the emitter would very quickly evaporate. ThO_2 evaporates much more slowly under these conditions, but the work function is higher.

In the HP lamp, as in the fluorescent lamp, a hot spot forms on the electrode. The end of the electrode has a temperature of about 2000 K, and the temperature at the hot spot is a few hundred degrees higher. In the SON lamp, on the other hand, the temperature profile is more homogeneous. The electrode temperature in this lamp is lower: 1800 K. This difference is probably due to the presence of a little sodium on the electrode surface, which plays an important role in the emission.

The electrode losses depend to a great extent on the geometry of the electrodes [2]. A different electrode is required for each of the nine types of HP lamp with power ratings from 50 to 2000 W. The losses in the HP 400-W lamp are 30 W, in the SON 400-W lamp they amount to 24 W.

The metal-iodide (HPI) lamp uses a different type of electrode (fig. 5), developed by Philips. This consists of a tungsten rod *W*, around which a single coil S_1 is first wound, followed by a smaller coil S_2 wound in two layers in the space between the rod and S_1. One reason for using this geometry is to prevent iodides from condensing on the electrode during the cooling of the lamp after switching off, which would interfere with the ignition later. The presence of iodine also explains why alkaline-earth oxides cannot be used for the emitter; they react too readily with iodine. The ThO_2 also used as emitter in this type of lamp has been found to meet the requirements, though it is only active during ignition. The temperature at the end of the electrode coil is 3000 K, from which it follows that the tungsten is the emissive material. The electrode losses in a 375-W lamp amount to 36 W.

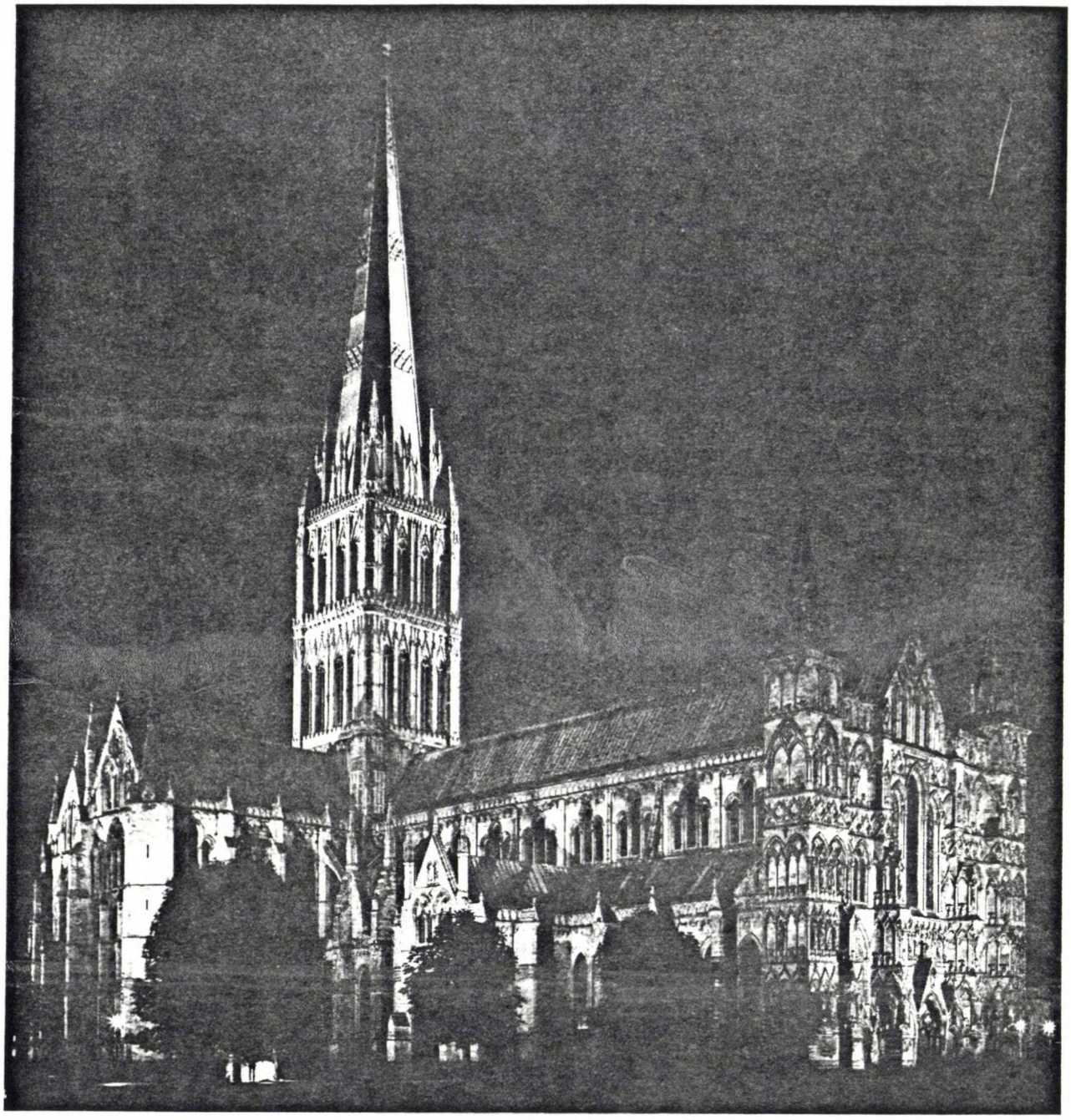
Finally, not even ThO_2 can be used in the tin-halide lamp, owing to its high reactivity with the chlorine gas present. So far tungsten has proved to be the only material that is reasonably resistant to such a reactive gas filling. The electrode temperature and losses are of the same order of magnitude as those in an HPI lamp.

It should be noted, incidentally, that tungsten can be used as an emitter in both these halide lamps because the presence of the halogen ensures that the tungsten evaporated and sputtered from the electrode is not deposited on the wall.

Continued research on electrodes for discharge lamps is expected to lead to a further reduction of the losses in SOX lamps, as opposed to the fluorescent lamps, where the limits for the improvement of losses and life appear to have been reached. In the case of the high-pressure lamps there are good prospects for the work on the improvement of the emitters and the reduction of losses.

Summary. The operation and properties of electrodes in discharge lamps are described, with particular reference to their influence on lamp life, energy losses and radio noise. Electrode design and choice of emitter must be adapted to the conditions so that the electrode is sufficiently resistant to ion bombardment during the cathode phase. The causes of radio noise are briefly discussed. The article concludes with a description of some types of electrode used in low-pressure lamps (fluorescent, SOX) and in high-pressure lamps (HP, HPI).

[6] H. J. J. van Boort, Philips tech. Rev. **25**, 114, 1963/64.



Salisbury Cathedral, England, illuminated by 23 metal-halide floodlights.

Optimum spectra for light sources

J. J. Opstelten, D. Radielović and J. M. P. J. Verstegen

Introduction

The various types of lamp discussed in the preceding articles illustrate the great variety that exists in lamps and their applications. Two of the characteristics of a lamp that determine its range of application are of general significance; the *luminous efficacy* η and the *colour rendering*, i.e. the extent to which the lamp renders the colours of objects naturally. The quality of the colour rendering is often expressed numerically by the *general colour-rendering index* R_a. We shall return to the definition of this quantity later.

Table I gives η and R_a for a number of well known lamps. The combinations of η and R_a found in Table I are given in the form of a diagram in *fig. 1*. It can be seen from this that, generally speaking, a higher luminous efficacy is often associated with a poorer colour rendering, and vice versa. The extremes are the incandescent lamp (1), which gives perfect colour rendering but a luminous efficacy of only 15 lm/W, and the low-pressure sodium lamp (13), which makes everything yellow, brown or black, but has a luminous efficacy of no less than 183 lm/W.

In view of the general aim of combining good colour rendering with high luminous efficacy, it would be interesting to know where the border-line of the theoretically possible lies in fig. 1. Knowledge of this borderline would enable us to see whether or not there is any point in looking for better light sources.

Closer consideration suggests that this question should be put somewhat differently. The luminous efficacy η is the product of three factors:

$$\eta = K\eta_e = K_m V_s \eta_e.$$

The factor η_e is the fraction of the electrical energy supplied that is converted into radiant energy in the visible region of the spectrum (380-780 nm). For the various types of lamp this factor depends on a wide variety of technical factors, and need not be considered here. Given a particular radiant flux of power in the visible region, however, what we are then concerned with is the *spectral distribution* of that power. Its concentration into one line at the wavelength to which the eye is most sensitive (555 nm) would give the highest luminous efficacy for the same value of η_e. In this case the luminous efficiency of the visible radiation V_s (we shall return presently to the definition of this quantity) is equal to 1, and the 'luminous efficacy of the radiation' $K = K_m V_s$ has its maximum value K_m at about 680 lm/W. There is then, however, no colour rendering. This would require the radiant power to be distributed more uniformly over the visible region, but that automatically means that V_s becomes less than 1, because some of the light is then emitted at wavelengths to which the eye is less sensitive. Table I also gives the value of V_s for the various light sources, and *fig. 2*

Drs J. J. Opstelten, Dipl.-Ing. D. Radielović and Dr J. M. P. J. Verstegen are with the Philips Lighting Division, Eindhoven.

Table I. Luminous efficacy η, luminous efficiency of visible radiation V_s, and general colour-rendering index R_a for a number of types of lamp. The chromaticities of these lamps are given in fig. 4.

	Lamp	Code	η(lm/W)	V_s(%)	R_a
1.	Incandescent lamp		15	32	100
	Fluorescent lamps				
2.	standard daylight	TL 54	66	42	77
3.	standard white	TL 33	80	48	65
4.	standard warm white	TL 29	78	50	52
5.	special daylight	TL 57	45	37	93
6.	special white	TL 37	45	37	93
7.	special warm white	TL 27	44	37	93
8.	High-pressure mercury fluorescent lamp	HPLN	54	50	41
9.	Metal-iodide lamp	HPI	80	50	67
10.	Dysprosium lamp		84	37	89
11.	Tin-halide lamp		60	38	91
12.	High-pressure sodium lamp	SON	120	60	23
13.	Low-pressure sodium lamp	SOX	183	76	—

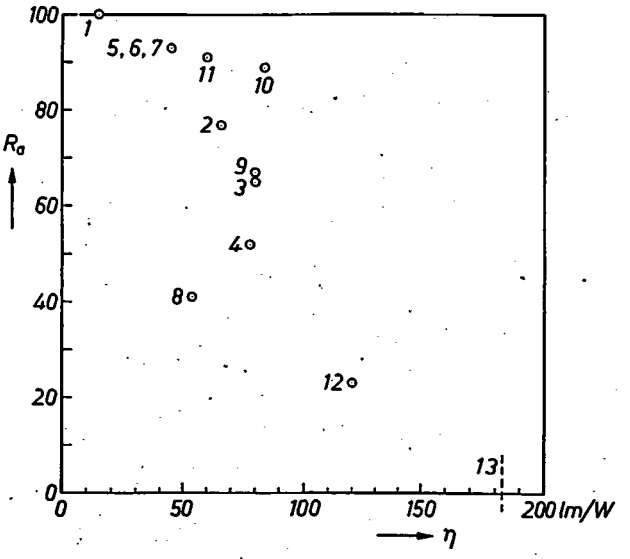


Fig. 1. Combinations of the general colour-rendering index R_a and luminous efficacy η in the lamps of Table I.

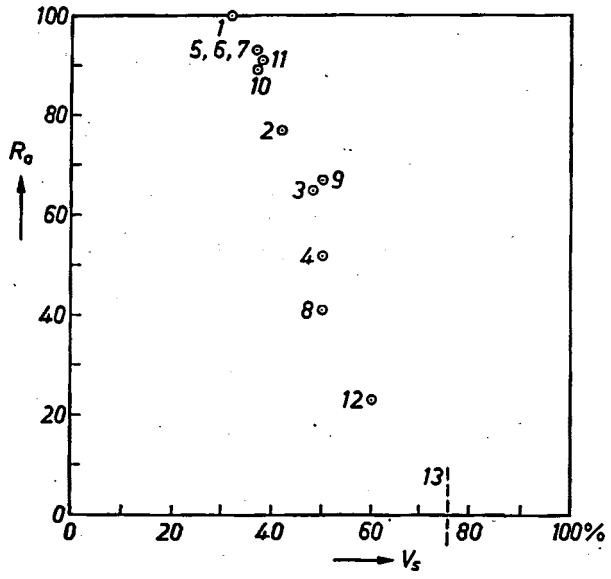


Fig. 2. Combinations of R_a, the general colour-rendering index, and V_s, the luminous efficiency of the visible radiation, in the lamps of Table I.

shows the R_a, V_s combinations in the form of a diagram.

Instead of looking for the 'theoretically possible' in fig. 1, we shall concern ourselves with the 'theoretically possible' in fig. 2. This amounts to calculating 'optimum spectra', that is to say spectral distributions of a given radiant power over the visible region such that R_a has a maximum at a given V_s, or vice versa.

To define the concept of 'colour rendering' more precisely we need to introduce another parameter. The 'natural colour' of an object is the colour it has in daylight. Now of course the spectral composition of daylight changes from day to day and from hour to hour, and therefore so does the spectral composition of the light reflected by objects. Nevertheless the visual impression remains the same under these changing conditions: we say that 'the objects keep their colour'. This is referred to as the 'law of colour constancy', based on the chromatic adaptation of the eye. The same applies to the light from incandescent lamps: in a room illuminated by incandescent lamps objects retain in general terms their natural colour. In spite of the difference in spectral composition, the various qualities of daylight and incandescent-lamp light all possess a 'perfect' colour rendering by virtue of the chromatic adaptability of the eye. In particular, the colour of the light itself is always perceived as 'white'. This by no means applies to all kinds of light; certainly not, for example, to the light from sodium, mercury or neon lamps. Physically, daylight and the light from incandescent lamps are distinguished from such other qualities of light by the fact that they approximate closely to 'Planckian radiation' — the radiation emitted by a black body at a given temperature. Planckian radiators themselves are differentiated in terms of their (*correlated*) *colour temperature*, the temperature of the corresponding black body. Now Planckian radiators of different colour temperature should not be used *together* as light sources: an incandescent lamp in daylight gives a yellow sensation, and conversely a daylight radiator in a space lit by incandescent lamps has a bluish colour. Consequently, when it is a question of light sources for general illumination — white light sources with good colour rendering — *it is always necessary to specify the colour temperature.*

In what follows we shall first give a more precise definition of the quantities V_s and R_a, in connection with some other basic concepts (luminous flux, chromaticity coordinates). We shall then discuss our calculations of optimum spectra and the results obtained. Finally we shall show how the knowledge thus obtained has led to a new generation of fluorescent lamps with an unprecedented combination of luminous efficacy and colour rendering.

Luminous flux, chromaticity coordinates and colour temperature

Light can be defined by three parameters, in the sense that the eye is unable to distinguish between two types of light with equal values of the corresponding parameters. The parameters chosen are usually the luminous flux and the two chromaticity coordinates in the chromaticity diagram or colour triangle (*fig. 3*). Since the calculation of these quantities from a given spectral distribution of radiant power in the visible region $P(\lambda)$ is an essential part of our investigation, we shall first briefly summarize the procedure [1].

In the trichromatic system laid down in 1931 by the International Commission on Illumination (Commission Internationale d'Eclairage, CIE), a number of functions of the wavelength λ were defined that characterize the sensitivity of the eye of the 'standard colorimetric observer' to various types of light. These are the colour-matching functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$. Of these $\bar{y}(\lambda)$ is the *spectral luminous efficiency* (for photopic vision), usually expressed by the symbol $V(\lambda)$. For $\lambda = 555$ nm the factor $V(\lambda)$ has its maximum value, which is equal to 1.

If we multiply the power $P(\lambda)\mathrm{d}\lambda$ in the wavelength interval λ, $\lambda + \mathrm{d}\lambda$ by the function $K_m V(\lambda)$, we obtain the luminous flux in that interval. The total luminous flux Φ_v is therefore $\Phi_v = K_m \int V(\lambda)P(\lambda)\mathrm{d}\lambda$.

The ratio of the integral in this expression to the power of the visible radiation, $\int_v P(\lambda)\mathrm{d}\lambda$, is by definition the luminous efficiency of the visible radiation V_s:

$$V_s = \int V(\lambda)P(\lambda)\mathrm{d}\lambda / \int_v P(\lambda)\mathrm{d}\lambda.$$

Here $\int_v$ is an integral that extends only over the

visible region (380-780 nm). For monochromatic radiation of wavelength λ_0 the value of V_s is equal to $V(\lambda_0)$; in particular V_s is equal to 1 for monochromatic radiation of 555 nm.

maticities of the various qualities of daylight (clear sky, cloudy, overcast, etc.) and of incandescent-lamp light lie on or close to this locus. The colour temperature T_c increases from right to left along the locus. Points

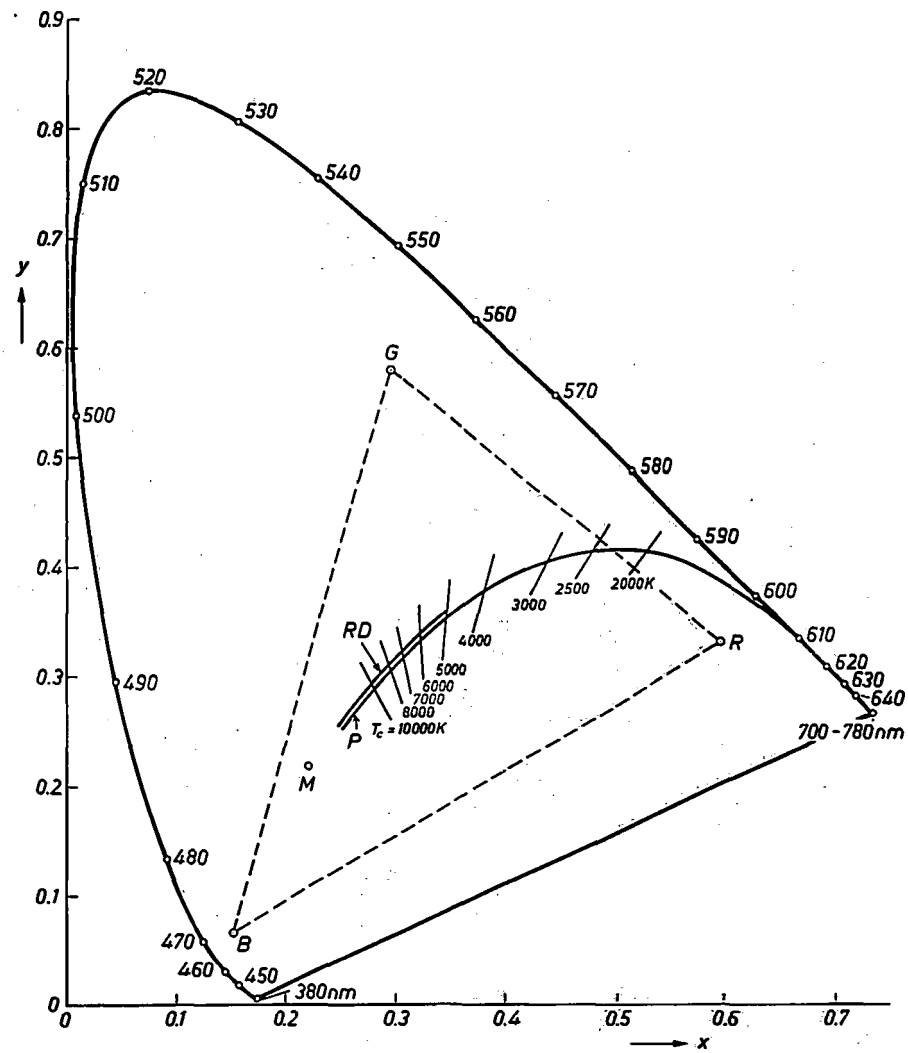


Fig. 3. The CIE chromaticity diagram (1931). The chromaticities of monochromatic light sources lie on the periphery. *P* is the Planckian locus, i.e. the curve along which the chromaticities of full (black-body) radiators lie. T_c colour temperature. *RD* chromaticity line of 'reconstituted-daylight radiators' as defined by the CIE. *M* chromaticity of the light from the low-pressure mercury discharge in a fluorescent lamp. *R*, *G*, *B* chromaticities of the phosphor emission together with the mercury light, for the three phosphors developed for the new generation of fluorescent lamps (see p. 369).

Multiplying $P(\lambda)$ by the functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ we obtain, after integration, the tristimulus values

$$X = \int \bar{x}(\lambda)P(\lambda)\mathrm{d}\lambda, \quad Y = \int \bar{y}(\lambda)P(\lambda)\mathrm{d}\lambda, \quad Z = \int \bar{z}(\lambda)P(\lambda)\mathrm{d}\lambda.$$

The chromaticity coordinates are then by definition:

$$x = X/(X + Y + Z), \quad y = Y/(X + Y + Z).$$

The line *P* in fig. 3 is the Planckian (black-body) locus, calculated from Planck's formula. The chromaticities near the locus are also assigned a colour temperature, which is that of the Planckian radiator whose colour best corresponds to the given chromaticity.

The chromaticities of the lamps in Table I are shown in *fig. 4*.

[1] More background information on these quantities can be found in A. A. Kruithof and J. L. Ouweltjes, Philips tech. Rev. 18, 249, 1956/57, and in the literature referred to in that article.

Colour rendering

The method of appraising the colour rendering of a light source consists in illuminating a number of representative 'test colours' (coloured surfaces) first with the light source and then with a reference illuminant of the same colour temperature. The colour rendering is a measure of the correspondence between the colour appearance under the light source investigated and the colour appearance under the reference source. dinates of the reflected light of both sources. Since colour differences judged to be equal are not represented as equal chromaticity distances in the x,y chromaticity diagram, we transform the x,y-coordinates into the u,v-coordinates of the UCS diagram (CIE 1960; UCS stands for Uniform Chromaticity Scale), in which the corresponding chromaticity distances are more nearly equal. From the distance between the two chromaticities found in this diagram — possibly cor-

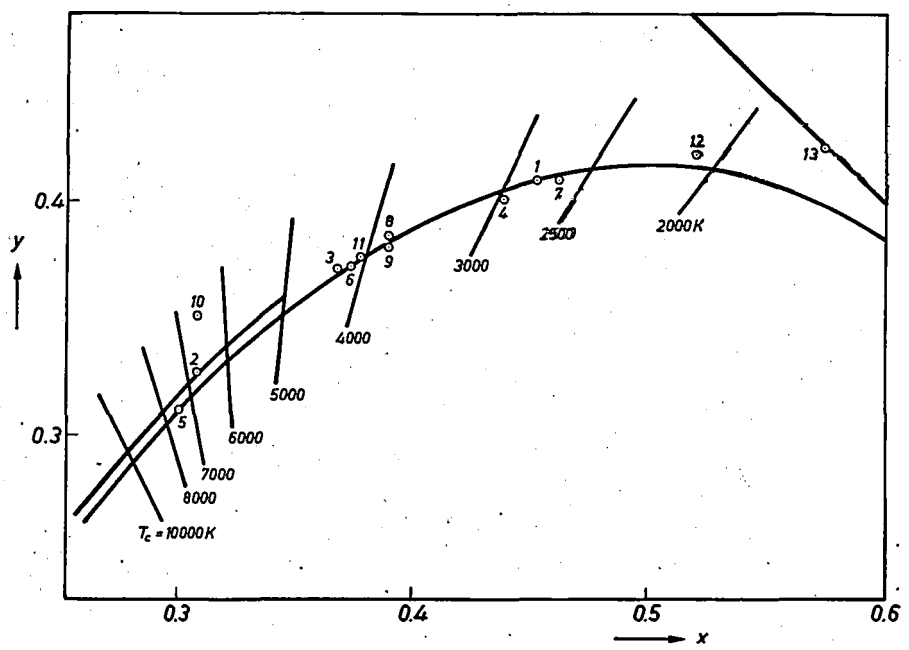


Fig. 4. Part of the chromaticity diagram, showing the Planckian locus and that of the RD radiators, and the chromaticities of the lamps in Table I.

Reference illuminants are defined by the CIE in terms of their spectra, and eight test colours are defined by their spectral radiance factors $\beta_i(\lambda)$. The reference illuminants recommended by the CIE [2] are Planckian radiators in so far as they relate to colour temperatures up to 5000 K. At temperatures of 5000 K and higher the recommended sources are 'reconstituted-daylight radiators' (RD radiators), defined by particular spectra [3]. Their chromaticity coordinates differ slightly from those of the Planckian radiators (see fig. 3). Examples of spectral radiance factors are shown in *fig. 5*, relating to the test colours Nos. 1 (red), 3 (green-yellow) and 6 (purplish blue).

If we know the spectral energy distribution of a (possibly theoretical) light source, we can now calculate the colour rendering in the following way [4]. From the spectral energy distributions of the lamp under investigation and the reference illuminant, together with the spectral radiance factor $\beta_1(\lambda)$ of test colour No. 1, we calculate the chromaticity coor- rected to allow, for example, for a difference in the state of adaptation of the eye resulting from a difference in chromaticity of the two light sources — we derive the 'special colour-rendering index' R_1. This is defined such that R_1 is equal to 100 at a corrected chromaticity distance of zero, and is smaller the greater the corrected chromaticity distance. The same procedure is repeated for the seven other test colours. The *general colour-rendering index* R_a is the average of R_1 to R_8. With ideal colour rendering R_a is thus equal to 100. For a fluorescent lamp of the 'standard warm white' type (TL 40 W/29) the value of R_a is 50.

In our calculations we used R_a as the criterion for defining the 'optimum' of a spectrum. In this context the restricted significance of R_a should be borne in mind. Although a light source with $R_a = 100$ renders the eight test colours in exactly the same way as the reference illuminant, if R_a is less than 100 the values of R_i may differ amongst themselves, and the rendering of one test colour could for example be much

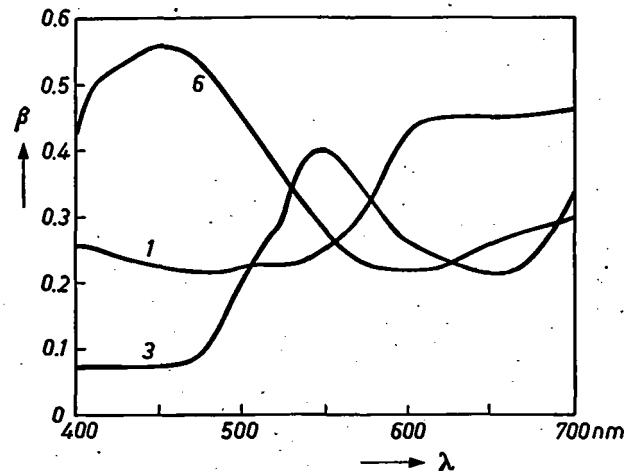


Fig. 5. Spectral radiance factor β as a function of wavelength for the CIE test colours Nos. 1 (red), 3 (green-yellow) and 6 (purplish blue).

poorer than that of another. A difference of five units in R_i is found in practice to be about the limit of the perceptible. A light source with $R_a = 95$ will therefore *in general* still give 'very good' colour rendering, but even at this value it is theoretically possible for one R_i to amount to only 60. The more R_a differs from 100 the less its significance is. Moreover, it may happen that the eight test colours chosen are well rendered but a ninth colour is not. To obtain a better appraisal of a light source the CIE has introduced a further six special test colours, consisting of saturated hues of red, yellow, green and blue, a skin colour and leaf green (Nos. 9 to 14). The spectral radiance factors of Nos. 9, 12 and 13 are given in *fig. 6*.

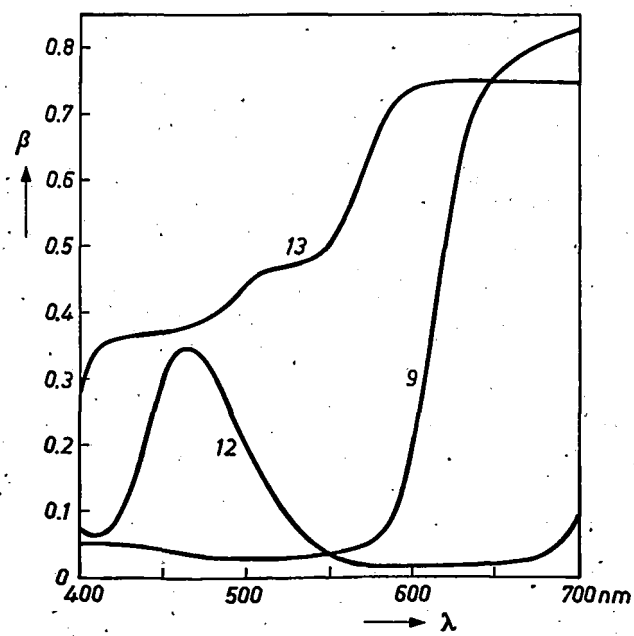


Fig. 6. Spectral radiance factor $\beta(\lambda)$ of the test colours Nos. 9 (saturated red), 12 (saturated blue) and 13 (skin colour).

Optimum spectra

In the preceding discussion we have already encountered one optimum spectrum. This is the spectrum obtained when there are no specifications for colour or colour rendering but V_s has to have the maximum value (100%): a single line at 555 nm.

One of the first requirements to be met by a lamp for general lighting is that the colour of the light itself should be white, which implies that the chromaticity coordinates should be those of the Planckian or RD radiator at the chosen colour temperature. Calculations have shown that the optimum spectrum for this single specification consists of two lines [5] in a particular intensity ratio: one blue line at 445 nm together with the colour-temperature-dependent complementary line (*Table II*). With increasing colour temperature the luminous efficiency attainable decreases.

Table II. Maximum attainable value of V_s for light sources whose chromaticities correspond to those of a particular Planckian or RD radiator (and are thus 'white'). The corresponding spectra consist of two lines, with the indicated wavelengths λ_1, λ_2.

Chromaticity	V_s(%)	λ_1 (nm)	λ_2 (nm)
P 2500 K	75.9	445	582
P 3000 K	74.5	445	578
P 4000 K	70.0	445	574
P 5000 K	65.6	445	571
RD 6000 K	63.1	445	569
RD 8000 K	58.1	445	566

The colour rendering of a two-line source is of course very limited, since the reflected light consists solely of the two given components — only the intensity ratio is changed — and therefore the chromaticy can only lie on the line joining the chromaticities of the two given spectral lines. The next step is obvious: with *three lines* a whole region of the chromaticity diagram can be covered. This principle is used in colour television.

In our calculations on optimum three-line spectra we proceeded as follows [6]. For a given power, a three-line spectrum has five degrees of freedom: the three wavelengths λ_1, λ_2, λ_3 and two intensity ratios, for example I_1/I_3 and I_2/I_3. At prescribed chromaticity coordinates (x, y) this is reduced to three. For each combination $(\lambda_1, \lambda_2, \lambda_3)$ we can therefore calculate the intensity ratios along with V_s and R_a. We varied λ_1, λ_2

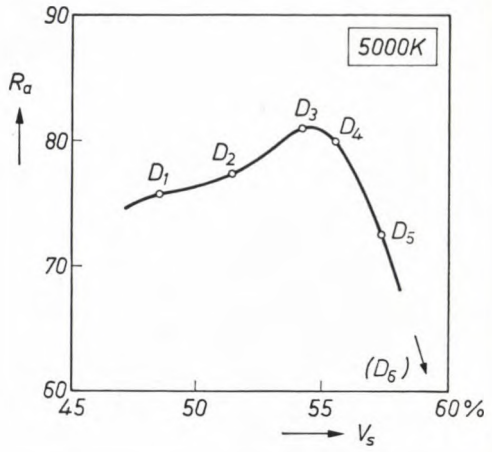
[2] CIE publication No. 13.2 (T.C.-3.2), 1974.
[3] CIE publication No. 15 (E-1.3.1), 1971.
[4] An accurate definition of the colour-rendering indices is given in the CIE publication referred to in note [2].
[5] D. L. MacAdam, J. Opt. Soc. Amer. 40, 120, 1950.
[6] M. Koedam and J. J. Opstelten, Lighting Res. Technol. 3, 205, 1971.
W. A. Thornton, J. Opt. Soc. Amer. 61, 1155, 1971.

and λ_3 in steps of 1 nm until a maximum was found for R_a. Next we selected from the neighbouring spectra with a given 'neighbouring' value of V_s the 'optimum' spectrum: the one for which R_a was largest. We found that in general a simultaneous shift of the three lines in one direction had little effect on R_a, whereas V_s was increased when the three lines were shifted in the direction of shorter wavelengths. Proceeding in this way from point to point we found the maximum value of R_a and the corresponding spectrum as a function of V_s.

We carried out this procedure for the chromaticities of the CIE reference sources P 2500 K, P 3000 K, P 4000 K, P 5000 K, RD 6000 K and RD 8000 K. The graph in *fig. 7 (above)* gives the R_a of the optimum spectrum as a function of the chosen V_s for the chromaticity of P 5000 K. The maximum of R_a is seen to be no less than 80. The value of R_a decreases rapidly, however, when V_s exceeds the value corresponding to the maximum. The lower diagrams in fig. 7 show how the three wavelengths of the optimum spectrum depend on V_s. At the maximum V_s attainable in this way for this particular chromaticity, $V_s = 65.6\%$, the spectrum is reduced to the two-line spectrum of Table II (spectrum D_6). The diagrams in fig. 7 also show in which range R_9 to R_{14} are situated.

Fig. 8 (above) shows the R_a-V_s curves for the six cases mentioned, and the

Fig. 7. *Right:* R_a as a function of V_s for the optimum three-line spectra at the chromaticity of the CIE reference source P 5000 K. *Below:* the optimum three-line spectra and the values of R_a, V_s and R_1 to R_{14} corresponding to points D_1, D_2, ... D_6 on the curve. For R_9 to R_{14} the diagram only indicates the region in which they lie. At the maximum V_s the spectrum (D_6) is reduced to two lines (see Table II).



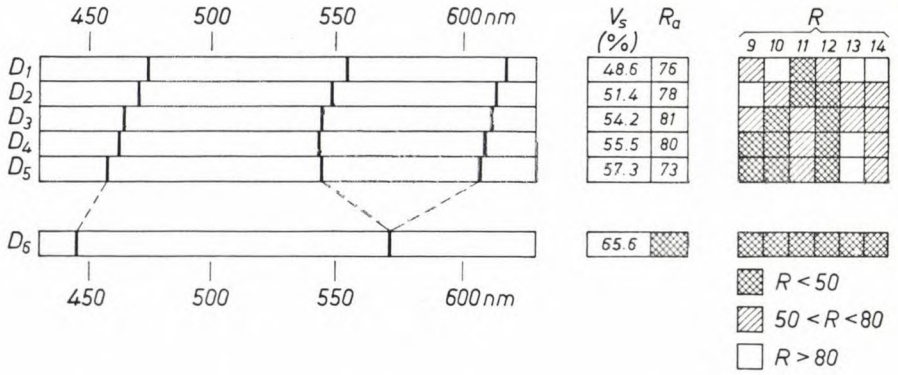
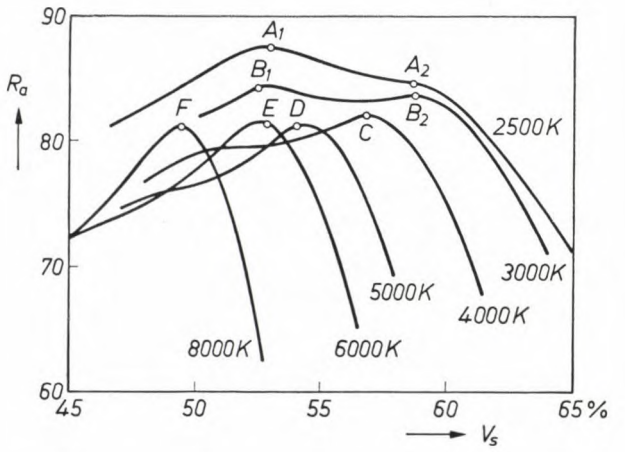
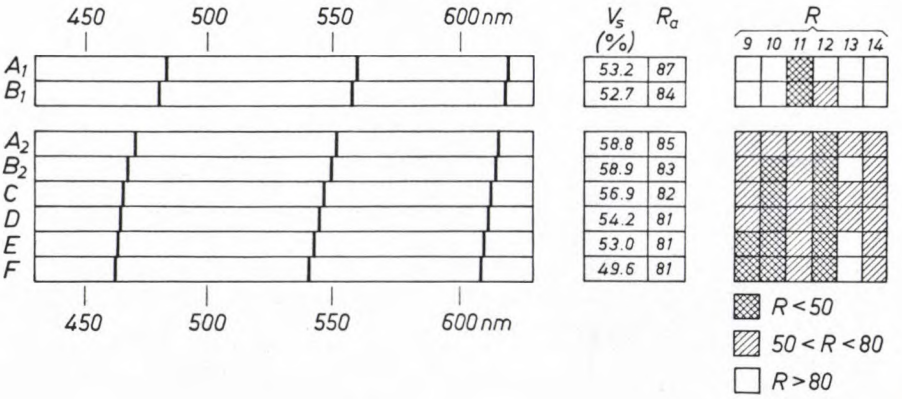


Fig. 8. *Right:* R_a-V_s curves for the optimum three-line spectra at the chromaticities of six different CIE reference sources. *Below:* Optimum spectra and values of V_s, R_a, R_9 to R_{14} corresponding to the maxima in the upper figure.





diagrams under the graph give the spectra and the R_9 to R_{14} regions that correspond to the maxima. The luminous efficiency corresponding to the maximum decreases as the colour temperature increases. The wavelengths required for the maximum shift somewhat towards smaller values with rising colour temperature.

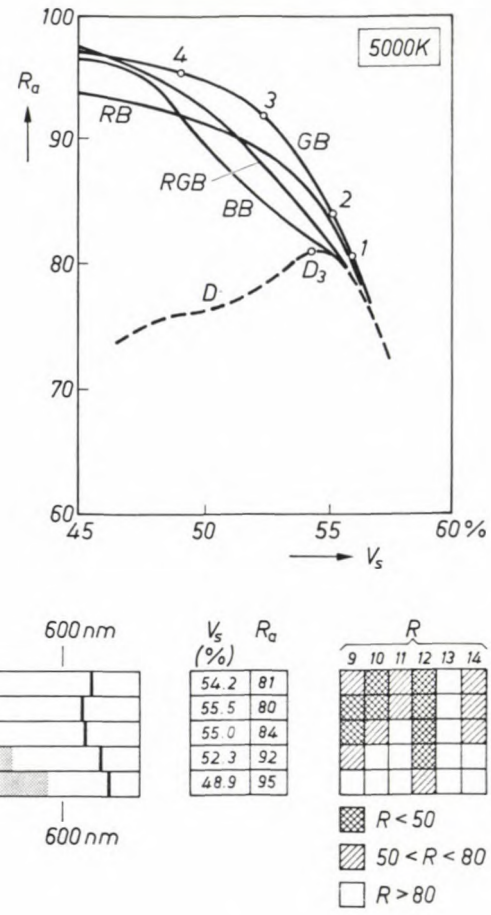
Summarizing the main results: with a three-line spectrum an R_a of more than 80 can be achieved, and the wavelengths required are about 465, 545 and 610 nm (lower diagrams in figs. 7 and 8).

Band spectra and spectra with more than three lines

The colour rendering can be considerably improved by broadening one or more lines to form bands [7]. For the energy distribution (on the wave-*number* scale) we adopted a gaussian curve. We now have one new variable: the bandwidth. By carrying out the procedure described above for one bandwidth at a time, we again found curves of the type shown in fig. 7 (above). Variation of the bandwidth yields a family of such curves. *Fig. 9 (above)* gives the envelope of families of curves for the chromaticity of P 5000 K, for the cases of a red, a green and a blue band, each combined with two lines, and for three bands of the same width. As can be seen, broadening of the green line gives by far the best results. Examples of some typical spectra of this kind are given in fig. 9 (*below*). More detailed calculations show that some further improvement on this situation can be obtained by broadening the blue line, but that broadening the red line causes a deterioration.

Finally, *fig. 10* illustrates what can theoretically be achieved with line splitting instead of line broadening [8]. The spectra themselves will not be dealt with here.

Fig. 9. *Right:* R_a-V_s curves for optimum spectra consisting of one band and two lines, and of three equally broad bands (chromaticity P 5000 K). *RB*, *GB*, *BB* denote a red band, a green band and a blue band respectively, each with two lines. *RGB* three bands of equal width. *D* bandwidth zero (three-line spectrum, see fig. 7). *Below:* optimum 'green-band spectra' and the V_s- and R-values corresponding to some of the points on the curve *GB*. The portion of the green band between the half-width wavelength is shown in grey. The spectrum D_3 from fig. 7 is included for comparison.



Taking our results as a whole, we come to the following conclusion. The 'limit of the possible' we were looking for in fig. 2 depends on the colour temperature, as appears for example in the case of three-line spectra in fig. 8; at a given R_a the attainable V_s increases as T_c decreases. If we compare fig. 10 and fig. 2 we see that,

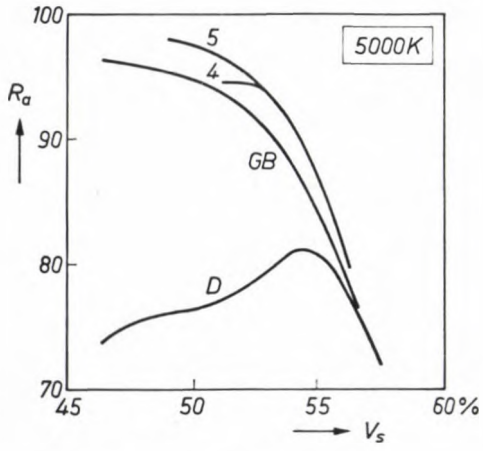


Fig. 10. R_a-V_s curves (*4* and *5*) for optimum spectra of four and five lines, obtained by line-splitting from a three-line spectrum. The curves *D* (three lines) and *GB* (green band plus two lines) from fig. 9 are included for comparison.

[7] M. Koedam, J. J. Opstelten and D. Radielović, J. Illum. Engng. Soc. **1**, 285, 1972.

[8] See [7], and also: K. Mahr, Extended Abstracts Spring Meeting Electrochem. Soc. 72-1 (Houston 1972), abstract No. 76.

theoretically, much better combinations of V_s and R_a are possible with three, four and five-line spectra than have so far been achieved. The transition from three to four lines means a considerable gain. The transition from four to five lines is so much less impressive that there is not much practical advantage to be gained from further refinements: there is little need to make R_a greater than 95. In other words, for all practical purposes the theoretically possible is given in fig. 10.

Until recently it was always assumed that it was necessary to simulate the continuous spectrum of Planckian radiators to obtain good colour rendering. The significance of our results is that they show that good colour rendering can also be achieved with a line spectrum, which will give greater efficacy. As we mentioned earlier, we took R_a as the criterion for colour rendering in our calculations. The good results found are partly due to the fact that the spectral radiance factors of the test colours, like those of most colours of practical interest, are relatively broad, smooth curves. In the rare cases where the spectral radiance factors have peaked curves it is possible that unexpected and undesirable results might arise.

In the rest of this article we shall discuss, as an example, the application of the above principle to the fluorescent lamp.

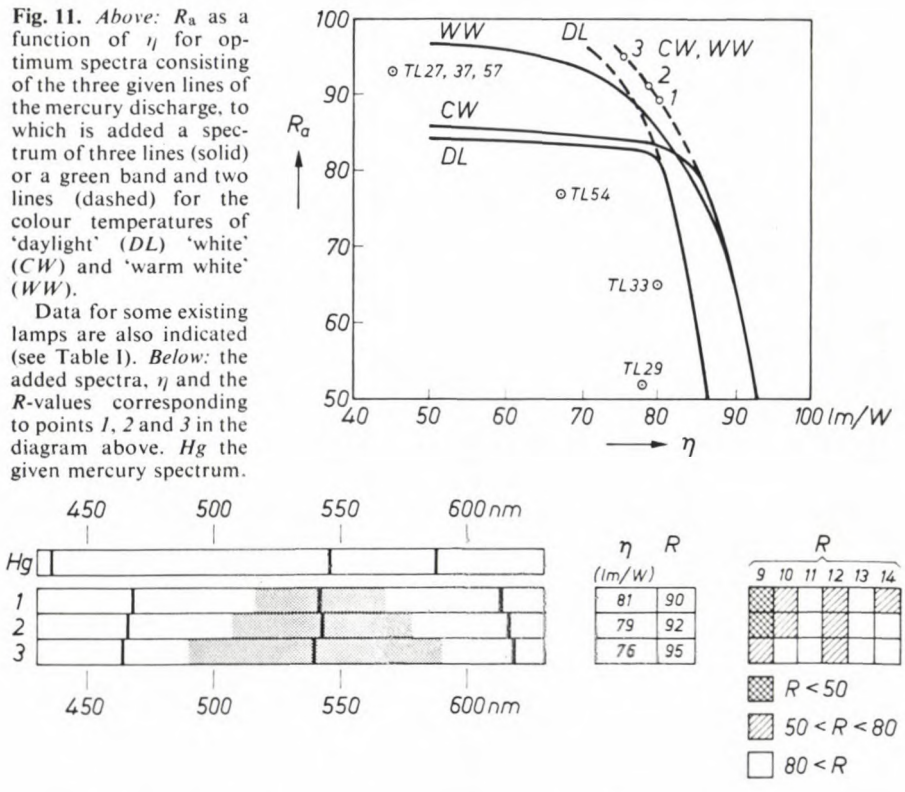


Fig. 11. *Above:* R_a as a function of η for optimum spectra consisting of the three given lines of the mercury discharge, to which is added a spectrum of three lines (solid) or a green band and two lines (dashed) for the colour temperatures of 'daylight' (*DL*) 'white' (*CW*) and 'warm white' (*WW*).

Data for some existing lamps are also indicated (see Table I). *Below:* the added spectra, η and the *R*-values corresponding to points *1*, *2* and *3* in the diagram above. *Hg* the given mercury spectrum.

Optimum spectra for fluorescent lamps

Our calculations for fluorescent lamps [9] differ in two respects from those discussed above. In the first place we considered the visible lines of the mercury discharge in the tube — a relatively strong blue line at 436 nm, a green one at 546 nm and a yellow one at 578 nm — as a given part of the ultimate spectrum. This means that we calculated the lines or bands that have to be *added* to obtain an optimum result. Secondly, in the case of a low-pressure mercury discharge lamp we know how much of the supplied electrical power is converted into UV radiation and how much is converted into visible radiation. This allows calculation with the luminous efficacy η, which is of more direct significance for practical purposes than the luminous efficiency of the visible radiation, V_s. We did however have to make an assumption about the fraction of the number of available UV quanta that are converted by the phosphors into visible quanta; we put this 'quantum efficiency' Q at 0.8, a not uncommon value with good phosphors.

In agreement with what has been said above, our results indicate that the mercury spectrum can be improved quite appreciably by the addition of three lines. For the colour temperature of 'warm white' a value of more than 90 can be reached in this way for R_a. Compared with the pure three-line spectrum the blue line is shifted towards a longer wavelength, as a result of the strong blue mercury line. If band broadening is also permitted, a green band with a red and a blue line again proved to be the most effective. The R_a-η curves for these cases, for the colour temperature of 'daylight' 'white' and 'warm white', and three of the corresponding optimum spectra are given in *fig. 11*.

Our results, as presented in fig. 11 (*above*), show clearly that there is room theoretically for an improvement in fluorescent lamps, and at the same time they indicate the limits of the attainable. Results like those shown in fig. 11 (*below*) give an important indication of the direction in which we should look for the desired improvement. Briefly, this amounts to supplementing the mercury spectrum with emission in the spectral regions around 465 nm (blue), 545 (green) and 615 nm (red), with the green emission in the form of a relatively

broad band, the blue in a narrow band and the red as sharp a line as possible.

The new generation of fluorescent lamps

Using the results just discussed as a guideline, we looked for phosphors with which better lamps might be produced. The phosphors commonly used in fluorescent lamps were ruled out because of their broad-band emission. The narrow-band emission required pointed in the direction of phosphors activated with rare-earth ions. Many such phosphors have been discovered and applied in the past ten years in illumination engineering and in colour television [10].

Taking into account the specified spectral regions around 465, 545 and 615 nm, we find the most promising activators to be, respectively, Eu^{2+}, Tb^{3+}, and Eu^{3+} [9]. *Fig. 12* gives the energy-level diagrams for these ions in an oxidic host lattice.

The red emission of Eu^{3+} is due to the transition between two 4f levels. The electrons are situated, and remain situated, in the deep 4f shell in the ion: as a result the levels are sharply defined and not much affected by the host lattice. The emission is consequently in a very narrow band, which is what we require. The emission of Eu^{2+} involves a transition from a $4f^65d$ state, where there is an electron in the outer shell, to the ground state $4f^7$. The spectral emission is therefore somewhat broader than in the case of Eu^{3+}, and moreover the wavelength is highly dependent on the nature of the host lattice. Although the emission from Tb^{3+} takes place between sharply defined $4f^8$ levels, the transitions here are in fact very close together, so that for practical purposes the emission may be regarded as a single relatively broad band. The wavelength, however, is not much affected by the host lattice.

Having chosen the activator ions, we had to look for suitable host lattices for them. The main thing was to obtain a sufficiently stable phosphor, that is to say one that could withstand the various processes during lamp manufacture as well as the intensive UV radiation during operation. In addition, the phosphor had to have a sufficiently high quantum efficiency. A good phosphor containing the Eu^{3+} ion was already known and in use for colour television and other applications; this is $Y_2O_3:Eu^{3+}$, which emits in a sharp line at 611 nm. In view of the stability of Al_2O_3-type compounds we looked among the aluminates [11] to find host lattices for Eu^{2+} and Tb^{3+}. By modifying the already known barium aluminate using Eu^{2+} as activator [12], we obtained a phosphor that emits at 450 nm with a half-width of about 50 nm, and meets the requirements for stability and high efficiency [13]. These requirements

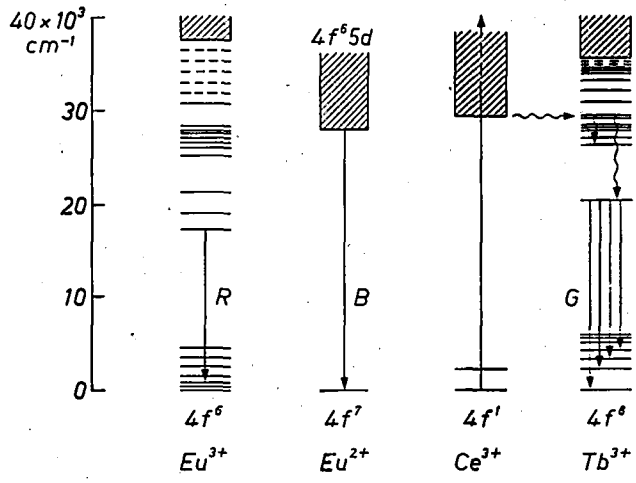


Fig. 12. Energy-level diagrams of the rare-earth ions used as activators in the phosphors for the new fluorescent lamps. The red emission of Eu^{3+} is a sharp line, the blue emission of Eu^{2+} is less sharp because a 5d state is involved. The green emission of Tb^{3+} consists of a number of sharp lines, and in practice corresponds to a band 80 nm broad. Ce^{3+} is used as a sensitizer in the newly developed green phosphor.

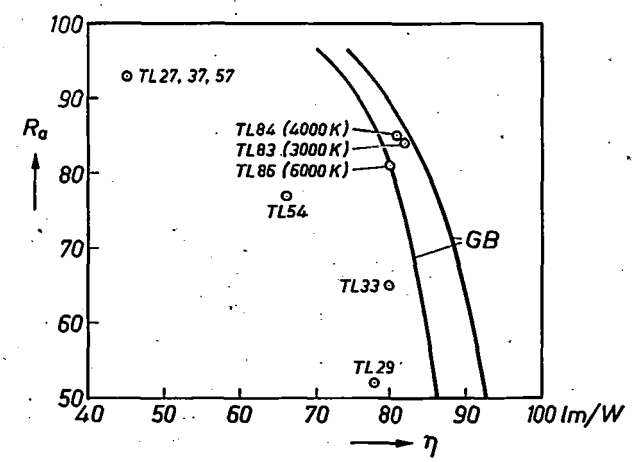


Fig. 13. The newly developed fluorescent lamps TL 83, TL 84 and TL 86, with the colour temperatures 3000, 4000 and 6000 K, together with some currently used lamps in the R_a-η diagram. *GB* curves indicating what is theoretically feasible with a green band and two lines (see fig. 11).

are also fulfilled by the cerium-magnesium aluminate with Tb^{3+}, which we have developed. This emits at 545 nm with an effective half-width of about 80 nm. As in other Tb phosphors with cerium [10], the cerium acts here as a 'sensitizer': the energy is absorbed in the form of UV quanta by the Ce^{3+}, transferred to the

[9] J. J. Opstelten, D. Radelović and W. L. Wanmaker, J. Electrochem. Soc. 120, 1400, 1973.
[10] See for example A. Bril and W. L. Wanmaker, Philips tech. Rev. 27, 22, 1966 and G. Blasse and A. Bril, Philips tech. Rev. 31, 304, 1970.
[11] J. M. P. J. Verstegen, J. Electrochem. Soc. 121, 1623, 1974.
[12] G. Blasse and A. Bril, Philips Res. Repts. 23, 201, 1968.
[13] J. M. P. J. Verstegen, D. Radelović and L. E. Vrenken, J. Electrochem. Soc. 121, 1627, 1974.

Tb^{3+}, and finally emitted by this ion as 'green' quanta (see fig. 12).

Fig. 3 gives the chromaticities of the emission from the three phosphors mentioned above, in each case together with the emission from the mercury (R, G, B). Any point inside the triangle RGB can in principle be realized by appropriate mixing, and in particular any colour temperature above 2300 K can be obtained.

In this way we have made fluorescent lamps with colour temperatures of about 2600, 3000, 4000 and 6000 K. It is clear from *fig. 13* that this new generation marks a significant advance. The figure shows the R_a-η combinations of the new lamps together with those of existing types. It can be seen that the new lamps surpass the 'standard' lamps in luminous efficacy, and at the same time are not much inferior to the 'special' type in colour rendering. The points lie more or less on the GB curves, which means that what is theoretically feasible with a green band and two lines by means of a phosphor with a quantum efficiency of 80% (fig. 11) has virtually been achieved.

Summary. From the spectral power distribution of the visible radiation the 'luminous efficiency of the visible radiation' V_s of a light source can be calculated, and also, in accordance with the recommended procedure of the Commission Internationale d'Eclairage, the 'general colour-rendering index' R_a. The quantity R_a is significant only for approximately 'white' light sources, whose colour temperature T_c always has to be specified. For ideal colour rendering R_a has a value of 100. 'Optimum spectra' are calculated, i.e. distributions in which V_s has a maximum for a given R_a, or vice versa. With three-line spectra the maximum R_a is more than 80; the corresponding V_s is 50% at $T_c = 8000$ K, and 57% at $T_c = 4000$ K. Better results can be achieved with band spectra or spectra consisting of more than three lines. Substantial improvements of fluorescent lamps are shown to be theoretically possible by the addition of a green band at about 545 nm, a red line at about 615 nm and a blue line at about 465 nm to the spectrum of the mercury discharge in the tube. This approach has been used in developing new fluorescent lamps whose colour rendering is nearly as good as that of 'special' lamps and which at the same time have a higher luminous efficacy than 'standard' lamps.

A street in Lyons (La part Dieu), illuminated at the centre by high-pressure sodium lamps (400 W) and along the pavements by high-pressure mercury-vapour lamps with fluorescent envelopes (125 W).

Volume 35, 1975, No. 11/12 pages 293-372 Published 29th April 1976

Subject index, Volumes 26-35

Figures in bold type indicate the volume number, those in ordinary type the page number. Subjects dealt with in volumes **1-25** are listed in the indexes included with volumes **10, 15, 20, 25** and **30**. The asterisk * indicates that the entry refers to a photograph and caption.

Aachen and Hamburg, Philips laboratories at . . . . . **26**, 89 -176
Accelerators for elementary particles, see **Proton synchrotrons**
Acoustics:
filters with acoustic surface waves **32**,179
magnetoacoustic effects in bismuth **32**,233
acoustic waves in piezoelectric semiconductors **33**,310
acoustoelectric effect **33**,310
see also **Audio**
Activators, see **Luminescence**
Adsorption on Si and Ge **32**,131
Agriculture, history of pests and disease control in **28**,121
Agrobiological research:
availability of organic compounds in higher plants . . . **28**, 93
phenol synthesis and photomorphogenesis **28**,101
Boekesteyn laboratory* . . . **28**,111
atomization of dispersions of pesticides **28**,112
Airfields:
computer for air-traffic control **26**,226
short-range radar for ground services **32**, 13
MADGE guidance system for aircraft landing . . . **34**,225 **35**,271
Air pollution, see **Environmental science**
Air traffic, see **Airfields**
Alkali metals, surface properties . . **30**,225
Alloys:
alloying behaviour of transition metals **33**,149
influence of ionicity **33**,196
ferromagnetic, electrical conduction **35**, 29
Alloy-diffusion technology (p.o.b. transistors) **28**,246
Alpha-particle spectrometry with semiconductor detectors **30**, 13
Amplifier:
small-signal, MOS transistor as **31**,216
integrated audio, with high input impedance and low noise **31**,245
power, for h.f. band, MOS transistor as **31**,251
integrated amplifier circuits . . **32**, 8
see also **Masers** and **Parametric amplifiers**
'Amplimat' (automatic exposure for X-ray diagnosis) **27**, 92
Analog computer for simulating one-dimensional antenna arrays . **31**, 2
Analog signals, shift register for ('bucket-brigade' delay line) . . **31**, 97 **31**,266
Analytical chemistry, see **Inorganic chemical analysis**
Anisotropy of electrical resistance **35**, 29
Anomalous oscillations with IMPATT diode **32**,361
Anomalous transmission of X-rays **29**,114
ANS, the Netherlands astronomical satellite:
general survey of project . . . **33**,116
attitude control **33**,162
onboard computer **34**, 1
UV experiment (Groningen University) **34**, 33
X-ray experiment (Utrecht University) **34**, 43
reaction wheels **34**,106
optical sensors:
sun sensors **34**,208
horizon sensor **34**,213
star sensor **34**,218
Antenna systems:
for satellite-launcher telemetry* **27**,131
simulation of one-dimensional electronically scanned, with *P-I-N* diodes **31**, 2 **32**,405
APT programming language . . . **28**,329
Argon-plasma torch, r.f., for emission spectroscopy **33**, 50
Argonne proton synchrotron . . . **30**,330
Astronomy, see **ANS** and **Radio astronomy**
Asynchronous motors, see **Induction motors**
Audio:
stereophonic radio broadcasting **26**,327 **27**, 62
measurements on hearing aids* **27**, 46
hearing aids (behind-the-ear types) **27**,258
motional feedback with loudspeakers **29**,148
audio tape cassettes **31**, 77
integrated audio amplifier . . **31**,245
see also **Acoustics**
Auger electron spectroscopy . . . **34**,359
Automation, see **Mechanization** and **Numerical control**
Avalanche photodiode **35**, 23
AVOID short-range high-definition radar **32**, 13

Ballast for long fluorescent lamps **31**, 57
Barium titanate:
PTC thermistors of **26**,119 **30**,170
ceramic, microstructure of . . **32**, 92
Battery of 'paper' **28**,298
Beam-plasma amplifier tubes . . . **27**,275
Bearings:
spiral-groove, grease-lubricated, for straight-through shaft **27**,107
lubrication with gases **29**,186
spiral-groove, in flywheel for space vehicles **30**, 2
hydrostatic, in precision lathe **30**,117
spiral-groove, grease-lubricated **34**,103 **35**,137
'push-pull' spiral-groove bearing **35**, 11
Binary systems in phase theory . **26**, 12 **26**, 52
Bismuth, magnetoacoustic effects in **32**,233
Bleustein-Gulyaev waves **33**,310
'BOL' nuclear research project* . . **30**, 35
Boron filament as constructional material **35**,125
Bowl cutter machine* **30**,306
Bragg reflections in electron photomicrographs* **26**, 48
Brittle materials, cutting* **30**,306
Brookhaven proton synchrotron . **30**,330
Brushless d.c. motors:
small, efficient **30**, 7
with integrated Hall device . . **31**,366
steady-state performance . . . **35**,106
applications:
ANS reaction wheels . . **34**,106
ANS horizon sensor . . . **34**,213
Bubble memory, iron-garnet . . . **35**, 1
Bucket-brigade delay line . . . **31**, 97 **31**,266
Burn-out of incandescent lamps **32**,155 **35**,296

Cadmium sulphide, monograin layers **29**,238
Camera tubes for television, see **'Plumbicon'**
Capacitors, electrolytic, non-polar **29**, 37
Capacitor motor, supply-voltage speed control **34**,180
Car-transmission control, electronic **28**,179
Carbon, glassy* **31**,369
Carbon-filament lamps with halogen filling **35**,316
Cardiograph, universal vector . . **29**, 42
Cardiovascular-examination installation* **29**,298
Carrier telephony: transistorization, conclave principle **26**,206
Cassettes, audio tape **31**, 77
Cathodes:
with small thermal delay (harp cathode) **26**,309
osmium dispenser cathodes . . **27**, 69
Cathodoluminescence for investigating crystal defects **35**,239
Ceramics technology:
continuous sintering under pressure **29**, 45
renaissance in (Prof. A. L. Stuijts's inaugural lecture) . . **31**, 44
influence of microstructure on properties of electroceramics **32**, 79
isostatic hot pressing without metallic encapsulating layer . **35**, 65
thick-film technology for hybrid-IC interconnection patterns **35**,144
sintering process **35**,188
ceramic-to-metal bonding . . **35**,209
CERN proton synchrotron **30**,330
Channel electron multiplier:
basic principles **28**,156
channel plates **30**,252
investigation of microchannel plates by scanning electron microscopy **34**,270
Characteristic luminescence, see **Luminescence**
Charge-coupled devices **31**, 97 **31**,266
Chemical analysis, see **Inorganic chemical analysis**

Chemiluminescence, detectors based on 34, 73
Chopper, integrated, with MOS transistors 31,248
Circuit blocks, digital, see **Digital circuit blocks**
CO_2 lasers, sealed-off, for high power 28,287
Coaxial and stripline microwave components 28,211
COBRA, see **Numerical control**
COD meter for water 34,123
Code modulation with digitally controlled companding 31,335
Coils:
  superconducting:
    in radio interferometer . . 27,313
    for magnet 29,309
  see also **Inductors**
Cold-cathode tube for flame-failure control 29,254
Colour rendering:
  of discharge lamps 35,347
  optimum spectra for light sources 35,361
Colour television:
  new phosphors for 27, 22
  transmission systems 27, 33
  camera with 'Plumbicon' tubes 28,336
  studio* 29,147
  delay line for PAL receivers . . 29,243
  light-weight camera with small 'Plumbicon' tubes 29,325
  degaussing colour-TV tube* . 29,368
  fast phosphors 32,125
  'VLP' video record, see separate entry
Commutator motors, see **D.C. motors**
Companding:
  companders with a high degree of compression 26,215
  digitally controlled, for code modulation 31,355
Compensation wall in ferromagnetic materials 34, 96
Complementary MOS circuits, LOCMOS technology 34, 19
Composites, growth of* 32,102
Compressor:
  application of rolling diaphragm seals 26,287
  two-stage, for helium liquefier 30, 51
Computers:
  recent developments 26, 67
  data checking by check characters 26,156
  analog, for simulating one-dimensional antenna arrays . 31, 2
  ANS onboard computer . . . 34, 1
  microprogramming for P1000 family of computers 34,132
  applications:
    air-traffic control 26,226
    monitoring nuclear reactor 30, 48
    calculation of cross-coupled waveguides 32,165
    calculation of Gunn effect 32,385
    automatic reading of hand-written numerals 33, 89 – 33,130
    analysis of emission spectra 34,322
    neutron activation analysis 34,330 – 34,351
    measurements from X-ray photographs 35,170
    testing digital circuits . . 35,261
    numerical control, see separate entry
    control of draughting machine, see **Drawing by computer**
  see also **Peripheral equipment for computers** and **Software**
Connection patterns:
  on flexible tape 34, 85
  in thick-film technology . . . 35,144
Constant-current source, integrated 32, 4
Continuous furnaces for rod material 27,337
Contour deposition 28,153
Control-rod monitoring for Dodewaard nuclear reactor 30, 48
Control techniques, see **Measurement and control**
'Controlled cascading' for frequency dividers 32,103
Core drilling with polygonal bit* . 30,307
Correlation measurement with stellar interferometer . . . . . 27,141 – 27,152
Counters, see **Radiation monitoring**
'Cryogem' miniature Stirling refrigerator 32, 51
Cryogenics:
  gas-refrigerating machine for 20 K and below 26, 1
  production of ultra-pure hydrogen 26, 27
  industrial refrigerating machine with hydraulic piston drive . 26,297
  Peltier cooling 27,113
  expansion ejector 28,243
  fast cryopump with two-stage gas-refrigerating machine for ultra-high vacuum 28,355
  heat exchanger, compact, of high thermal efficiency . . . 29,158
  helium liquefier with gas-refrigerating machines . . 29,197 – 30, 51
  cooling of parametric amplifier 32, 27
  miniature refrigerators for electronic devices 32, 49
Crystals:
  growth of, by chemical transport reactions 26,117
  crystal chemistry and magnetism of oxide materials (crystal field theory; spinel structures) . . . . . 28, 13 – 28, 23
  single, of ferrites and garnets, research on 28,135
  for studies in electronics* . . 28,145
  sawing of, for record-player cartridges* 28,297
  growth of, by temperature-alternating methods 28,316
  highly perfect, anomalous X-ray transmission in 29,114
  structure determination with PAILRED automatic single-crystal diffractometer . 29,165 – 29,176
  hydrothermal growth 30, 89
  single, various* 30,263
  vacancy clusters in dislocation-free Si and Ge 34,244
  transport of crystal defects in solids 35,181
  investigation of crystal defects by cathodoluminescence . . 35,239
  see also **Silicon carbide**
Cuprous-chloride single crystals . . 30,261
Current-limiting circuits for transistorized power supplies 28,251
Cutting brittle materials* 30,306
Cyclotron, development of (Prof. N. F. Verster's inaugural lecture) 26,321

Data processing, see:
  Computers
  Peripheral equipment for computers
  Software
Data transmission:
  with digital circuits 30, 71
  in MADGE aircraft-guidance system 35,271
D.C. motors:
  general 33,230
  with speed control 34,163
  small, designing 35, 96
  see also **Brushless d.c. motors**
Defford radio-astronomy interferometer 27,313
Deflection coils for TV picture tubes 32, 61
Degaussing colour-TV tube* . . . 29,368
Delay line, ultrasonic, for PAL TV receivers 29,243
Delta modulation 31,338
Demonstration and test vehicle for exterior lighting* 35,307
Density of small particles, measurement 28, 30
Deposition by electron beam . . . 28,174
Diameter variations in wire, measurement. 31,111
Diamond, drilling by electron beam 28,177
Dielectric transmission lines for mm waves . . . . . 26,342 – 26,350
Diffractometry, see **X-rays**
Diffusion:
  anomalous, in Si 28,149
  alloy-diffusion technology . . 28,246
  of donors and acceptors in Si 35,181
Digital circuits:
  for data transmission 30, 71
  integrated, with MOS transistors 31,277
  'structural' testing. 35,261
  integrated, with low dissipation 35,312
Digital circuit blocks:
  system based on 1 active and 1 passive logic circuit (NOR and AND) 26,164
  10-series and 20-series 28, 44
  application of 10-series and 20-series 29, 19
Digital systems:
  unit for tuning musical instruments 26,170
  direction finder 28,226
  for displacement measurements by phase grating 30,149
  for displacement measurements by laser interferometer . . . 30,160
  control of companding for code modulation. 31,335
  tone generation for keyboard instrument 31,354
Dimming fluorescent lamps . . . 27,263
Direction finder, digital 28,226
Discharge lamps:
  containing iodides 29,353
  general 35,308
  with metal halides 35,347
  getter for metal-iodide mercury lamps 35,354
  electrodes 35,356
  see also **Fluorescent lamps** and **Sodium lamps**
Discharge tube, cold-cathode, for flame-failure control 29,259
Disease control in agriculture, history of 28,121
Dishwashers, fluidics control for . 33, 29
Dislocations:
  detection of, by anomalous X-ray transmission 29,114

formation and movement of (theory) 32,250
vacancy clusters in dislocation-free Si and Ge 34,244
Dispenser cathodes, osmium . . . 27, 69
Dispersions of pesticides, atomization of 28,112
Displacement measurements:
digital, with phase grating . . 30,149
digital, with laser interferometer 30,160
Display tubes, see **Television, picture tubes**
Distance measurement by microwaves (MADGE) 35,271
Dodewaard nuclear reactor, monitoring control-rod operation . . . 30, 48
Domain structure:
Lorentz microscopy 29,218
compensation wall 34, 96
bubble memories 35, 1
Doping:
in the epitactic growth of Si and Ge layers 26,194
with Si, anomalous diffusion . 28,149
alloy-diffusion technology (for p.o.b. transistors) 28,246
by ion implantation 31,267
Dosage of oxygen in gas systems . 31,112
Double glazing with good thermal insulation 34,242
Drawing by computer:
using COBRA to make stereo drawings 29,102
stereographic presentation of 'Duphaston' molecule . . . 29,110
stereo representation of solution of Van der Pol equation 29,112
of masks for integrated circuits 30, 29
Drilling:
of diamonds by electron beams 28,177
of core with hollow polygonal bit* 30,307
Droplet interferometry for investigating smooth surfaces 33, 74
Duphar, see Philips-Duphar
'Duphaston' molecule, stereographic presentation of, by computer . . 29,110

Ear-lobe clip for physiological transducers 33,102
Eddy-current coupling as variable-speed drive 27, 15
Elastic waves 33,310
Electric motors:
small, with ferroxdure rotor . . 26,143
small induction motors with thyristor control 28, 1
miniature motor for experiments on vision 28,320
small brushless d.c. motor . . 30, 7
electrostatic motors 30,178
brushless d.c., with integrated Hall device 31,366
general 33,215
stability of small single-phase synchronous motors 33,235
moving-coil motors 33,244
electromagnetic vibrators . . 33,249
frequency-analog speed control 33,260
for ANS reaction wheels . . . 34,106
for ANS horizon sensor . . . 34,213
torque measurements on induction motors 34,153
d.c. motor with speed control 34,163
high-speed solid-rotor induction motors 34,170
speed control for capacitor motor 34,180
early history 35, 77
designing small d.c. motor . . 35, 96
applications of stepping motors 35,104
d.c., with electronic commutation 35,106
scaling laws 35,116
Electro-elastic waves 33,310
Electrolytic capacitors, non-polar 29, 37
Electromagnetism:
theory 32,150
scattering characteristics of crossed waveguides 32,165
electromagnetic waves 33,310
Electrons, surface analysis with . . 34,357
Electron beam:
in plasma 27,275
chemical deposition by . . . 28,174
drilling diamond by 28,177
in IC manufacture 35, 41
Electron microprobe 34,370
Electron microscope:
stereography with 27,231
with TV display 29,294
EM 300 29,370
photographs:
Bragg reflections* 26, 48
glass fracture surface* . . 26, 85
SiC whiskers 26,187, 29,252
frosting of glass 29,142
see also **Scanning electron microscope**
Electron multipliers, see **Channel electron multiplier**
Electron resists in IC manufacture 35, 41
Electronic devices, innovation in . 32,117
Electronic keyboard instrument with digital tone generation . . . 31,354
Electrostatic motors 30,178
Electrostatically focused 'Plumbicon' tubes 28, 60
Ellipsometry 29,135
Emission spectrometry:
in multielement analysis, general 34,305
automatic analysis of emission spectra 34,322
Endoscope, experimental, with miniature TV camera 35,166
Energy paper 28,298
Energy spectra, three-dimensional 27,132
Environmental science:
SO_2 monitoring network for Rhine estuary region 32, 33
monitoring centre for air pollution* 33,194
ozone, NO and NO_2 detectors 34, 73
monitoring the quality of surface water 34,113
measuring the oxygen demand of water 34,123
automatic extractor 35,196
Enzyme, substrate and product . . 27,160
Epitactic growth of GaAs 32,380
ESCA (electron spectroscopy for chemical analysis) 34,361
Etching with ions 35,199
Evaporation, see **Vapour deposition**
EVOLUON 31,187
Expansion ejector 28,243
Exposure control, automatic, for X-ray diagnosis 27, 92
Extractor, automatic 35,196
Extrusion of glass 32, 96

Feedback, motional, in loudspeakers 29,145
Ferrimagnetic materials:
compensation wall 34, 96
epitaxial growth of iron garnets for bubble memories 35, 1
Ferrites:
single-crystal research 28,135
in inductor cores 28,184, 28,186
ferroxplana, crystal growth of* 28,365
ferroxcube for proton synchrotrons 30,312
thermogravimetric analysis , . 31, 24
in kicker magnets for proton beams 31, 66
control of properties via microstructure 32, 83
microwave integrated circuits on ferrite substrate 32,315
sintering 35,191
see also **Ferroxdure**
Ferroelectrics:
ceramic, microstructure of . . 32, 90
for detecting infrared 35,247
Ferromagnetic metals, electrical conduction in 35, 29
Ferroxcube for proton synchrotrons 30,312
Ferroxdure:
in small electric motors . . . 26,143
in miniature electric motor . . 28,320
in d.c. motor for washing machine 34,163
Ferroxplana, crystal growth of* . 28,365
Field-effect transistors:
thin-film transistor 27,200
development of FET electronics (Prof. F. M. Klaassen's inaugural lecture) 33,203
Field emission of electrons and ions 33,277
Filament, boron, for composite materials 35,125
Films, X-ray, determining noise in 31,117
Filters:
based on acoustic surface waves 32,179
YIG, for microwaves 32,322
inductorless 33,294
Flame-failure control with UV detector 29,254, 29,260
Flash lamps, combustion, fluorine reactions in 26,113
Fluidics in dishwasher control . . 33, 29
Fluorescence analysis, X-ray . . 34,339, 34,353
Fluorescent lamps:
continuous drawing of glass tube for* 27,238
dimming 27,265
long, electronic starter for . . 31, 54
low-pressure gas discharges, general 35,321
optimum colour rendering . . 35,361
Fluorescent screen for colour TV 27, 22
Fluorine:
chemistry 26,111
cyclic processes in halogen lamps 35,302, 35,316
profiling filaments 35,332
Flywheel for stabilizing space vehicles 30, 2
Frequency-analog systems:
conversion into frequency-analog signals 29,189
for speed control 33,260
Frequency dividers with 'controlled cascading' 32,103
Frosting of glass 29,142
Furnaces, continuous, for rod material 27,337

Gallium-antimonide/metal contacts, tunnelling in 32,214
Gallium arsenide:
tunnelling in contacts with metal 32,213
epitaxial 32,350
diode, combined with IR phosphors 34, 31
investigation of crystal defects 35,239
Gallium phosphide:
light sources and photocells . . 26,136
radiating centres in 32,261
investigation of crystal defects 35,239
Gallium selenide, stoichiometric analysis 34,350
Gamma radiation, see **Radiation monitoring**
Gamma spectrometry:
in prospecting for minerals . . 30, 97
neutron activation analysis 34,330; 34,351
Garnets:
research on single crystals . . 28,135
YIG microwave filters 32,322
iron-garnet films for bubble memories 35, 1
magnetostriction of Ru and Ir ions in YIG 35,225
Gas discharges:
general 35,310
low-pressure 35,321
investigation by light scattering 35,344
Gas-discharge lamps, see **Discharge lamps**
Gas kinetics 33, 43
Gas laser, see **Lasers**
Gas lubrication 29,186
Gearbox, automobile, electronically controlled 28,179
'Gems' of lithium niobate* . . . 31, 23
Germanium:
investigations on the interface with electrolyte 26,127
epitaxial, gas doping of . . . 26,194
chemical behaviour of clean surfaces 32,131
dislocation-free, vacancy clusters in 34,244
see also **Transistors**
Getter for metal-iodide mercury lamps 35,354
Glass:
fracture surface (electron micrograph)* 26, 85
continuous drawing, for fluorescent lamps* 27,238
frosting of 29,142
extrusion of 32, 96
Glassy carbon* 31,369
Gold nuggets, growing 30, 89
Graphite, pyrolitic 28,231
Group theory 32,152
Gunn effect:
Gunn diode 32,370
computer calculations 32,385
oscillators and amplifiers . . . 32,397
Gyrator 33,305

Hall devices:
integrated, for brushless d.c. motors 31,366
for measuring induction-motor torque 34,153
Halogen lamps, see **Incandescent lamps**
Hamburg and Aachen, Philips laboratories at . . . . . 26, 89; -176
Handwriting analysis and synthesis 32, 73
Harp cathode 26,309
Hearing aids:
measurement of* 27, 46
worn behind the ear 27,258
Heart catheterization* 29,298
Heartbeat, recording with ear-lobe clip 33,102
Heat exchanger, compact, of high thermal efficiency 29,158
Heat-flux transformers (with heat pipe) 33,141
Heat pipes:
operation and characteristics . 33,104
applications 33,138
application in Stirling engine for vehicles 31,168; 31,176
Heating elements, PTC thermistors as 30,170
Heating quartz glass with a plasma torch* 34, 60
Helical-groove bearings, see **Bearings**
Helicon waves 33,310
Helium II, vortices in 32,256
Helium liquefier with gas-refrigerating machines:
heat exchanger 29,158
design and characteristics . . 29,197
compressor for 30, 51
Hemiplegics, muscle stimulator for 30, 23
H.F. amplifiers, MOS transistors as 31,251
H.F. generator for argon-plasma torch 33, 54
H.F. heating, control with variable inductor 27,134
High-precision spark machining . . 30,195
High-speed oscilloscope 30,256
History:
early history of telegraphy . . 26,268
of pest and disease control in agriculture 28,121
40 years of workshop technology 31,127
electric motors 35, 77
Holographic strain analysis . . . 35, 53
Hormone action, retrosteroids with 28, 70
Hot-gas engine, see **Stirling engine**
Hot pressing:
continuous 29, 45
of ceramics, without metallic encapsulating layer 35, 65
Hot-spot model for filament burnout 35,296
HPI lamps 29,353; 35,347
Hybrid ICs, connection patterns . 35,144
Hydrogen:
ultra-pure, by cryogenic method 26, 27
in discharge lamps 35,354
Hydrostatic bearings and drive, precision lathe with 30,117
Hydrothermal crystal growth and crystal synthesis 30, 89

Image intensifiers:
research on, at Mullard . . . 28,161
image converter as high-speed shutter 30,241
see also **TITUS tube**
Impact-mounting of components . 27, 28
IMPATT diodes:
general 32,328
in microwave oscillators . . . 32,345
anomalous oscillations 32,361
Inaugural lectures:
Prof. N. F. Verster 26,321
Prof. A. L. Stuijts 31, 44
Prof. F. M. Klaassen 33,203
Incandescent lamps:
transport reactions with fluorine in 26,114
gas-filled, containing bromine and chlorine 27,173
non-destructive measurement of pressures in 31, 93
burn-out 32,155; 35,296
gas-filled, temperature distribution 32,206
research 35,295
halogen, cyclic processes in . . 35,302
carbon-filament, with halogen filling 35,316
temperature profiling of filament 35,332
Indium antimonide and arsenide for infrared perception 30,264
Indium-oxide filter in low-pressure sodium lamps 29, 17
Induction motors:
thyristor control 28, 1
general 33,226
torque measurements on . . 34,153
high-speed, with solid rotor . . 34,170
Inductors:
variable, for r.f. heating control 27,134
ferrite-cored, variability in . . 28,184
Q contour charts 28,186
simulated, in filters 33,294
see also **Coils**
Industrial research:
innovation in electronic devices 32,117
and theoretical physics . . . 32,149
Inertial navigation 29, 69
Infrared:
reflection in, as method of investigation 29,129
far, image perception in . . . 30,264
television 30,270
television, passive, at 4 and 10 μm 30,278
objective lenses for 30,290
phosphors 34, 24
horizon sensor in ANS satellite 34,213
improved thermal insulation of double glazing 34,242
pyroelectric detectors 35,247
Injection logic:
integrated (I^2L) 33, 76
low-dissipation 35,217
Inorganic chemical analysis:
general 34,298
multielement analysis by optical emission spectrometry . . 34,305
automatic analysis of emission spectra 34,322
neutron activation analysis 34,330; 34,351
line-intensity calculations for X-ray fluorescence analysis . 34,339
mass-spectrographic determination of C, O and N . . . 34,344
stoichiometric analysis of GaSe 34,350
determining the content of active oxygen in oxides* . . 34,356
surface analyses 34,357
electron microprobe 34,370
simple extractor 35,196
Insecticides, atomization of . . . 28,112
Integrated circuits:
introduction 27,180
evaporated circuits 27,182
monolithic solid circuits . . . 27,192
manufacture* 27,207
tunable 27,264

applications of contour deposition (MOS transistors, etc.) . 28,153
bucket-brigade delay line, shift register for analog signals 31, 97, 31,266
digital, with MOS transistors 31,277
MOST store with discretionary wiring 31,286
Hall device for brushless d.c. motors 31,366
linear basic 32, 1
integrated injection logic (I^2L) 33, 76
mounted with leads on tape . . 34, 86
hybrid, connection patterns for 35,144
digital, low-dissipation 35,212
testing LSI circuits 35,261
see also:
LOCOS technology
Masks for integrated circuits
MOS technology

Integrated microwave circuits:
general 32,292
with lumped elements 32,305
on ferrite substrate 32,315

Interconnection patterns:
on tape 34, 86
in thick-film technology . . . 35,144

Interferometry:
Narrabri stellar interferometer 27,141, 27,153
radio astronomy at Defford . . 27,313
direction finding by multiple interferometer 28,226
for measuring displacements 30,149, 30,160
droplet, for investigating smooth surfaces 33, 74
in MADGE aircraft-landing guidance system 34,225
Intersynchronization of TV cameras 30,300
Iodide discharge lamps 29,353

Ions:
implantation in MOS transistors 31,267
surface analysis with 34,362
ion etching 35,199
Iron and steel, silicon, nitrogen in 26, 81
Isothermal space (with heat pipe) 33,139

Keyboard instrument, transposing, with digital tone generation . . . 31,354
Kicker magnet for proton beams . 31, 66
Kick-sorter spectral display . . . 27,132

Klystron:
multiplier for 0.8 mm 26, 61
reflex, simple, for 8 mm . . . 27,310
reflex, for 1.5 mm 28,284

Laboratories:
of an X-ray equipment factory* 27,322
Cyclotron and Isotope, at Petten* 28, 50
agrobiological: 'Boekesteyn'* 28,111
Waalre complex of Philips Research Laboratories ('Nat. Lab.') 31,153
see also **Research at Philips**
Lacquers, electron-sensitive, for IC fabrication 35, 41

Lamps, see:
Discharge lamps
Incandescent lamps
Mercury lamps
Sodium lamps

Lasers:
thickness measurement by laser beam 27,208
sealed-off CO_2, for high power 28,287
in displacement measurements 30,160
single crystals for* 30,263
gas, optical polarization in . . 32,190
in scanning 'VLP' (video long-play) system 33,186
Lateral skin effect in a flat conductor 32,221
Lathe, precision, with hydrostatic bearings and drive 30,117
LED (light-emitting diode) 32,261
LEED (low-energy electron diffraction) 34,358
LEP 30,213-308

Levitation:
diamagnetic, in micromanometer 30,231
theory 34, 67
Light amplification, fundamentals 28,165
Light generation with selective thermal radiators 26, 33
Light scattering in investigation of gas discharges 35,344
Light sources, optimum colour rendering 35,361

Lighting:
GaP light sources 26,136
of tunnel entrances 27, 76
dimming fluorescent lamps . . 27,265
of roads, design method . . . 29,299
of roads, demonstration and test vehicle* 35,307
of roads, open-air laboratory* 35,320
see also **Colour rendering** and Lamps
Linear basic circuits, integrated . . 32, 1
Lithium niobate, 'gems' of * . . . 31, 23
Live-stock feeds, vitamins in . . . 28, 85

LOCOS technology
for MOS transistors 31,234
for bipolar transistors* . . . 31,276
for complementary MOS circuits: LOCMOS 34, 19

Logic circuits:
fast, with low energy consumption 29,363
fluidics, for dishwasher control 33, 29
integrated injection logic (I^2L) 33, 76
large, structural test methods for 35,261
Lorentz microscopy 29,218
Loudspeakers with motional feedback 29,148
LPE (liquid-phase epitaxy) . . . 32,380, 35, 1

Luminescence:
characteristic:
absorption and emission spectra of important activators 31,304
efficiency of phosphors excited in the activator 31,314
energy transfer and efficiency 31,324
electroluminescence 32,261
screen with pillar structure* . 33,161
detectors based on chemiluminescence 34, 73
cathodo-, in investigating crystal defects 35,239
see also **Phosphors**
Lumped-element microwave components 32,305

MADGE guidance system for aircraft landing 34,225, 35,271
Magnets, permanent 34,193
Magnets, superconducting (with addendum, 30, 147) 29,309
Magnetic deflection in TV picture tubes 32, 61
Magnetic liquids* 33,293

Magnetic materials:
Mössbauer spectra 28, 33
with $(BH)_{max}$ of 18.5 million gauss oersteds 29,336
for permanent magnets . . . 34,193
anisotropy and magnetostriction of Ru and Ir ions in YIG 35,225
see also **Ferrimagnetic materials and Ferrites**

Magnetic recording:
non-professional video tape recorder 26,177
for telegraphy 26,250
audio tape cassettes 31, 77

Magnetism:
crystal chemistry and . . . 28, 13, 28, 23
research on ferrites and garnets 28,135
deposition of magn. thin films* 28,196
investigation of domain structure 29,218
photomagnetic effects 31, 33
magnetoacoustic effects in bismuth 32,233
magnetic liquids* 33,293
levitated magnets 34, 67
the compensation wall in ferrimagnetic materials 34, 96
bubble memories of iron garnet 35, 1
Magnetometer, vibrating-reed . . 31, 40
Magneto-optical memories 34, 96
Magnetostriction of Ru and Ir ions in YIG 35,225
Magnetron for frequency-agile radar 28,323
Manometers, see **Pressure measurements**

Masers:
for radio-astronomy (Defford interferometer) 27,313
physics of materials for . . . 28,146

Masks for integrated circuits:
automated drawing of 30, 29
step-and-repeat camera . . . 30, 57
making photomasks with the Opthycograph 34,257

Mass spectrometry:
analysis of solid surfaces . . 27,344, 34,362
determination of C, O and N in semiconductors 34,344
Mathieu equation 34, 61

Measurement and control:
industrial equipment with digital circuit blocks 29, 19
with frequency-analog signals 29,189, 33,260, 34,288
attitude control of ANS satellite 33,162, 34,208
control systems in 'VLP' player 33,186, 33,190
reaction wheels for ANS satellite 34,106
servocontrol of Opthycograph 34,263

Mechanization:
eddy-current coupling for variable speed drive 27, 15
impact mounting of components 27, 28
mounting ICs and conductor patterns on tape 34, 85
see also **Numerical control**

Medical technology:
- automatic exposure control for X-ray diagnosis 27, 92
- X-ray colour pictures 27,228
- hearing aids (behind-the-ear types) 27,258
- universal vectorcardiograph . 29, 42
- muscle stimulator for hemiplegic patients 30, 23
- infrared thermography. . . . 30,278
- ear-lobe clip for physiological transducers 33,102
- experimental endoscope with miniature TV camera 35,166
- automatic measurements on X-ray pictures 35,170
- see also **Pharmaceutical research**

Memories:
- development for computers . . 26, 67
- in automatic telegraph exchanges. 26,240
- magnetic-tape, for telegraph characters, with static read-out 26,250
- fast, parameters of evaporated NiFe films for 28,188
- magnetic thin films for* . . . 28,196
- bucket-brigade delay line, shift register for analog signals 31, 97; 31,266
- MOST store with discretionary wiring 31,286
- bubble memory of ion garnet 35, 1

Mercury lamps:
- low-pressure, with fluorescence, see **Fluorescent lamps**
- sorting quartz tubing for . . 32, 57
- high-pressure, with metal iodides, hydrogen getter for 35,354

Metals:
- simple model for alloys . . 33,149; 33,196
- surface imaging by field emission. 33,277
- ferromagnetic, electrical conduction in 35, 29
- metallization technique for ceramic-to-metal bonding . . 35,209

Metal-halide lamps 35,347
Metal-inert-gas (MIG) welding, with plasma 33, 21
'Micro Cryogem' miniature Stirling refrigerator 32, 51
Microelectronic processing, see **Electron beam**
Microfilms by PD process 33, 1
Microfractography of thin films . 27, 87
Microprobe, electron 34,370
Microprogram control of P1000 family of computers 34,132

Microscopy:
- imaging metal surfaces by field emission 33,277
- development of avalanche photodiodes. 35, 23
- see also **Electron microscope** and **Scanning electron microscope**

Microstructure of electroceramic materials 32, 79

Microwaves:
- surface-wave transmission lines 26,342; 26,350
- coaxial and stripline components 28,211
- solid-state electronics for . . 32,282
- integrated microwave circuits 32,292
- lumped-element circuits . . . 32,305
- integrated circuits on ferrite substrate 32,315
- YIG filters 32,322
- IMPATT diodes 32,328
- IMPATT oscillators. 32,345
- anomalous oscillations in IMPATT diodes 32,361
- Gunn diode 32,370
- epitaxial GaAs for microwave devices 32,380
- computer calculations of Gunn effect 32,385
- Gunn effect oscillators and amplifiers 32,397
- *P-I-N* diodes in phase-shifters for antenna arrays 32,405
- MADGE aircraft-guidance system 34,225; 35,271
- see also:
  - **Klystron**
  - **Radar**
  - **Waveguides, oversized**

Minerals, prospecting for, with neutrons 30, 97
Mobility of carriers in MOS transistors 31,237
Modulation, code, with digitally controlled companding 31,335
Monitoring centre for air pollution* 33,194
Monitoring network for SO_2 in Rhine estuary region 32, 33
Monograin layers 29,238
Monolithic circuits 27,192
Mortar made of SiC 26,340

MOS technology:
- general 31,225
- with ion implantation 31,267
- in thin monocrystalline Si layers 31,271
- store with discretionary wiring 31,286
- LOCMOS, for complementary circuits 34, 19
- ICs with low dissipation . . . 35,212
- see also **MOS tetrodes** and **MOS transistors**

MOS tetrodes:
- general 30,134
- with short channel for UHF 31,259
- in bucket-brigade delay line . . 31,266

MOS transistors:
- complementary pairs 28,154
- general 31,206
- operation and d.c. behaviour 31,209
- as small-signal amplifier . . . 31,216
- LOCOS technology 31,234; 31,276
- carrier mobility 31,237
- in integrated audio amplifiers 31,245
- in integrated chopper circuit . 31,248
- for power amplification in HF band 31,251
- in digital integrated circuits . . 31,277
- see also **Field-effect transistors**

Mosaic printers for computers, etc. 29,205
Motional feedback in loudspeakers 29,148
Motors, see **Electric motors** and **Stirling engine**
Moving-coil motors 33,244
Mullard Research Laboratories 28,129-230
Multiplier tubes, see **Photomultipliers**
Muscle stimulator for hemiplegic patients 30, 23

Musical instruments:
- with intervals produced digitally 26,170
- digital tone generation for transposing keyboard instrument 31,354

Narrabri stellar interferometer . 27,141; 27,152
Navigation, inertial 29, 69
Network theory for inductorless filters. 33,294

Neutrons:
- prospecting with 30, 97
- activation analysis 34,330; 34,351

Nickel-iron films for fast memories 28,188
Night vision 28,161
'Nimrod' proton synchrotron . . . 30,330
NIRMS (inert-gas-ion reflection mass spectrometry). 34,363
Nitrogen in silicon iron and steel 26, 81
NO and NO_2 detectors 34, 73
Noise in X-ray films, determining 31,117
Non-polar electrolytic capacitors . 29, 37
NTSC colour-TV system 27, 33
Nuclear reactor at Dodewaard, monitoring of control-rod operation 30, 48
Nuclear research with the 'BOL' project* 30, 35
Numerals, handwritten, optical character recognition of . . . 33, 89; 33,130

Numerical control:
- with the small COBRA computer 27,285
- APT language 28,329
- see also **Drawing by computer**

Onsager relations 33,310
Opthycograph, for optical drawing of photomasks 34,257

Optical character recognition:
- general 28,197
- equipment* 28,203
- for handwritten numerals . 33, 89; 33,130

Opto-acoustic cross-correlator for radar 28,217

Opto-electronics:
- GaP light sources and photocells 26,136
- physics of radiative centres in GaP 32,261

Oscilloscope, high-speed 30,256
Osmium dispenser cathodes . . . 27, 69
Oxides, determining active-oxygen content of* 34,356
Oxide materials, crystal chemistry and magnetism of 28, 13; 28, 23

Oxygen:
- monitoring low partial pressures 31,112
- oxygen demand of water . . 34,123
- determining active-oxygen content of oxides* 34,356

Ozone detectors 34, 73

PAILRED:
- short description 26,316
- principles 29,165
- automatic control circuits . . 29,176

PAL system for colour TV:
- comparison with other systems 27, 33
- delay line 29,243

Parametric amplifiers:
- low-noise 28,204
- for radio astronomy 32, 20

PD process:
- general 33, 1
- connection patterns for ICs . . 34, 85

Peltier cooling 27,113

Perception:
miniature electric motor for experiments on vision . . . 28,320
analysis and synthesis of handwriting 32, 73
Peripheral equipment for computers:
problems and trends in development 26,148
fast punched-tape reader . . . 28,259
high-speed printers 29,205
see also **Optical character recognition**
Permanent magnets, materials for 34,193
Permittivity tensor 33,310
Pesticides, atomization of dispersions of 28,112
Pharmaceutical research:
retrosteroids 28, 70
virus-vaccine preparation . . 28, 81
Phase grating, displacement measurements with 30,149
Phase-shifters with *P-I-N* diodes, for antenna arrays 32,405
Phase theory:
unary systems 26, 12
binary systems 26, 12; 26, 52
ternary systems 27,213
thermogravimetric analysis applied to ferrites 31, 24
simple model for alloys . . 33,149; 33,196
Phenol synthesis and photomorphogenesis 28,101
Philips-Duphar . . . . . 28, 69; -128
Phosphors:
for colour TV 27, 22; 32,125
infrared 34, 24
see also **Luminescence**
Photoconductivity:
principles: recombination process, sensitivity, speed of response; inhomogeneity effects 27, 47; 29,221
monograin layers 29,238
for image perception in the infrared 30,264
targets for infrared TV camera tubes 30,270
see also **'Plumbicon'**
Photodiodes of GaP 26,136
Photoemission:
of semiconductors 29, 54
of alkali metals 30,225
Photography:
flash lamps, combustion-type, with fluorine reactions . . . 26,113
high-speed camera shutters . . 28,161
ultra-high-speed 30,241
PD process 33, 1
Photomagnetic effects 31, 33
Photomasks, see **Masks for integrated circuits**
Photomorphogenesis 28,101
Photomultipliers:
general 29,267
application of the transfer technique 30,236
hybrid 30,236
with channel plate 30,236
Photons, surface analysis with . . 34,358
Photoresistors, see **Photoconductivity**
Photosynthesis 27,241
Phototitus optical converter . . . 34,274
Physiological transducers, ear-lobe clip for 33,102
Phytopharmaceutical research, see **Agrobiological research**
Pictures, transmission of simple . . 32, 42
Picture tubes, see **Television, picture tubes**
Piezoelectrics:
ceramic, microstructure of . . 32, 91
waves in 33,310
P-I-N diodes in phase shifters for antenna arrays 32,405
PINPOINT radio system for locating and monitoring vehicles . . . 35, 15
Pirani gauge for pressures up to 1000 torr and higher 30,166
Plasma:
amplifier tubes with electron beam in 27,275
MIG welding 33, 21
r.f. argon-plasma torch for emission spectroscopy . . . 33, 50
torch for heating quartz glass* 34, 60
'Plumbicon' television camera tubes:
with improved sensitivity to red light 26, 49
comparison with other tubes 27, 1
experimental electrostatically focused 28, 60
application in colour camera 28,336
in electron-microscope display 29,294
for portable colour camera . . 29,325
P.O.B. transistors 28,246
Pockels effect 30,292; 34,274
Polarization, optical, in gas laser . . 32,190
Polytypism of SiC 30, 36
Power amplifiers, high-frequency, MOS transistors as 31,251
Power supplies, transistorized, current-limiting circuits in 28,251
Precision lathe with hydrostatic bearings and drive 30,117
Pressure measurements:
Pirani gauge for pressures up to 1000 torr and higher . . . 30,166
non-destructive, in lamps . . . 31, 93
monitoring oxygen partial pressures down to 10^{-30} bar 31,112
miniature pressure transducers with silicon diaphragm . . . 33, 14
Pressure sintering, see **Hot pressing**
Princeton proton synchrotron . . 30,330
Printed circuits, impact mounting of components 27, 28
Printer, high-speed mosaic 29,205
Programs, see **Software**
Proton synchrotrons:
ferroxcube for 30,312
AGS, CPS, ZGS, Nimrod and PPA 30,330
kicker magnets for 31, 66
stabilization of proton orbits 34, 64
PTC thermistors:
of $BaTiO_3$ 26,119
as thermostatic heating elements 30,170
Pulse-code modulation 31,335
Pulse-height analyser, spectral display 27,132
Punched-tape reader for 2500 characters/s without buffer store . . . 28,259
PW 1100 single-crystal diffractometer 33, 61
Pyroelectric infrared detectors . . 35,247
Pyrolytic graphite 28,231

Q contour charts for ferrite-cored inductors 28,186
Quartz tubing, sorting, for mercury lamps 32, 57

Radar:
harbour surveillance, for the Elbe* 27,169
ultrasonic* 28,173
opto-acoustic cross-correlator 28,217
frequency-agile 28,323
simulation of antenna arrays by analog computer 31, 2
traffic-flow analysis by 31, 17
AVOID, short-range high-definition radar 32, 13
Radiation monitoring:
sensitive monitor for X- and gamma rays 26,190
semiconductor detectors:
general 27,323
for α-radiation 30, 13
Radio astronomy:
masers for interferometer . . 27,313
simulation of antenna arrays by analog computer 31, 2
parametric amplifiers for radio interferometer 32, 20
Radio broadcasting, stereophonic 26,327; 27, 62
Radio-frequency, see R.F.
Radio system for vehicle location (PINPOINT) 35, 15
RAMP inertial navigation system 29, 69
Reaction wheels for ANS satellite 34,106
Record-player cartridges, sawing crystals for* 28,297
Recording, see **Magnetic recording** and **'VLP' video long-playing system**
Refractometer for thin films . . . 35,142
Research at Philips:
at Philips Research Laboratories ('Nat. Lab.') 24,339-414
at Philips Aachen and Hamburg Laboratories 26, 89-176
at Philips-Duphar 28, 69-128
at Mullard Research Laboratories (MRL) 28,129-230
at LEP Laboratories, Limeil-Brévannes 30,213-308
Research Laboratories, Waalre complex 31,153
Retrosteroids 28, 70
R.F. amplifiers, MOS transistors as 31,251
R.F. generator for argon-plasma torch 33, 54
R.F. heating, control with variable inductor 27,134
Rhine estuary region, SO_2 monitoring network for 32, 33
Road lighting, see **Lighting**
Rolling diaphragm seal:
sealing high pressure differences with 26,287
application:
in Stirling engine 26,292
in industrial gas refrigerating machine 26,297
in compressors 26,294; 30, 51

Samarium-cobalt magnet 29,336
Satellite, astronomical, see **ANS**
Sawing crystals for record-player cartridges* 28,297
Scaling laws for electric motors . . 35,116
Scanning electron microscope:
investigation of microchannel plates 34,270
electron microprobe 34,370

PSEM 500 35,153
investigation of crystal defects by cathodoluminescence . . 35,239
see also **Electron microscope**
Scanning microscope for development of avalanche photodiodes . . 35, 23
Seals, see **Rolling diaphragm seal**
SECAM colour-TV system 27, 33
Selective thermal radiators in generating light 26, 33
SEM, see **Scanning electron microscope**
Semiconductors:
for radiation detectors . . . 27,323; 30, 13
photoemission 29, 54
optical investigation of surfaces 29,129
single crystals for circuit elements* 30,263
ceramic, effect of microstructure 32, 92
tunnelling in metal-semiconductor contacts under pressure 32,211
in microwave electronics . . . 32,282
mass-spectrographic determination of C, N and O in . . 34,344
see also:
Gallium arsenide
Gallium phosphide
Germanium
Gunn effect
Photoconductivity
Silicon
Semiconductor technology:
epitaxial gallium arsenide . . 32,380
with electron resists 35, 41
transport of crystal defects in solids 35,181
ion etching 35,199
see also:
Integrated circuits
Integrated microwave circuits
MOS technology
Transistors
Shaver, electric, vibrator for . . . 33,249
Shift register for analog signals (bucket-brigade delay line) . . 31, 97; 31,266
Silicon:
epitaxial, spark doping of . . 26,194
anomalous diffusion of impurities in 28,149
alloy-diffusion method 28,246
thin-film MOS transistors . . 31,271
chemical behaviour of clean surfaces 32,131
diaphragm, in miniature pressure transducers 33, 14
dislocation-free, vacancy clusters in 34,244
diffusion of donors and acceptors 35,181
see also **Integrated circuits** and **Transistors**
Silicon carbide:
whiskers 26,187; 29,252
mortar made of 26,340
polytypism 30, 36
Silicon iron and steel, nitrogen in 26, 81
SIMS (secondary-ion mass spectrometry) 34,366
Single crystals for semiconductor devices and lasers* 30,263
Sintering, see **Ceramics technology**
Skin effect:
introduction 28,271
at high frequencies 28,300
in superconductors 28,366
lateral, in flat conductors . . 32,221
SO_2 monitoring network for Rhine estuary region 32, 33
Sodium lamps:
thermal insulation for . . . 26,105
low-pressure, with In_2O_3 filter 29, 17
low-pressure 35,321
high-pressure 35,334
Software:
calculation of traffic-light sequences 27,298
APT language 28,329
analysis of methods 29,338
drawing masks for integrated circuits 30, 29
calculations on cross-coupled waveguides 32,165
control of X-ray diffractometer 33, 61
microprogram control for P1000 family of computers . 34,132
Solar cells, monograin layer . . . 29,238
Solderability measurements . . . 28,362
Soldering iron with PTC thermistor as heating element 30,176
Solid circuits, see **Integrated circuits**
Solid-state electronics for microwaves 32,282
Solid-state research:
transport phenomena in solids 26, 99
Mössbauer spectra 28, 33
magnetoacoustic effects in bismuth 32,233
transport of crystal defects . . 35,181
see also:
Crystals
Gallium arsenide
Gallium phosphide
Photoconductivity
Surface
Transistors
Solidification, directional, in growing composites* 32,102
Sorting quartz tubing for mercury-vapour lamps 32, 57
Sound, see **Acoustics** and **Audio**
Space science:
telemetry antenna* 27,131
flywheel for stabilizing space vehicles 30, 2
boron filament as constructional material 35,125
spiral-groove bearings with grease 35,137
see also ANS
Spark machining, high-precision . 30,195
Spectra, optimum, for light sources 35,361
Spectroscopy, emission, r.f. argon-plasma torch for 33, 50
Spinel structure and magnetism . . 28, 23
Spiral-groove bearings, see **Bearings**
Stabilization by oscillation 34, 61
Starter for long fluorescent lamps . 31, 54
Statistical aspects of alarm level in flame-failure control 29,260
Stellar interferometer at Narrabri 27,141; 27,152
Step-and-repeat camera for making photomasks 30, 57
Stepping-motor applications . . . 35,104
Stereo drawings, see **Drawing by computer**
Stereophony in radio broadcasting 26,327; 27, 62
Stereophotography with electron microscope 27,231
Stirling cycle, see **Cryogenics** and **Stirling engine**
Stirling engine:
application of rolling diaphragm seal 26,287
for vehicular propulsion, prospects 31,169
Stoichiometry:
existence region 30, 52
point defects and control of their concentrations 30,107
determination of existence regions 30,142
analysis of GaSe 34,350
Strain gauges, miniature, in an Si diaphragm 33, 14
Stripline and coaxial microwave components 28,211
Structural testing of digital circuits 35,261
Structure analysis, see **Crystals**
Superconductivity:
current adjustment in superconducting magnet 27,313
and skin effect 28,366
general 29, 1
superconducting magnet coils (with addendum, 30, 147) . . 29,309
vorticity in superconductors . 32,252
Surface:
germanium-electrolyte interface 26,127
solid, analysis by mass spectrometer 27,344
semiconductor, optical investigations of 29,129
of pure alkali metals, properties 30,225
of clean Si and Ge, chemical behaviour 32,131
metal, imaging by field emission 33,277
analysis, general 34,357
Surface waves:
electromagnetic, transmission lines for 26,342; 26,350
acoustic, in filters 32,179
acoustic, in piezoelectric materials 33,310
Synchronous motors:
general 33,217
small single-phase, stability of 33,235
Synthesis, hydrothermal chemical 30, 89

Tape cassettes, audio 31, 77
Tape recorders:
non-professional, for TV . . . 26,177
for audio cassettes 31, 87
moving-coil motors for . . . 33,248
Telegraphy:
automatic exchanges with electronic stores 26,240
magnetic-tape store for . . . 26,250
history 26,268
Telephony:
carrier 26,206
companders with high degree of compression 26,215
Television:
picture tubes:
phosphors for colour TV 27, 22
degaussing* 29,368
magnetic deflection . . . 32, 61
recording:
non-professional tape recorder 26,177
see also **'VLP' video long-playing system**
transmitters for bands IV and V 26,256
electron microscopy with TV display 29,294
projection on large screen . . 30,292
outside-broadcast system (mobile transmitters with intersynchronization and automatically trained antennas) . . . 30,300

MOS tetrode for UHF band 31,259
miniature TV camera in experimental endoscope 35,166
see also **Colour television** and **'Plumbicon'**
Temperature distribution in gas-filled incandescent lamps 32,206
Ternary systems in phase theory . 27,213
Tetrodes, see **MOS tetrodes**
Thermal insulation, good, for double glazing 34,242
Thermal radiators, selective, in generating light 26, 33
Thermistors, $BaTiO_3$ 26,119
Thermodynamics, see **Phase theory**
Thermography 30,278
Thermogravimetric analysis applied to ferrites 31, 24
Thick-film technology 35,144
Thickness measurement by laser beam 27,183
Thin films:
thermal insulation of sodium lamps 26,105
microfractography of 27, 87
thin-film transistor 27,200
magnetic, deposition of* . . . 28,196
indium-oxide filter for low-pressure sodium lamps . . . 29, 17
Si, MOS transistors in 31,271
evaporated, fast refractometer for 35,142
see also **Vapour deposition**
Thyristors for controlling induction motors 28, 1
'Ticonal', Mössbauer spectra . . . 28, 41
Tin-halide lamps 35,347
Titrimetric determination of GaSe 34,353
TITUS tube:
introduction 30,292
single, in polychrome data display 34,129
Phototitus optical converter . 34,274
Torque measurements on induction motors 34,153
Torque meter, sensitive, for high frequencies 27,342
Traffic control:
calculation of traffic-light sequences 27,298
traffic-flow analysis by radar . 31, 17
radio system for vehicle location (PINPOINT) 35, 15
see also **Airfields** and **Lighting**
Transfer technique 30,234 30,236
Transistors:
p.o.b., alloy-diffusion method for 28,246
bipolar, LOCOS technology for* 31,276
see also **Field-effect transistors** and **MOS transistors**
Transition metals, alloying behaviour 33,149
Transmission:
of speech by code modulation with digitally controlled companding 31,335
of simple pictures 32, 42
Transmission system, automobile, with electronic control 28,179
Transmitters for television bands IV and V 26,256
Transmitting valves, cathode with low thermal inertia for 26,309
Transport phenomena in solids . . 26, 99
Transport reactions, chemical, for crystal growth 26,117 28,316
Transport reactions, chemical, in incandescent lamps 35,296 35,302 35,316 35,332
Triglycine sulphate (TGS) for infrared detectors 35,247
Tuning of musical instruments . 26,170 31,354
Tunnel entrances, lighting 27, 76
Tunnelling in metal-semiconductor contacts under pressure 32,211
Turn-to-turn diffusion in incandescent lamps 35,296

Ultrasonic delay line for PAL TV receivers 29,243
Ultrasonic radar* 28,174
Unary systems in phase theory . . 26, 12

Vacancy clusters in dislocation-free Si and Ge 34,244
Vaccines, virus, preparation by tissue cultures 28, 81
Vacuum evaporation of reflectors for projection lamps* 30,106
Vacuum technology:
fast cryopump 28,355
Pirani gauge for pressures up to 1000 torr and higher . . . 30,166
absolute micromanometer with diamagnetic levitation . . . 30,231
flow of highly rarefied gases . 33, 43
Van der Pol equation, stereographic representation of solution 29,112
Vapour deposition:
of integrated circuits 27,182
technique for thin and thick films 28, 57
parameters of NiFe films for fast memories 28,188
of magnetic thin films* . . . 28,196
fast refractometer for evaporated thin films 35,142
Vectorcardiograph, universal . . . 29, 42
Vehicular propulsion, prospects of Stirling engine 31,169
Vibrating-reed magnetometer for microscopic particles 31, 40
Vibrators, electromagnetic 33,249
Videorecorder, non-professional . 26,177
Virus vaccines, preparation by tissue cultures 28, 81
Vision:
at very low light levels 28,161
miniature electric motor for experiments on 28,320
Vitamins in live-stock feeding . . 28, 85
'VLP' video long-playing system:
general 33,178
signal processing 33,181
scanning system of player . . 33,186
control mechanism of player . 33,190
Vortices 32,247

Washing machine, speed-controlled d.c. motor for 34,163
Water pollution, see **Environmental science**
Waves: electromagnetic, elastic and electro-elastic 33,310
Waveguides, oversized:
rectangular, for mm waves . . 28, 86
cross-coupled, scattering characteristics 32,165
Weed control: atomization of pesticides 28,112
Weiss domains, see **Domain structure**
Welding, plasma-MIG 33, 21
Whiskers, SiC 26,187 29,252
Windows, double-glazed, with good thermal insulation 34,242
Windows, vacuum-tight, transparent for wide spectral range . . . 30,299
Wire:
thickness measurement by laser beam 27,183
measurement of diameter variations in 31,111
Workshop technology, 40 years of 31,127
Writing, muscular activity in . . . 32, 73

X-rays:
sensitive radiation monitor . . 26,190
automatic exposure control . . 27, 92
colour pictures 27,228
automatic spectrometer . . . 27,300
equipment factory, laboratory* 27,322
anomalous transmission in highly perfect crystals . . . 29,114
PAILRED single-crystal diffractometer 29,165 29,176
noise in X-ray films 31,117
PW 1100 . . . . . 33, 65
fluorescence analysis . . . 34,339 34,353
automatic measurement of medical images 35,170

YIG, see **Garnets**

Author index, Volumes 26-35

Figures in bold type indicate the volume number, and those in ordinary type the page number. Articles published in volumes 1-25 are given in the author indexes at the end of volumes 10, 15, 20, 25 and 30.

Aa, H. H. M. van der, see Rietdijk, J. A.
Aalders, J. W. G., R. J. van Duinen and P. R. Wesselius
The Groningen ultraviolet experiment with the Netherlands astronomical satellite (ANS) 34, 33
Acket, G. A., R. Tijburg and P. J. de Waard
The Gunn diode 32,370
Adams, J. and B. W. Manley
The channel electron multiplier, a new radiation detector 28,156
Adriaanse, R. P. and P. van der Laan
Flame-failure control with a UV-sensitive cold-cathode tube, II. Statistical aspects of the detection process and choice of the alarm level 29,260
Adriaansz, M., see Vriens, L.
Aitchison, C. S.
Low noise parametric amplifiers 28,204
Lumped components for microwave frequencies . . 32,305
Albers, W. and C. Haas
Stoichiometry,
I. Existence region 30, 82
II. Point defects and the control of their concentrations 30,107
III. The determination of existence regions 30,142
Albrecht, C. and J. Proper
A method of determining noise in X-ray films . . . 31,117
Alcock, R. N.
A digital direction finder 28,226
—, D. A. Lucas and R. P. Vincent
'MADGE', a microwave aircraft digital guidance equipment, I. General principles and angle-measuring units 34,225
Ammers, H. van and J. Hesselink
A sensitive monitor for X-rays and gamma rays . . . 26,190
Andrich, E.
PTC thermistors as self-regulating heating elements . 30,170
— and K. H. Härdtl
Investigations on $BaTiO_3$ semiconductors 26,119
Andrieux, G. and C. Loty
A high-speed oscilloscope for real-time use 30,256
Appels, J. A., H. Kalter and E. Kooi
Some problems of MOS technology 31,225
—, see Nielen, J. A. van
Arink, G. J. A.
The onboard computer of the Netherlands astronomical satellite (ANS) 34, 1
Asselman, G. A. A. and D. B. Green
Heat pipes,
I. Operation and characteristics 33,104
II. Applications 33,138
Bacchi, H. and G. Eschard
Ultra-high-speed photography 30,241
Backers, F. Th.
A delay line for PAL colour television receivers . . . 29,243
Backmark, N., J. E. V. Krim and F. Sellberg
Frequency-agile radar 28,323
Bähler, W. and W. van der Hoek
An eddy-current coupling employed as a variable-speed drive 27, 15
Banga, U. H. and W. Mesman
Continuous furnaces for rod material 27,337
Barneveld, E. J. van
Digital circuit blocks 28, 44
Bastings, L. C., see Bruninx, E.
Beckmann, K. H.
Optical investigations of semiconductor surfaces . . 29,129
Beek, L. K. H. van
The PD photographic process 33, 1
Beekmans, N. M. and L. Heyne
An instrument for monitoring low oxygen pressures . 31,112
Beer, A. F.
A MOS transistor store with discretionary wiring . . 31,286
—, J. B. Coughlin and P. J. Daniel
Some applications of contour deposition 28,153
Beijer, L. B., C. A. J. Jacobs and T. Tol
The iodide discharge lamp 29,353
Belevitch, V.
The lateral skin effect in a flat conductor 32,221
Berben, T. J. and C. W. Berghout
Lubrication with gases 29,186
Berghout, C. W., see Berben, T. J.
Berth, M. and J.-J. Brissot
Targets for infra-red television camera tubes 30,270
Berz, F., see Murphy, N. St. J.
Beukering, H. C. J. van, see Rietdijk, J. A.
Beun, M.
A flexible method for automatic reading of handwritten numerals,
I. General description of the recognition method . 33, 89
II. The thinning procedure and determination of the special points 33,130
Beuvens, H. J. H. and J. H. Dettingmeyer
A non-destructive measurement of pressures in incandescent lamps 31, 93
Bhargava, R. N.
The physics of radiative centres in GaP 32,261
Blasse, G.
Crystal chemistry and magnetism of oxide materials,
II. Magnetic compounds with spinel structure . . . 28, 23
— and A. Bril
Characteristic luminescence,
I. The absorption and emission spectra of some important activators 31,304
II. The efficiency of phosphors excited in the activator 31,314
III. Energy transfer and efficiency 31,324
—, see Bril, A.
Bleekrode, R., M. Koedam and L. Rehder
Discharge lamps 35,308
Bloemendal, W. and C. Kramer
The Netherlands astronomical satellite (ANS) . . . 33,117
Blom, D., H. W. Hanneman and J. O. Voorman
Some inductorless filters 33,294
Boer, B. de
Behind-the-ear hearing aids 27,258
Boer, F. J. de, see Boumans, P. W. J. M.
Boers, P. M. and L. J. M. Bollen
A fast scanning microscope used in the development of avalanche photodiodes 35, 23
Bollée, B.
A miniature electric motor for experiments on vision . 28,320
Electrostatic motors 30,178
Bollen, L. J. M., see Boers, P. M.
Bongers, P. F.
Crystal chemistry and magnetism of oxide materials,
I. Principles and application of crystal field theory . 28, 13
Boort, H. J. J. van and R. Groth
Low-pressure sodium lamps with indium-oxide filter 29, 17
Booth, P. L., see Smith, F. W.
Bootsma, G. A., see Meijer, F.
Borne, J.
High-speed printers for numerical data processing equipment 29,205
Bos, J. G. G.
Spiral-groove bearing systems with grease 35,137
Bosch, G. and J. H. H. Janssen
Integrated circuit with Hall device for brushless d.c. motors 31,366
Bosgra, O. and J. H. G. Roerink
Preparation of virus vaccines by means of tissue cultures 28, 81
Boter, P. A. and M. D. Wijnen
Energy paper 28,298

Boucher, A. and B. C. Easton
Epitaxial growth of gallium arsenide 32,380
Boulou, M., see Schiller, C.
Boumans, P. W. J. M.
Multielement analysis by optical emission spectrometry — rise or fall of an empire? 34,305
—, F. J. de Boer and J. W. de Ruiter
A stabilized r.f. argon-plasma torch for emission spectroscopy 33, 50
Boutry, G.-A. and H. Dormont
Some surface properties of pure alkali metals 30,225
Bouwer, A. G., R. H. Bruel, H. F. van Heek, F. T. Klostermann and J. J. 't Mannetje
The Opthycograph 34,257
Bouwhuis, G. and P. Burgstede
The optical scanning system of the Philips 'VLP' record player 33,186
—, see Lang, H. de
Bouwkamp, C. J.
Scattering characteristics of a cross-junction of oversized waveguides 32,165
Bouwknegt, A., H. Nienhuis, D. J. Schipper and P. A. W. Tielemans
Electrodes in discharge lamps 35,356
Bowers, Brian
The early history of the electric motor 35, 77
Brandt, B. B. M., W. Steinmaier and A. J. Strachan
LOCMOS, a new technology for complementary MOS circuits 34, 19
Bregman, F.
Representations of three-dimensional energy spectra 27,132
Breimer, H., W. Holm and S. L. Tan
A colour television camera with 'Plumbicon' camera tubes 28,336
Bril, A., G. Blasse, A. H. Gomes de Mesquita and J. A. de Poorter
Fast phosphors for colour television 32,125
— and W. L. Wanmaker
New phosphors for colour television 27, 22
—, see Blasse, G.
—, see Sommerdijk, J. L.
Brinkman, A. C., J. Heise and C. de Jager
Observation of cosmic X-ray sources with the Netherlands astronomical satellite (ANS) 34, 43
Brissot, J.-J. and A. Lemogne
Method for growing single crystals of cuprous chloride 30,261
—, see Berth, M.
Brockman, F. G., H. van der Heide and M. W. Louwerse
Ferroxcube for proton synchrotrons 30,312
Broerse, P. H., A. C. van Dorsten and H. F. Premsela
Electron microscopy with television display system . . 29,294
—, J. H. T. van Roosmalen and S. L. Tan
An experimental light-weight colour television camera 29,325
Brongersma, H. H., F. Meijer and H. W. Werner
Surface analysis, methods of studying the outer atomic layers of solids 34,357
Bronnes, R. L., R. C. Hughes and R. C. Sweet
Ceramic-to-metal bonding with sputtering as a metallization technique 35,209
Brouwer, G., see Witmer, A. W.
Brouwer, H. J., S. M. de Veer and H. Zeedijk
The SO_2 monitoring network in the Rhine estuary region 32, 33
Browne, A.
The stellar interferometer at Narrabri, Australia, II. Measurement of the correlation 27,152
Bruel, R. H., see Bouwer, A. G.
Bruijning, H. G., see Goorissen, J.
Bruin, O. A. de, H. F. L. Schöler and J. N. Walop
Retro-steroids, a new class of compounds with sex-hormone action 28, 70
Bruining, H. and S. Duinker
The research programme of the Philips laboratories in Germany 26, 93
Bruninx, E. and L. C. Bastings
Stoichiometric analysis of gallium selenide 34,350
Burgstede, P., see Bouwhuis, G.
Buschow, K. H. J., W. Luiten, P. A. Naastepad and F. F. Westendorp
Magnetic material with a $(BH)_{max}$ of 18.5 million gauss oersteds 29,336
Bussche, W. van den, A. H. Hoogendijk and J. H. Wessels
Signal processing in the Philips 'VLP' system . . . 33,181
Butterweck, H. J. and F. C. de Ronde
Oversized rectangular waveguide components for millimetre waves 29, 86
Butzelaar, P. F. and L. P. J. Hoogeveen
A new method of measuring the oxygen demand of water 34,123
Casimir, H. B. G.
Electrostatic machines, particle accelerators and industry 30,309
Theoretical physics and industrial research 32,149
— and J. Ubbink
The skin effect,
I. Introduction; the current distribution for various configurations 28,271
II. The skin effect at high frequencies 28,300
III. The skin effect in superconductors 28,366
Cath, P. G. and J. Ladell
Principles and design of the automatic single-crystal diffractometer system PAILRED,
I. Principles of the diffractometer 29,165
II. The automatic control circuits 29,176
Cayzac, J.
An outside-broadcast equipment with mobile television cameras 30,300
Christis, W. J.
The optical sensors of the Netherlands astronomical satellite (ANS), III. The star sensor 34,218
Clegg, J. B. and E. J. Millett
The determination of carbon, oxygen and nitrogen *in semiconductors* by spark-source mass spectrography 34,344
Compaan, K. and P. Kramer
The Philips 'VLP' system 33,178
Corbett, B. D.
'MADGE', a microwave aircraft digital guidance equipment, II. The data link: data transmission and distance measurement 35,271
Coughlin, J. B., see Beer, A. F.
Crousel, L.
Computer programming 29,338
Crucq, J.
The reaction wheels of the Netherlands satellite ANS 34,106
Daniel, P. J., see Beer, A. F.
Daniels, A. and F. K. du Pré
Miniature refrigerators for electronic devices 32, 49
Davies, R. and R. E. Pearson
Parametric amplifiers for a radio-astronomy interferometer 32, 20
Day, P. E., see Janssen, P. J. M.
Dekkers, A. J., A. van Duuren, F. A. Lootsma and J. Vlietstra
Calculation of traffic-light sequences 27,298
Desvignes, F., J. Revuz and R. Zeida
Photoelectric solid-state devices and the perception of images in the infra-red 30,264
Dettingmeijer, J. H., G. Dittmer, A. Klopfer and J. Schröder
Research on incandescent lamps, III. Regenerative chemical cycles in tungsten-halogen lamps 35,302
—, see Beuvens, H. J. H.
Dijk, P. van
The optical sensors of the Netherlands astronomical satellite (ANS), II. The horizon sensor 34,213
Dijken, R. H.
Designing a small d.c. motor 35, 96
Dimigen, H. and H. Lüthje
An investigation of ion etching 35,199
Dinklo, J. A. and E. B. de Vries
The microprogram control of the Philips P1000 family of computers 34,132
Dittmer, G., see Dettingmeijer, J. H.
Donjon, J. and G. Marie
Polychrome data display using a single TITUS tube . 34,129
Doorn, A. G. van
The 'Plumbicon' compared with other television camera tubes 27, 1
Dorleijn, J. W. F., see Miedema, A. R.
Dormont, H., see Boutry, G.-A.
Dorsten, A. C. van, see Broerse, P. H.

Douma, F. and J. Guldemond
A universal vectorcardiograph 29, 42
Drift, A. van der, W. G. Gelling and A. Rademakers
Integrated circuits with leads on flexible tape 34, 85
Drop, P. C., E. Fischer, F. Oostvogels and G. A. Wesselink
Metal-halide discharge lamps 35,347
Dros, A. A.
An industrial gas refrigerating machine with hydraulic piston drive 26,297
Ducot, C.
The Laboratoires d'Electronique et de Physique Appliquée at Limeil-Brévannes near Paris. Origin — Function — Activities 30,213
Duinen, R. J. van, see Aalders, J. W. G.
Duinker, S., see Bruining, H.
Duis, J. A. ten and E. van der Meulen
Measurement of the solderability of components . . 28,362
Dumont, F., J. P. Hazan and D. Rossier
The Phototitus optical converter 34,274
Duuren, A. van, see Dekkers, A. J.
Duyfjes, W.
The atomization of concentrated dispersions of pesticides 28,112
Easton, B. C., see Boucher, A.
Eisler, P. L., see Wylie, A. W.
Elgström, H. and E. Zieler
An automatic exposure control system for X-ray diagnostics 27, 92
Engelsma, G.
Phenol synthesis and photomorphogenesis 28,101
Enz, U. and R. W. Teale
Photomagnetic effects 31, 33
Eschard, G. and R. Polaert
The production of electron-multiplier channel plates 30,252
—, see Bacchi, H.
Essers, W. G., G. Jelmorini and G. W. Tichelaar
Plasma-MIG welding 33, 21
Evrard, R.
Absolute micromanometers with diamagnetic levitation 30,231
Fast, J. D.
Nitrogen in silicon iron 26, 81
Ferguson, E. T., see Lang, H. de
Fischer, E., J. Fitzgerald, W. Lechner and W. Lems
Research on incandescent lamps, II. Transport and burn-out in incandescent lamps 35,296
—, see Drop, P. C.
Fitzgerald, J. and H. Hörster
The temperature distribution in gas-filled incandescent lamps 32,206
—, see Fischer, E.
Flinn, I., see Murphy, N. St. J.
Franssen, N. V. and C. J. van der Peet
Digital tone generation for a transposing keyboard instrument 31,354
Fuller, K. L.
AVOID, a short-range high-definition radar 32, 13
— and A. J. Lambell
Traffic-flow analysis by radar 31, 17
Funk, W.
Thick-film technology 35,144
Garratt, G. R. M.
The early history of telegraphy 26,268
Gasser, R. and R. Hug
Current-limiting circuits for transistorized power supplies 28,251
Gast, J. G. C. de, see Kraakman, H. J. J.
Gelling, W. G., see Drift, A. van der
Gerlach, H. G.
The harp cathode, a cathode with low thermal inertia for small transmitting valves 26,309
Gerwen, P. J. van
The use of digital circuits in data transmission . . . 30, 71
—, see Greefkes, J. A.
Gibson, R. W.
PINPOINT — a radio system for locating and monitoring vehicles 35, 15
Gieles, A. C. M. and G. H. J. Somers
Miniature pressure transducers with a silicon diaphragm 33, 14
Goddard, N. E.
Solid-state microwave electronics 32,282
Gomes de Mesquita, A. H.
'PAILRED', an automatic single-crystal diffractometer 26,316
The polytypism of silicon carbide 30, 36
—, see Bril, A.
Gool, G. H. van, see Witmer, A. W.
Goorissen, J. and H. G. Bruijning
Doping methods for the epitaxial growth of silicon and germanium layers 26,194
Gossel, D.
Generation of musical intervals by a digital method 26,170
The frequency-analog signal as a basis for measurement and control 34,288
—, G. Kaps and W. Schott
A new system of digital circuit blocks for industrial measuring and control equipment 26,164
Gouiran, R.
Five major proton synchrotrons 30,330
Graaff, H. C. de and H. Koelmans
The thin-film transistor 27,200
Greebe, C. A. A. J.
Electromagnetic, elastic and electro-elastic waves . . 33,309
Greefkes, J. A., P. J. van Gerwen and F. de Jager
Companders with a high degree of compression of speech level variations 26,215
— and K. Riemens
Code modulation with digitally controlled companding for speech transmission 31,335
Green, D. B., see Asselman, G. A. A.
Grijseels, R. A.
An electronic computer for air traffic control . . . 26,226
Grimmeiss, H. G., W. Kischio and H. Scholz
Gallium phosphide light sources and photocells . . . 26,136
Groenhuis, H. and W. L. L. Lenders
Impact-mounting of components on printed wiring panels 27, 28
Groot, J. de and A. Mircea
Computer calculations of the Gunn effect 32,385
Groot, J. J. de, J. A. J. M. van Vliet and J. H. Waszink
The high-pressure sodium lamp 35,334
Groth, R. and E. Kauer
Thermal insulation of sodium lamps 26,105
—, see Boort, H. J. J. van
Guétin, P. and G. Schréder
Tunnelling in metal-semiconductor contacts under pressure 32,211
Guldemond, J., see Douma, F.
Haaijman, P. W.
Integration of electronic circuits 27,180
Haan, E. F. de, F. M. Klaassen and P. P. M. Schampers
An experimental 'Plumbicon' camera tube with increased sensitivity to red light 26, 49
Haanstra, H. B.
Stereophotography with the electron microscope . . 27,231
— and W. F. Knippenberg
Crystal growth of silicon carbide (II) 26,187
— and S. C. Rademaker
The frosting of glass 29,142
— and H. Zijlstra
Electron-microscope investigation of the domain structure of magnetic particles 29,218
—, see Nieuwenhuizen, J. M.
Haarhuis, G. J.
The Philips helium liquefier 29,197
Haas, C., see Albers, W.
Haas, G.
Problems and trends in the development of peripheral equipment for computers 26,148
Haas, L. A. de and S. S. Wadman
The Waalre complex of Philips Research Laboratories 31,153
Haeringen, W. van, see Lang, H. de
Hanbury Brown, R.
The stellar interferometer at Narrabri, Australia, I. The principles of the intensity interferometer; general description 27,141
Hanneman, H. W., see Blom, D.
Hansen, H. N.
Modern carrier telephone systems 26,206

Hansen, P.
Magnetic anisotropy and magnetostriction of Ru and Ir ions in yttrium iron garnet 35,225
— and J.-P. Krumme
The compensation wall 34, 96
Härdtl, K. H.
A simplified method for the isostatic hot pressing of ceramics 35, 65
—, see Andrich, E.
Harkema, P.
Automatic telegraph exchanges with electronic stores 26,240
Harrison, F. W., R. F. Pearson and K. Tweedale
Single crystal research on ferrites and garnets 28,135
Hart, C. M. and A. Slob
Integrated injection logic (I^2L) 33, 76
Hart, P. A. H. and F. M. Klaassen
The MOS transistor as a small-signal amplifier . . . 31,216
Harten, H. U., R. Memming and G. Schwandt
Investigations on the germanium-electrolyte interface 26,127
Havas, P. G.
Measurement of wire-diameter variations 31,111
Sorting of quartz tubing for high-pressure mercury-vapour lamps 32, 57
Hazan, J. P., see Dumont, F.
Hector, F.
The RAMP inertial navigation system 29, 69
Heek, H. F. van, see Bouwer, A. G.
Heide, H. van der
Stabilization by oscillation 34, 61
—, see Brockman, F. G.
Heijne, L., is the same person as Heyne, L.; see below
Heise, J., see Brinkman, A. C.
Helden, G. W. M. T. van, see Velde, T. S. te
Hentley, E. L., see Smith, F. W.
Hesselink, J., see Ammers, H. van
Heusden, S. van
Air-pollution monitors based on chemiluminescence 34, 73
Heuven, J. H. C. van
P-I-N switching diodes in phase-shifters for electronically scanned aerial arrays 32,405
— and A. G. van Nie
Microwave integrated circuits 32,292
Heyne, L.
Physical principles of photoconductivity,
II. Kinetics of the recombination process; sensitivity and speed of response 27, 47
III. Inhomogeneity effects 29,221
— and A. T. Vink
A Pirani gauge for pressures up to 1000 torr and higher 30,166
—, see Beekmans, N. M.
Hily, C.
Objective lenses for infra-red image formation . . . 30,290
Hoek, W. J. van den and W. A. Klessens
Carbon-filament lamps with a chemical transport cycle 35,316
Hoek, W. van der, see Bähler, W.
Hof, A. P. M. van 't, see Oosterkamp, W. J.
Hofker, W. K.
Semiconductor detectors for ionizing radiation . . . 27,323
—, K. Nienhuis and J. C. Post
α-particle spectrometry with semiconductor detectors 30, 13
Hofmeester, J. H. M. and J. P. Koutstaal
Moving-coil motors 33,244
Holm, W., see Breimer, H.
Holster, P. L., C. J. Th. Potters and H. F. G. Smulders
A water-pressure operated control system for dishwashers 33, 29
Hoogendijk, A. H., see Bussche, W. van den
Hoogeveen, L. P. J., see Butzelaar, P. F.
Horn, B. L. ten
Forty years of workshop technology 31,127
Hornstra, J. and H. Vossers
The Philips PW 1100 single-crystal diffractometer . . 33, 61
Hörster, H., E. Kauer and W. Lechner
The burn-out mechanism of incandescent lamps . . 32,155
—, see Fitzgerald, J.
Hug, R., see Gasser, R.
Hughes, R. C., see Bronnes, R. L.
Hurck, N. van, F. L. H. M. Stumpers and M. Weeda
Stereophonic radio broadcasting,
I. Systems and circuits 26,327
II. Susceptibility to interference 27, 62
Iperen, B. B. van and W. Kuypers
A klystron multiplier for generating 0.8 mm waves . 26, 61
Jacobs, C. A. J., see Beijer, L. B.
Jager, C. de, see Brinkman, A. C.
Jager, F. de, see Greefkes, J. A.
Jansen, J. A. J., see Witmer, A. W.
Janssen, J. H. H., see Bosch, G.
Janssen, P. J. M. and P. E. Day
Control mechanisms in the Philips 'VLP' record player 33,190
Jatteau, M.
Infra-red thermography equipment for medical applications 30,278
Jelmorini, G., see Essers, W. G.
Jochems, P. J. W. and E. Kooi
Metallurgical aspects of the alloy-diffusion method in transistor technology 28,246
Jonker, G. H. and A. L. Stuijts
Controlling the properties of electroceramic materials through their microstructure 32, 79
Josephy, R. D.
MOS transistors for power amplification in the HF band 31,251
Kalis, H. and J. Lemmrich
Frequency-analog speed control 33,260
Kalter, H., see Appels, J. A.
Kamerbeek, E. M. H.
Electric motors 33,215
Torque measurements on induction motors using Hall generators or measuring windings 34,153
Scaling laws for electric motors 35,116
Kampen, H. van
Magnetic tape store for telegraph characters 26,250
Kaps, G.
'Controlled cascading', a new open-loop control principle for adjustable frequency dividers 32,103
—, see Gossel, D.
Kauer, E.
Generating light with selective thermal radiators . . 26, 33
—, see Groth, R.
—, see Hörster, H.
Kessel, Th. J. van and R. J. van de Plassche
Integrated linear basic circuits 32, 1
Keve, E. T.
Pyroelectric materials based on triglycine sulphate (TGS) for infrared detection 35,247
King, H. N. G.
Electron beam processes 28,174
Kischio, W., see Grimmeiss, H. G.
Klaassen, F. M.
The development of field-effect transistor electronics 33,203
—, see Haan, E. F. de
—, see Hart, P. A. H.
Klaassen, J. A. and S. H. de Koning
Motional feedback with loudspeakers 29,148
Klerk, M.
The electron microprobe 34,370
Klessens, W. A., see Hoek, W. J. van den
Klopfer, A., see Dettingmeijer, J. H.
Klostermann, F. T.
A step-and-repeat camera for making photomasks for integrated circuits 30, 57
—, see Bouwer, A. G.
Knippenberg, W. F.
Inorganic chemical analysis 34,298
—, B. Lersmacher, H. Lydtin and A. W. Moore
Pyrolytic graphite 28,231
— and G. Verspui
Crystal growth of silicon carbide (III) 29,252
—, G. Verspui and J. Visser
A silicon carbide mortar 26,340
—, see Haanstra, H. B.
Knudsen, B., see Vermeulen, J.
Kock, A. J. R. de
Vacancy clusters in dislocation-free silicon and germanium 34,244

Koedam, M.
Determination of small dimensions by diffraction of a laser beam 27,208
—, see Bleekrode, R.
Koelmans, H., see Graaff, H. C. de
Koning, S. H. de, see Klaassen, J. A.
Kooi, E., see Appels, J. A.
—, see Jochems, P. J. W.
Kooi, H. van der
Monitoring the control-rod operation of the nuclear reactor at Dodewaard (Netherlands) 30, 48
Koster, W. G., see Vredenbregt, J.
Köstlin, H.
Double-glazed windows with very good thermal insulation 34,242
Koten-Hertogs, M. van, see Wessels, J. S. C.
Koutstaal, J. P., see Hofmeester, J. H. M.
Kraakman, H. J. J. and J. G. C. de Gast
A precision lathe with hydrostatic bearings and drive 30,117
Kramer, C., see Bloemendal, W.
Kramer, P., see Compaan, K.
Kreuwels, W. G. J.
Structural testing of digital circuits 35,261
Krim, J. E. V., see Backmark, N.
Kroon, D. J. and M. Q. Mengarelli
Monitoring the quality of surface water 34,113
Kruishoop, J. C. W.
Non-polar electrolytic capacitors 29, 37
Krul, L. G., and P. Reijnierse
Transmission of simple pictures 32, 42
Krumme, J.-P., see Hansen, P.
Kuus, G.
A getter for metal-iodide high-pressure mercury-vapour lamps 35,354
Kuypers, W. and J. C. Tiemeijer
The Philips PSEM 500 scanning electron microscope 35,153
—, see Iperen, B. B. van
Laan, P. van der, see Adriaanse, R. P.
Laar, J. van and J. J. Scheer
Photoemission of semiconductors 29, 54
Ladell, J., see Cath, P. G.
Lakerveld, H. G.
High-speed solid-rotor induction motors 34,170
Lambell, A. J., see Fuller, K. L.
Lang, H. de, E. T. Ferguson and G. C. M. Schoenaker
Accurate digital measurement of displacements by optical means, I. Displacement measurement with a reflection phase grating 30,149
— and G. Bouwhuis
Accurate digital measurement of displacements by optical means, II. Displacement measurement with a laser interferometer 30,160
—, D. Polder and W. van Haeringen
Optical polarization effects in a gas laser 32,190
Lange, H. K. A. de
A video tape recorder for non-professional use . . . 26,177
Lechner, W.
Peltier cooling 27,113
—, see Fischer, E.
—, see Hörster, H.
Leenhouts, J. I.
Stereo drawings made with a digital computer, II. Stereographic presentation of the 'Duphaston' molecule 29,110
Leeuwen, H. J. van and J. Vredenbregt
A muscle stimulator for hemiplegic patients 30, 23
Legoux, R.
The transfer technique, a new method for activating cathodes of phototubes 30,234
Transparent windows for a wide spectral range . . . 30,299
Lely, P. van der and G. Missriegler
Audio tape cassettes 31, 77
Lemke, M. and W. Schilz
Microwave integrated circuits on a ferrite substrate . 32,315
Lemmrich, J., see Kalis, H.
Lemogne, A., see Brissot, J.-J.
Lems, W., see Fischer, E.
Lenders, W. L. L., see Groenhuis, H.
Lersmacher, B., see Knippenberg, W. F.
Lewin, I. H.
Drilling of diamonds using electron beams 28,177
Lindop, R. W.
An experimental electronically controlled car transmission 28,179
Lootsma, F. A., see Dekkers, A. J.
Lopitzsch, K. H.
A variable inductor for r.f. heating control 27,134
Loty, C., see Andrieux, G.
Louwerse, M. W., see Brockman, F. G.
Lucas, D. A., see Alcock, R. N.
Luiten, A. L.
Superconducting magnets 29,309
(see also addendum, Vol. 30, p. 147)
Luiten, W., see Buschow, K. H. J.
Lüthje, H., see Dimigen, H.
Lydtin, H., see Knippenberg, W. F.
Maaren, A. C. van, O. Schob and W. Westerveld
Boron filament: a light, stiff and strong material . . 35,125
Magarshack, J.
Gunn-effect oscillators and amplifiers 32,397
Man, Th. J. de
Vitamins in live-stock feeding 28, 85
Manley, B. W., see Adams, J.
Mannetje, J. J. 't, see Bouwer, A. G.
Marie, G.
Large-screen projection of television pictures with an optical-relay tube based on the Pockels effect . . . 30,292
—, see Donjon, J.
Meijer, A.
An analogue computer for simulating one-dimensional aerial arrays 31, 2
Meijer, F. and G. A. Bootsma
Investigation of the chemical behaviour of clean silicon and germanium surfaces 32,131
—, see Brongersma, H. H.
Meijer, R. J.
Prospects of the Stirling engine for vehicular propulsion 31,169
—, see Rietdijk, J. A.
Meijering, J. L.
Phase theory,
I. Introduction to unary and binary systems . . . 26, 12
II. Quantitative considerations on binary systems . 26, 52
III. Ternary systems 27,213
Memming, R., see Harten, H. U.
Mengarelli, M. Q., see Kroon, D. J.
Mesman, W., see Banga, U. H.
Meulen, E. van der, see Duis, J. A. ten
Meyer, D.
A process for converting quantities to be measured into frequencies 29,189
Miedema, A. R.
A simple model for alloys,
I. Rules for the alloying behaviour of transition metals 33,149
II. The influence of ionicity on the stability and other physical properties of alloys 33,196
— and J. W. F. Dorleijn
Electrical conduction in ferromagnetic metals . . . 35, 29
Millett, E. J., see Clegg, J. B.
Mircea, A., see Groot, J. de
Missriegler, G., see Lely, P. van der
Mitchell, R. F.
Acoustic surface-wave filters 32,179
Moerkens, J. C.
Dimming fluorescent lamps 27,265
An electronic starter for long fluorescent lamps . . . 31, 54
Moore, A. W., see Knippenberg, W. F.
Mouthaan, K.
IMPATT-diode oscillators 32,345
Munk, E. C. and A. Rademakers
Integrated circuits with evaporated thin films 27,182
Murphy, N. St. J., F. Berz and I. Flinn
Carrier mobility in MOS transistors 31,237
Naastepad, P. A., see Buschow, K. H. J.
Newcomb, C. V. and E. C. Snelling
Analysis of variability in ferrite cored inductors . . . 28,184
Nicholas, K. H.
Studies of anomalous diffusion of impurities in silicon 28,149
Nie, A. G. van, see Heuven, J. H. C. van
Nie, C. P. van
Improved ear-lobe clip for physiological transducers 33,102

Nielen, J. A. van
Operation and d.c. behaviour of MOS transistors . . 31,209
—, M. J. J. Theunissen and J. A. Appels
MOS transistors in thin monocrystalline silicon layers 31,271
Nienhuis, H., see Bouwknegt, A.
Nienhuis, K., see Hofker, W. K.
Nienhuis, R. J.
Integrated audio amplifiers with high input impedance and low noise 31,245
A MOS tetrode for the UHF band with a channel 1.5 μm long 31,259
Niessen, C. and H. E. J. Wulms
Automatic drawing of masks for integrated circuits . 30, 29
Nieuwenhuizen, J. M. and H. B. Haanstra
Microfractography of thin films 27, 87
Nijenhuis, W. and H. van de Weg
Developments in the field of electronic computers during the last decade 26, 67
Nobel, D. de and M. T. Vlaardingerbroek
IMPATT diodes 32,328
Nussli, J.
Photomultipliers and the transfer technique 30,236
—, see Piétri, G.
O'Hanlon, H.
Ferrite-cored kicker magnets 31, 66
Okkerse, B. and P. Penning
Anomalous transmission of X-rays in highly perfect crystals 29,114
Okumura, T.
The MOS tetrode 30,134
Ommering, R. Ch. van and G. C. M. Schoenaker
The 'COBRA', a small digital computer for numerical control of machine tools 27,285
Oosterhout, G. W. van
Measurement of the density of small particles . . . 28, 30
Oosterkamp, W. J., A. P. M. van 't Hof and W. J. L. Scheren
X-ray pictures in colour 27,228
Oostrom, A. van
Field emission of electrons and ions 33,277
Oostvogels, F., see Drop, P. C.
Opstelten, J. J., D. Radielović and J. M. P. J. Verstegen
Optimum spectra for light sources 35,361
Orton, J. W., D. H. Paxman and J. C. Walling
The physics of maser materials 28,146
Osenbruggen, C. van
High-precision spark machining 30,195
Otterloo, P. van
Attitude control for the Netherlands astronomical satellite (ANS) 33,162
Oudemans, G. J.
Continuous hot pressing 29, 45
Overbeek, A. J. W. M. van and W. A. J. M. Zwijsen
Tunable integrated circuits 27,264
Overgoor, B. J. M.
An integrated chopper circuit with MOS transistors . 31,248
Palfreeman, J. S.
An opto-acoustic cross-correlator in radar signal detection 28,217
Pannenborg, A. E.
History of Philips research laboratories in Germany . 26, 89
Paxman, D. H., see Orton, J. W.
Peacock, R. V.
Critical parameters of evaporated NiFe films for fast stores 28,188
Pearson, R. E., see Davies, R.
Pearson, R. F., see Harrison, F. W.
Peet, C. J. van der, see Franssen, N. V.
Penning, P., see Okkerse, B.
Périlhou, J.
An experimental endoscope with miniature television camera 35,166
Piétri, G. and J. Nussli
Design and characteristics of present-day photomultipliers 29,267
Pistorius, J. A., J. M. Robertson and W. T. Stacy
The perfection of garnet bubble materials 35, 1
Plantinga, G. H., J. W. Rommerts and Th. J. Westerhof
An 8 mm reflex klystron of simple design 27,310
— and Th. J. Westerhof
An experimental reflex klystron for 1.5 mm wavelength 28,284
Plassche, R. J. van de, see Kessel, Th. J. van
Ploos van Amstel, J. J. A.
A technique for depositing thin and thick films . . . 28, 57
Polaert, R. and J. Rodière
Investigation of microchannel plates by scanning electron microscopy 34,270
—, see Eschard, G.
Polder, D., see Lang, H. de
Polman, J., H. van Tongeren and T. G. Verbeek
Low-pressure gas discharges 35,321
Poorter, T.
Flame-failure control with a UV-sensitive cold-cathode tube, I. Design of the tube and circuit . . . 29,254
Poorter, J. A. de, see Bril, A.
Post, J. C., see Hofker, W. K.
Potters, C. J. Th., see Holster, P. L.
Prast, G.
A gas refrigerating machine for temperatures down to 20 °K and lower 26, 1
Pré, F. K. du, see Daniels, A.
Premsela, H. F., see Broerse, P. H.
Proper, J., see Albrecht, C.
Rabenau, A.
Chemical transport reactions 26,117
— and H. Rau
Crystal growth and chemical synthesis under hydrothermal conditions 30, 89
Rademaker, S. C., see Haanstra, H. B.
Rademakers, A., see Drift, A. van der
—, see Munk, E. C.
Radielović, D., see Opstelten, J. J.
Radziwill, W.
A highly efficient small brushless d.c. motor 30, 7
Steady-state performance of a class of electronically commutated d.c. machines 35,106
Raes, R. and J. Schellekens
A speed-controlled d.c. motor for a washing machine 34,163
Rakels, C. J., J. C. Tiemeijer and K. W. Witteveen
The Philips electron microscope EM 300 29,370
Rathenau, G. W.
Innovation in electronic devices 32,117
Rau, H., see Rabenau, A.
Rehder, L., see Bleekrode, R.
Reijnen, P. J. L.
Thermogravimetric analysis applied to ferrites . . . 31, 24
Reijnierse, P., see Krul, L. G.
Reinhoudt, J. P.
A flywheel for stabilizing space vehicles 30, 2
Remmers, G.
Grease-lubricated spiral groove bearing for a straight-through shaft 27,107
Grease-lubricated helical-groove bearings of plastic . 34,103
Renelt, G. and J. Schröder
Data checking during input and transmission by means of one or two check characters 26,156
Rennicke, K.
Supply-voltage speed control for capacitor motors . 34,180
Revuz, J., see Desvignes, F.
Riemens, K., see Greefkes, J. A.
Rietdijk, J. A.
The expansion ejector, a new cryogenic device . . . 28,243
—, H. C. J. van Beukering, H. H. M. van der Aa and R. J. Meijer
A positive rod or piston seal for large pressure differences 26,287
Roberts, E. D.
Electron resists for the manufacture of integrated circuits 35, 41
Robertson, J. M., see Pistorius, J. A.
Robinson, S. J. and P. T. Saaler
A survey of coaxial and strip-line microwave components 28,211
Rodière, J., see Polaert, R.
Roeder, E.
Extrusion of glass 32, 96
Roerink, J. H. G., see Bosgra, O.
Rommerts, J. W., see Plantinga, G. H.
Ronde, F. C. de, see Butterweck, H. J.
Roosmalen, J. H. T. van
Experimental electrostatically focused 'Plumbicon' tubes 28, 60

—, see Broerse, P. H.
Röschmann, P.
YIG filters 32,322
Rossier, D., see Dumont, F.
Ruiter, J. W. de, see Boumans, P. W. J. M.
Saaler, P. T., see Robinson, S. J.
Sangster, F. L. J.
The 'bucket-brigade delay line', a shift register for analogue signals 31, 97
Integrated bucket-brigade delay line using MOS tetrodes 31,266
Saraga, P., J. A. Weaver and D. J. Woollons
Optical character recognition 28,197
Schagen, P. and A. W. Woodhead
Image converter and intensifier research 28,161
Schampers, P. P. M., see Haan, E. F. de
Scheer, J. J., see Laar, J. van
Schellekens, J., see Raes, R.
Schemmann, H.
Stability of small single-phase synchronous motors . 33,235
Scheren, W. J. L., see Oosterkamp, W. J.
Schiller, C. and M. Boulou
Investigation of crystal defects by cathodoluminescence 35,239
Schilz, W., see Lemke, M.
Schipper, D. J., see Bouwknegt, A.
Schmidt-Tiedemann, K. J.
Transport phenomena in solids 26, 99
Schmitz, A.
Solid circuits 27,192
Schob, O., see Maaren, A. C. van
Schoenaker, G. C. M.
Stereo drawings made with a digital computer, I. Making stereo drawings with the COBRA-controlled drawing machine 29,102
—, see Lang, H. de
—, see Ommering, R. Ch. van
Schoenmakers, W. J.
A sensitive torque meter for high frequencies 27,342
Schöler, H. F. L., see Bruin, O. A. de
Scholz, H.
Crystal growth by temperature-alternating methods . 28,316
—, see Grimmeiss, H. G.
Schott, W., see Gossel, D.
Schouten, J. F.
The EVOLUON, a permanent Philips exhibition . . 31,187
Schréder, G., see Guétin, P.
Schreuder, D. A.
Physiological aspects of the lighting of tunnel entrances 27, 76
Schröder, J.
Examples from fluorine chemistry and possible industrial applications 26,111
Temperature profiling of tungsten filaments in incandescent lamps by a chemical transport reaction . . 35,332
—, see Dettingmeijer, J. H.
Schröder, J., see Renelt, G.
Schulten, G.
Surface-wave transmission lines for microwave frequencies, II. Applications of the dielectric line . . . 26,350
Schwandt, G., see Harten, H. U.
Sellberg, F., see Backmark, N.
Severin, H.
Surface-wave transmission lines for microwave frequencies, I. The various types of transmission line . 26,342
Shaievitz, S.
Cryogenic production of ultra-pure hydrogen . . . 26, 27
Shannon, J. M.
Ion-implanted high-frequency MOS transistors . . . 31,267
Slob, A.
Fast logic circuits with low energy consumption . . 29,363
—, see Hart, C. M.
Slofstra, C.
The use of digital circuit blocks in industrial equipment 29, 19
Sluyterman, L. A. Æ.
The relation between enzyme, substrate and product 27,160
Smets, A. J.
The optical sensors of the Netherlands astronomical satellite (ANS), I. The sun sensors 34,208
Smith, F. W., P. L. Booth and E. L. Hentley
Masers for a radio astronomy interferometer 27,313
Smulders, H. F. G., see Holster, P. L.
Snelling, E. C.
Q contour charts 28,186
—, see Newcomb, C. V.
Somers, G. H. J., see Gieles, A. C. M.
Sommerdijk, J. L. and A. Bril
Phosphors for the conversion of infrared radiation into visible light 34, 24
Spiesberger, W. and M. Tasto
The automatic measurement of medical X-ray photographs 35,170
Stacy, W. T., see Pistorius, J. A.
Steen, L. M. van der
Digital integrated circuits with MOS transistors . . 31,277
Steinmaier, W., see Brandt, B. B. M.
Strachan, A. J., see Brandt, B. B. M.
Stratum, A. J. A. van, see Zalm, P.
Stuijts, A. L.
Renaissance in ceramic technology 31, 44
—, see Jonker, G. H.
Stumpers, F. L. H. M., see Hurck, N. van
Sweet, R. C., see Bronnes, R. L.
Tan, S. L., see Breimer, H.
—, see Broerse, P. H.
Tasto, M., see Spiesberger, W.
Teale, R. W., see Enz, U.
Thees, R.
Small electric motors 26,143
Theunissen, M. J. J., see Nielen, J. A. van
Tichelaar, G. W., see Essers, W. G.
Tielemans, P. A. W., see Bouwknegt, A.
Tiemeijer, J. C., see Kuypers, W.
—, see Rakels, C. J.
Tijburg, R., see Acket, G. A.
Timmerman, J.
Two electromagnetic vibrators 33,249
T'jampens, G. R. and M. H. A. van de Weijer
Gas-filled incandescent lamps containing bromine and chlorine 27,173
Tol, T., see Beijer, L. B.
Tongeren, H. van, see Polman, J.
Tooren, A. van
A simple and flexible automatic extractor 35,196
Trier, P. E.
The Mullard Research Laboratories, an outline of their growth and function 28,129
Troye, N. C. de
Digital integrated circuits with low dissipation . . . 35,212
Tummers, L. J.
MOS transistors 31,206
Tweedale, K., see Harrison, F. W.
Ubbink, J., see Casimir, H. B. G.
Veen, R. van der
The control of agricultural pests and diseases through the ages 28,121
Veer, S. M. de, see Brouwer, H. J.
Velde, T. S. te and G. W. M. T. van Helden
Monograin layers 29,238
Velzel, C. H. F.
Holographic strain analysis 35, 53
Venema, A.
A fast cryopump system for ultra-high vacuum . . . 28,355
The flow of highly rarefied gases 33, 43
Verbeek, H. J.
A two-stage compressor with rolling diaphragm seals 30, 51
Verbeek, T. G., see Polman, J.
Verbunt, J. P. M.
'Droplet interferometry' for investigating smooth surfaces 33, 74
Verheijke, M. L.
Neutron activation analysis 34,330
— and A. W. Witmer
Line-intensity calculations for X-ray fluorescence analysis 34,339
Verloop, A.
Availability of organic compounds in higher plants . 28, 93
Vermeulen, J. and B. Knudsen
The design of road lighting for given luminance and uniformity 29,299

Verspui, G., see Knippenberg, W. F.
Verstegen, J. M. P. J., see Opstelten, J. J.
Verster, N. F.
The exploration of the unknown 26,321
Vincent, R. P., see Alcock, R. N.
Vink, A. T., see Heyne, L.
Vink, H. J.
Crystal defects and their transport in solids; industrial applications 35,181
Visscher, J. M.
A high-speed punched-tape reader 28,259
Visser, J., see Knippenberg, W. F.
Vlaardingerbroek, M. T. and K. R. U. Weimer
Beam-plasma amplifier tubes 27,275
—, see Nobel, D. de
Vliet, J. A. J. M. van, see Groot, J. J. de
Vlietstra, J.
The APT programming language for the numerical control of machine tools 28,329
—, see Dekkers, A. J.
Volger, J.
Superconductivity 29, 1
Vortices 32,247
Volman, H. J. W. M.
The 'push-pull' spiral-groove bearing — a thrust bearing with self-adjusting internal preloading 35, 11
Vonk, G.
A compact heat exchanger of high thermal efficiency 29,158
Vonk, R.
Magnetic deflection in television picture tubes . . . 32, 61
Voorman, J. O., see Blom, D.
Vorm Lucardie, J. A. van der
Television transmitters for the ultra-high frequency band 26,256
Vossers, H., see Hornstra, J.
Vredenbregt, J. and W. G. Koster
Analysis and synthesis of handwriting 32, 73
—, see Leeuwen, H. J. van
Vriens, L. and M. Adriaansz
Investigation of gas discharges by light scattering . . 35,344
Vries, E. B. de, see Dinklo, J. A.
Vrijer, F. W. de
Colour television transmission systems 27, 33
Waard, P. J. de
Anomalous oscillations with an IMPATT diode . . 32,361
—, see Acket, G. A.
Wadman, S. S., see Haas, L. A. de
Wal, J. van der
A fast refractometer for evaporated thin films . . . 35,142
Walling, J. C., see Orton, J. W.
Walop, J. N., see Bruin, O. A. de
Walraven, A.
Controlling the speed of small induction motors by means of thyristors 28, 1
Walther, K.
Magnetoacoustic effects in bismuth 32,233
Wanmaker, W. L., see Bril, A.
Waszink, J. H., see Groot, J. J. de
Weaver, J. A., see Saraga, P.
Weeda, M., see Hurck, N. van
Weg, H. van de, see Nijenhuis, W.
Weijer, M. H. A. van de, see T'jampens, G. R.
Weimer, K. R. U., see Vlaardingerbroek, M. T.
Werner, H. W.
Mass spectrometer analysis of a solid surface 27,344
—, see Brongersma, H. H.
Wesselink, G. A., see Drop, P. C.
Wesselius, P. R., see Aalders, J. W. G.
Wessels, J. H., see Bussche, W. van den
Wessels, J. S. C. and M. van Koten-Hertogs
Photosynthesis. A survey of the present state of research 27,241
Westendorp, F. F., see Buschow, K. H. J.
Westerhof, Th. J., see Plantinga, G. H.
Westerveld, W., see Maaren, A. C. van
Wieringen, J. S. van
The Mössbauer effect and its application in solid state research 28, 33
Wijnen, M. D., see Boter, P. A.
Witmer, A. W., J. A. J. Jansen, G. H. van Gool and G. Brouwer
A system for the automatic analysis of photographically recorded emission spectra 34,322
—, see Verheijke, M. L.
Witteman, W. J.
Sealed-off high-power CO_2 lasers 28,287
Witteveen, K. W., see Rakels, C. J.
Woodhead, A. W., see Schagen, P.
Woollons, D. J., see Saraga, P.
Wulms, H. E. J., see Niessen, C.
Wylie, A. W. and P. L. Eisler
Prospecting with neutrons 30, 97
Wytzes, S. A.
An automatic X-ray spectrometer 27,300
Zalm, P. and A. J. A. van Stratum
Osmium dispenser cathodes 27, 69
Zeedijk, H., see Brouwer, H. J.
Zeida, R., see Desvignes, F.
Zieler, E., see Elgström, H.
Zijlstra, H.
A vibrating-reed magnetometer for microscopic particles 31, 40
Materials research for permanent magnets 34,193
—, see Haanstra, H. B.
Zonneveld, J. A.
Stereo drawings made with a digital computer, III. A solution of the Van der Pol equation, represented in three dimensions 29,112
Zwijsen, W. A. J. M., see Overbeek, A. J. W. M. van